Advance Access publication June 17, 2013

Political Analysis (2013) 21:298–313
doi:10.1093/pan/mpt002

Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark

Will Lowe
MZES, University of Mannheim
e-mail: will.lowe@uni-mannheim.de

Kenneth Benoit
Department of Methodology, London School of Economics and the Department of Political Science, Trinity College, Dublin
e-mail: kbenoit@lse.ac.uk (corresponding author)

Edited by R. Michael Alvarez

Automated and statistical methods for estimating latent political traits and classes from textual data hold great promise, because virtually every political act involves the production of text. Statistical models of natural language features, however, are heavily laden with unrealistic assumptions about the process that generates these data, including the stochastic process of text generation, the functional link between political variables and observed text, and the nature of the variables (and dimensions) on which observed text should be conditioned. While acknowledging statistical models of latent traits to be "wrong," political scientists nonetheless treat their results as sufficiently valid to be useful. In this article, we address the issue of substantive validity in the face of potential model failure, in the context of unsupervised scaling methods of latent traits. We critically examine one popular parametric measurement model of latent traits for text and then compare its results to systematic human judgments of the texts as a benchmark for validity.

A vast amount of effort in political science focuses on estimating characteristics of political actors—parties, legislators, candidates, voters, and so on—that may be estimated, but never directly observed. Whether we call them "ideal points," policy preferences, topics, or issue emphases, these latent traits and latent classes are not only fundamentally unobservable but also exist in a dimensional space that is fundamentally unknowable.[1] This has hardly prevented political researchers from attempting to identify and estimate such quantities, however, and a variety of such methods are widely used. Many, such as the analysis of roll call votes, suffer from problems of data censorship and selection that produce biased estimates of the quantities desired. Not all actors vote, co-sponsor bills, or return our questionnaires, but there is one activity that always accompanies political action: speech. This simple fact, coupled with a revolution in the availability of vast quantities of recorded text and speech, has spurred the development of a wide range of methods for analyzing textual data, most of which are surveyed in Grimmer and Stewart (2013).

Every statistical model applied to data—textual or otherwise—requires assumptions. As Grimmer and Stewart (2013) point out, such assumptions are always wrong. For textual data, these assumptions concern the process that generates the observed textual data, including the stochastic process of text generation; the functional model linking political variables of interest and observed text; and the nature of the variables (and dimensions) on which observed text should be conditioned. The reality is that even though we know that these assumptions are "wrong," we have no real benchmark by which to assess the *consequences* of or the degree of wrongness, because

Authors' note: Replication materials for this article are available from the *Political Analysis* dataverse at http://hdl.handle.net/1902.1/20387. Supplementary materials for this article are available on the *Political Analysis* Web site.

[1] For a good survey with examples of this problem, see Benoit and Laver (2012).

the data-generating process for natural language cannot ever be really "known" and therefore can never be reliably simulated. Moreover, the quantities we seek to estimate from text, such as latent traits for scaling models, or latent classes for topic models, are fundamentally unobservable. They can only be defined as inferences from, and therefore as inherently dependent on, the assumptions of some model. This core problem leaves us in a position of applying models of an unknown conditional data-generating process to estimate latent quantities in unknown dimensional spaces that can never be directly checked.

In what follows, we contend that the key question is not how wrong models of text are, but rather how *useful* these models are in helping us obtain valid measures of real political quantities. Furthermore, we focus on the hardest class of text models to validate, unsupervised scaling models, which we agree with Grimmer and Stewart (2013) have fewer direct methods for validation. Their challenge is that to validate the results from unsupervised methods, "scholars must combine experimental, substantive, and statistical evidence to demonstrate that the measures are as conceptually valid as measures from equivalent supervised methods." Although in many contexts, this is both appropriate and feasible—as shown in the analysis of credit-claiming, for instance, in Grimmer and Stewart (2013; Fig. 6)—there are many cases where hand coding is practically infeasible. The scaling of continuous latent traits such as ideology, for instance, may be both costly and conceptually difficult for texts of any significant size without resorting to a system of strict coding rules, and existing examples of such schemes have been shown to suffer extreme reliability problems (see Mikhaylov, Laver, and Benoit 2012). Yet, scaling ideology is a core activity in political measurement, because no model of political competition can be tested without reliable and valid information on the relative preferences of political actors.

Scaling ideology is not as simple as comparing supervised models. In general, supervised models assume much *less* about the data-generating process than unsupervised scaling models while assuming more about the quantity estimated (see, e.g., Jordan 1995). Consequently, a successful supervised model confirms that enough information is available in the data to allow the relevant imposed distinctions, whereas a well-validated unsupervised model demonstrates that the assumption of specific additional process structure, here conditional independence, is sufficient to recover it without external imposition. For these models, the core question is instead what Grimmer and Stewart (2013) call "semantic validity": whether the quantity being scaled reflects the quantity that the analyst intends to measure. Similar to the approach used by Grimmer and King (2011) to assess the semantic validity of unsupervised clustering, we apply experimentally elicited human input to validate an unsupervised scaling of relative ideological preferences. By demonstrating a detailed research design for validating estimates from quantitative text, we not only agree with Grimmer and Stewart (2013) but also show precisely *how* such a validation framework can be deployed. Rather than comparing the unsupervised results to those from a supervised model, we compare the unsupervised scaling results directly to elicited human judgments of the texts. Furthermore, we not only use human perceptions of *relative position* to validate the point estimates from the scaling model but also use human perceptions of *difference* to evaluate the estimates of uncertainty from statistical methods for scaling textual data. This second aspect returns directly to the point on which Grimmer and Stewart (2013) and we agree, namely, that our oversimplified models of the textual data-generating process are wrong, but provides a concrete assessment of *how* wrong these are by comparing the consequences for inference.

Our practical example comes from a set of speeches made and against a historic austerity budget debate in the Irish parliament in late 2009. For testing, we compare the results of the Poisson scaling model of Slapin and Proksch (2008) to a systematic rating of these same texts by twenty human readers. We have chosen the Poisson scaling model because its assumptions about the word data-generating process are the most explicit, and existing validation methods borrowed from the computer science literatures are not directly applicable, as we discuss further below. In addition, scaling latent traits such as policy positions are of central importance to empirical and theoretical political science, since without reliable and valid measurements of these concepts, it is impossible to test models of party competition, representation, or policy outcomes. The validation design we outline is easily understandable and can be easily replicated and extended to almost any research problem involving quantitative text, subject to sufficient human resources.

1 A Validation Design for Textual Data Analysis

Text generates unique quantitative data in that prior to being converted into numbers, text can be interpreted directly through a qualitative process of human reading. After all, the central purpose of text is to communicate a message to a reader or listener. Judging the message contained in a text, whether written or spoken, is something that humans do every day. This makes textual data fundamentally different from other forms of data, such as test item responses, the coded responses from completed survey questionnaires, or simple continuous quantitative data such as years of schooling, which do not convey a direct message through a careful construction that includes a rich vocabulary, a meaningful sequence, and a grammar and syntax. Other forms of quantitative data are often most meaningfully interpreted when aggregated and summarized, whereas with text as data, this process is reversed. When disassembled into quantitative information, typically a term-document matrix of word-type frequencies, humans can no longer make sense of textual data as they can in its raw form. Furthermore, few quantitative scales for measuring textual data have a natural metric whose summary interpretation is self-evident. This poses challenges for validating quantitative models of text as data, but also presents unique opportunities.

Our validation framework sets the following expectations: First, *valid positional estimates* from quantitative scaling, for a given dimension, should match a human reader's placement of these texts with respect to identifying relative differences along this dimension. Furthermore, this dimension should be clear to the reader and self-evident from reading of the texts. Second, *meaningful estimates of uncertainty* from a quantitative text scaling model should yield statistical conclusions of similarity or difference that correspond to a human reader's perceptions of difference between pairs of the same set of texts. A model that is "useful" in statistical decision-making will at least roughly correspond to this benchmark of human judgment of similarities and differences.

In the natural language processing, human qualitative interpretation has long formed the benchmark for validating automated content analysis methods. "Supervised" methods that assign complete documents to classes on the basis of qualitative human category assignments have long been validated by specificity and sensitivity (equivalently precision and recall; see Manning, Raghavan, and Schütze 2008). "Unsupervised" versions of these methods that attempt to simultaneously learn categories and their assignment to documents, such as clustering methods, as well as those that attempt to learn categories and to decompose documents into category proportions—known as "topic models"—present greater validation challenges while still using category-oriented methods (Wallach et al. 2009; Grimmer and King 2011). These cannot be applied directly to the validation of "unsupervised" scaling models, however, because we lack any kind of category labeling to work with, and there typically is no "test" set whose values are known. Moreover, because the latent traits that political scientists typically wish to estimate are often fundamentally unobservable—such as left-right ideology or some positive or negative level of affect—although they are selected to correspond to some set of characteristics or categories that humans have defined as meaningful and interesting. For unsupervised methods, human validation is all the more important to establish not only the correctness of the estimates but also the semantic validity of the scales or classes.

2 How Text Models Are "Wrong"

Statistical methods for scaling latent traits have received widespread attention in political science in the last decade. Here, we focus on the parametric scaling model formulated by Slapin and Proksch (2008) because it contains explicit, strong assumptions. As we will demonstrate through our validation exercise, while there are many linguistic reasons to recognize the drastically simplifying assumptions of this model's word generation process as wrong, the model nonetheless produces sufficiently valid results to be extremely useful as a measurement model of latent political traits.

In the scaling model—a reparameterization of Goodman's (1979) row–column (RC) association model—the count of the jth word in the ith document, C_{ij}, is a Poisson process with rate conditional on the document's position θ_i:

$$\begin{aligned} C_{ij} &\sim \text{Poisson}\,(\lambda_{ij}) \\ \log \lambda_{ij} &= \alpha_i + \psi_j + \theta_i \beta_j. \end{aligned} \tag{1}$$

Word parameters β and ψ, incidental document-level parameters α, and the speaker positions θ are jointly estimated by alternating conditional maximum likelihood.[2] Confidence intervals for θ_i can be estimated either asymptotically using the information matrix of the likelihood conditional on the word parameter estimates, by using a parametric bootstrap as explained by Slapin and Proksch (2008), or by alternative nonparametric bootstrap methods described below.

This model rests on several strong statistical assumptions to produce valid point estimates and confidence intervals for θ that we investigate in what follows.

2.1 *Unidimensional Latent Trait*

One important assumption of the Poisson scaling model in equation (1) is that the set of texts contains word counts generated from a single dimension of relevant variation. The difficulties involved in identifying dimensionality and the consequence of over- and underestimating dimensionality in exploratory factor analysis and related models are well known and no easier when text is involved. One text-specific issue does arise, however: the presence of extra dimensions is confounded with the existence of the kind of topic or frame structure that topic models are designed to extract. For these models, a "topic" consists roughly of a subset of vocabulary with strongly intercorrelated usage. Scaling models will instead treat such correlations as indications of position and scale them accordingly. This implies that the dimensions of difference identified through scaling may differ from the dimensions of difference intended by the researcher, and this is a possibility for which we must be vigilant.

2.2 *Conditional Independence*

The assumption that observed word counts are conditionally independent means that each word is generated independently from others according to a Poisson process with rate specified by the model parameters according to equation (1). Word count variation from causes other than expressed position is assumed to be noise. When this assumption is false, then information from one observed word provides information about the probability that another word is observed. These residual correlations mean that each new word observed provides less information than if it had been independently generated. Consequently, parameter uncertainty, in particular uncertainty about θ, will be underestimated.

There are two related problems with respect to conditional independence: unmodeled lexical associations that cause serial dependence and contemporaneous correlation in the form of a hierarchical document structure.

First, to the extent that text scaling models (at least those that treat only words as data) do not account for lexical associations, for example, collocations, compound nouns, and names, these sorts of conditional independence failure are to be expected. In defense of the conditional independence assumption, Laver, Benoit, and Garry (2003) point out that some single words *do* have strong directional associations—the word "tax" and its variants, for instance, is used almost exclusively by more right-leaning parties (who prefer to cut taxes). However, this fails to distinguish all sorts of politically interesting differences among taxes, such as "income taxes," "taxes on banks," "carbon taxes," "inheritance taxes," and "capital gains taxes." Not all words, however, tend to have such

[2] The model is also straightforward to implement in a Bayesian-MCMC framework using random effects for all parameters, and indeed was fit for two dimensions this way by Monroe and Maeda (2004).

stable associations. The word "free," for instance, is used for both "free enterprise" (a right-leaning phrase) as well as "GMO-free" and "free health care" (left-leaning phrases). From this perspective, it may seem remarkable that text scaling models based purely on the relative frequencies of atomic words—what linguists call the "bag of words" approach—work at all.[3]

2.3 *Words as a Poisson Process*

The assumption that word-count rates are conditionally distributed as Poisson implies that the variance of this rate is equivalent to the expected rate—a strong assumption that may not hold in natural language word counts. This may be because variation in θ may be present in manifestos as different ideological wings of a party add their own sections of text, a fact modeled explicitly by Lo, Proksch, and Slapin (2011). Unmodeled variation in the rate of word occurrence for fixed θ may also result from linguistic features. Interestingly, this may be over- *or* underdispersion. For example, in English, each sentence contains on average about one instance of the word "the." This regularity is very strong: in the Irish budget debate speeches we examine in more depth later, the rate per hundred words of the word "the" is 7.28 with variance 0.75, about ten times smaller than even a Poisson model with no covariates would predict. Structural zeros are another frequently encountered feature of term-frequency matrixes, caused when a word has *no* chance of occurring in some documents, for example, the term "European Union" prior to the 1980s (when the EU was still called the European Economic Community), or in the party manifestos of Australia where EU policy was simply never a feature of the political discourse. The counterpart of structural absences is when informative words that occur may trigger additional occurrences of the same words. This dependency may be either the result of a real dependency between nearby observations—"true contagion" (or "burstiness"; see Church and Gale 1995)—or when there is merely "apparent contagion" due to variation or serial correlation in values of θ (see Cameron and Trivedi 1998, 106 for a review).

To conclude, the problems with conditional independence point to a fundamental observation about applying measurement models to the text scaling task: We have very little idea about what the functional form of the relationship between Y and θ is. The best we can do is identify the model assumptions that fail, be realistic (but sanguine) about the limits of our models given these assumption failures, and—if possible—seek ways to correct them. The dilemma is that we simply have no "true" benchmark to compare the point estimates and confidence intervals from scaling models' estimates from natural language texts as data. Furthermore, standard techniques designed to test estimator assumption failure, such as Monte Carlo simulation, offer little respite because we have no realistic model for simulating natural language text. We can simulate quantitative data that looks like the quantitative form of natural language text, but only by relying on the very assumptions whose appropriateness we would like to test. For what appears to be a very useful model, in other words, we have no means through standard methods of validating the estimates it produces or of assessing how confidence about the estimates. This points to the need for some other form of validating the strong assumptions of quantitative models of text as data: one based on a human rating of the texts in their original, qualitative forms.

3 Data and Methods

Parliamentary speech has been analyzed previously with an aim to locating legislators' policy preferences (e.g., Monroe and Maeda 2004; Proksch and Slapin 2010), but the dimensions of policy measured in these applications have been less clear.[4] Such problems point to a need to choose texts where the topic is plausibly limited to a single dimension of difference, and where a

[3] Zhang (2004) argues that tight collocations like the ones above need not compromise inferences about θ in models with strong conditional independence assumptions, although he admits they will still bias uncertainty estimates downward.

[4] In Monroe and Maeda (2004), for instance, the primary dimension that emerged from a two-dimensional scaling model of US Senate speeches was labeled the "workhorse/showhorse" dimension, for want of a better interpretation. Proksch and Slapin (2010) had to interpret their single estimated dimension from the European Parliament by resorting to correlations with roll-call vote analysis and independent expert surveys.

lot of external information exists on speaker positions that can be used to assess the validity of the text scaling results. For political texts, this suggests a debate where the format and content of text are limited to a single topic: in our case, a budget debate dominated by a single dimension of willingness to accept the burden of austerity measures.

3.1 *Texts: Legislative Debates over the 2010 Irish Budget*

The set of texts we use for comparing human to Poisson-scaled estimates comes from the debate following the presentation of the Irish budget of 2010, taking place in December 2009 in the Irish *Dáil*, the lower house of the Irish parliament. At the time, this budget was widely acknowledged to be the harshest budget in Irish history. In a total of fourteen speeches by key members of each of five political parties, speakers urged either adoption of the harsh fiscal measures as a necessary measure to get the economy back on track, or rejection of the budget as unfair, unnecessary, or unworkable, along with a rejection of the government proposing it. On the government side, speeches by the *Taoiseach* (Prime Minister) Brian Cowen of the governing Fianna Fáil party, and Finance Minister Brian Lenihan of the same party, represented the most pro-budget positions. Three speeches from Green party ministers (Gormley, Cuffe, and Ryan) provided support for the budget, but somewhat more reluctantly, as many in the Green party regretted the austerity measures but felt bound to support the budget by the terms of their party's coalition agreement with Fianna Fáil. On the opposition side, the leaders of the Fine Gael and Labour parties—in addition to two deputies from the opposition, anti-system Sinn Féin party—showed the greatest opposition to the budget, and had allied in an opposition pact to replace the governing coalition in the next election. In all, the budget debates provide a good example of text expressing positions that plausibly reflect a single dimension of relative preference for fiscal austerity versus social protection, and also directly relate to the approval or rejection of specific legislation.

Table 1 lists the fourteen texts we analyze, by speaker and political party, along with the number of total words (tokens) and unique words (types). The median text had 3629 total words, although the shortest contained just 919, and 361 unique words (total 1644). Overall, the corpus contained 49,019 tokens and 4840 different word types. Our construction of the term-document matrix did not exclude any words, such as "stop words" thought *a priori* to be politically uninformative or very low-frequency words such as the sixty-seven hapaxes found in the corpus. Nor did we apply

Table 1 Quantitative summary of 2010 budget debate texts

Speaker	*Party*	*Tokens*	*Types*
Brian Cowen	FF	5842	1466
Brian Lenihan	FF	7737	1644
Ciaran Cuffe	Green	1141	421
John Gormley	Green	919	361
Eamon Ryan	Green	1513	481
Richard Bruton	FG	4043	947
Enda Kenny	FG	3863	1055
Kieran O'Donnell	FG	2054	609
Joan Burton	LAB	5728	1471
Eamon Gilmore	LAB	3780	1082
Michael Higgins	LAB	1139	437
Ruairi Quinn	LAB	1182	413
Arthur Morgan	SF	6448	1452
Caoimhghin O'Caolain	SF	3629	1035
All texts		49,019	4840
Minimum		919	361
Maximum		7737	1644
Median		3704	991
Hapaxes		67	

a stemmer to the words, although tests showed virtually identical results when applied to the set of stemmed words.[5]

Working with these fourteen texts, we selected a panel of human raters to read and evaluate the position expressed in each text, an exercise described next.

3.2 *The Qualitative Coding Exercise*

The objective of the statistical scaling model is to estimate the latent positions θ_i from a term-document matrix, along a single dimension of difference. The challenges for such a model lie in knowing whether the dimension of difference on which positions are estimated in the scaling model, in fact, correspond to the dimension expected by the researcher, and whether the estimated positions driven by the relative differences in term frequencies accurately reflect each text's position on this dimension. Related to this is the question of meaningfully measuring our *uncertainty* about these point estimates: whether statistical estimates of uncertainty reflect human confidence in perceived differences between texts. Our validation methodology targets both position and uncertainty by asking human readers to assess both, by reading the original texts and answering a series of structured questions about texts and pairs of texts. This takes human judgment beyond the realm of post hoc face validity, into generating independent judgments against which statistical estimates of textual traits may be compared directly.

Our experiment consisted of printing a collection of the budget debates as an eighty-four-page booklet, prefaced by detailed instructions (with full details and examples available in the supplementary materials to this article). The first speech was clearly identified as the introduction of the budget by Finance Minister Lenihan, abridged to remove from the middle of the speech some of the (tedious and specific) technical details of the budget measures. Prime Minister Cowen's speech followed, also slightly abridged to reduce the length caused by technical detail in the middle section. The remaining twelve speeches were referred to only by number. The instructions encouraged readers to make notes in or on the margins of each text, or in a box for notes provided at the end of each text. Most readers made moderate to extensive notes, as judged by the booklets returned to us for coding.

Each speech was followed by at least one question asking the reader to compare the speech just read to a previous speech. We did not ask readers to attempt all ninety-one possible pairwise comparisons, but were able to gauge their perception of differences on twenty-five of these pairs, including all pairs of party leaders, many key intraparty pairs, and several additional pairs. If the reader did perceive one of the two texts in the pairwise comparison to be more pro-budget than the other, then the questionnaire asked the reader to assess his or her confidence in there being a difference, on a scale from one ("Not at all confident") to ten ("Very confident").

At the end of the qualitative exercise, readers were asked to use their notes to place each text on scale of 0–100, with zero indicating complete support for ("lack of opposition to") the budget, and one hundred indicating complete opposition to the budget. Speeches 1 and 2 from Lenihan and Cowen, respectively, were fixed to zero. Readers had a choice of indicating a point estimate, drawing an interval, or doing both. Most chose to use intervals or some combination of intervals and point estimates, indicating that they understood the instructions and examples and were using the intervals to approximate a confidence region in which they felt the speaker's true position lay.

A total of nineteen readers completed the questionnaire, in exchange for a modest (fixed) honorarium. The readers consisted of five government postgraduate students from a course in qualitative text analysis from the London School of Economics, one post-doctoral fellow from the London School of Economics (LSE) associated with text analysis, nine PhD students and

[5] One reason for not excluding common words ("stop words") is that this presumes that we know which words are uninformative. For instance, the pronouns "he" and "she" emerge as highly discriminant words, with Poisson scaling parameter estimates of $\beta = -0.54$ and $\beta = -1.68$, respectively. This is similar to the results of Monroe, Quinn, and Colaresi (2008), who found that uncorrected partisan association measures for female pronouns (of the kind that the Poisson scaling model uses) indicated that they were Democrat words.

three MSc students in political science from Trinity College Dublin, and one PhD student from New York University.[6]

4 Results: Human Qualitative Placements of the Texts

We present the results of the human judgment exercise first because it forms the benchmark by which we will judge the validity of the automated scaling model's estimates of the position of each text.

In Fig. 1a, we plot the mean positions according to the direct human placement of each budget speech along the 0–100 scale, after reflecting the scale (so that one hundred indicates complete support rather than the original zero) and rescaled the means to the unit normal. The bars indicate the 95% confidence intervals assuming the distribution sampling means are *t*-distributed (and the Fianna Fáil deputies Cowen and Lenihan have no intervals because their positions were fixed in the questionnaire at zero). These positions represent our confidence that the point estimate of the speaker's position, on the scale we have asked respondents to place each speech, is contained in this interval. By aggregating the judgment of each human expert into a mean across experts, we have averaged out individual differences in scale usage as well as perceptions of difference to form a "consensus" position. Similar methods have been advocated and defended at length for the use of expert ratings of party positions by Benoit and Laver (2006).

The results shown in Fig. 1a confirm what almost any knowledgeable observer of Irish politics would expect. The degree of antipathy toward the budget, according to the human expert readers, neatly separates opposition parties on the left from government parties on the right, indicated in Fig. 1a by the dashed red line. Also, in accord with expectations, we see strong similarities in intraparty positions, where speakers of the same party tended to cluster together in their positions. Taking far, pro-budget positions are the *Fianna Fáil* deputies Brian Cowen and Brian Lenihan, Taoiseach and Finance Minister respectively. Slightly less pro-budget were the Greens, a position we would expect given their reluctant identification with the governing Fianna Fáil whose economic policies were widely seen to be the cause of the financial crisis necessitating the austerity budget being debated. On the opposition side were Fine Gael and Labour, two would-be governing partners whose speakers' positions overlapped, although the median Labour speaker was more opposed to the budget than the median Fine Gael speaker, in line with expectations. Since they reject not only the government and its budgets but also the system itself, it is also entirely plausible that Sinn Féin would be the most opposed to the government's austerity budget, a result also affirmed by the human placements.

Since we will lean heavily on the human results, we also checked their internal consistency by comparing explicit positions to pairwise judgments. Figure 1b shows the result of applying a Bradley-Terry model for the paired comparisons.[7] The Bradley-Terry model provides an alternative scaling method of the budget positions. Using $>$ to mean "is more pro-budget than" and denoting two speakers to be compared by a and b, this model assumes that $\text{logit}\, P(a > b) = \theta_a - \theta_b$, with θ suitably normalized for identification.[8] The results indicate that the human judgments themselves are very consistent across the two types of questions we asked regarding the positions of the speeches, as the placement of people and parties is nearly exactly the same as the direct scaling discussed previously.[9]

[6] In order to test for any effect of prior knowledge of Irish politics, coder education, or prior coding experience, we also compared the results according to a variety of coder-specific variables, but found no statistically significant differences in any factors.

[7] For estimation, we used the R package BradleyTerry2 (Firth 2005).

[8] Following Firth's (2005) suggestions, responses indicating a pair of speeches that could not be distinguished were distributed evenly between the other two responses. Removing pair judgments marked as uncertain perfectly preserved the ordering of speakers.

[9] The uncertainty estimates in Fig. 1b are wide for three reasons. First, we asked a rather small subset of possible comparisons to reduce respondent fatigue, second because pairwise comparisons are inherently less informative than, for example, scalar judgments, and third because we do not make use of the uncertainty information that respondents provided.

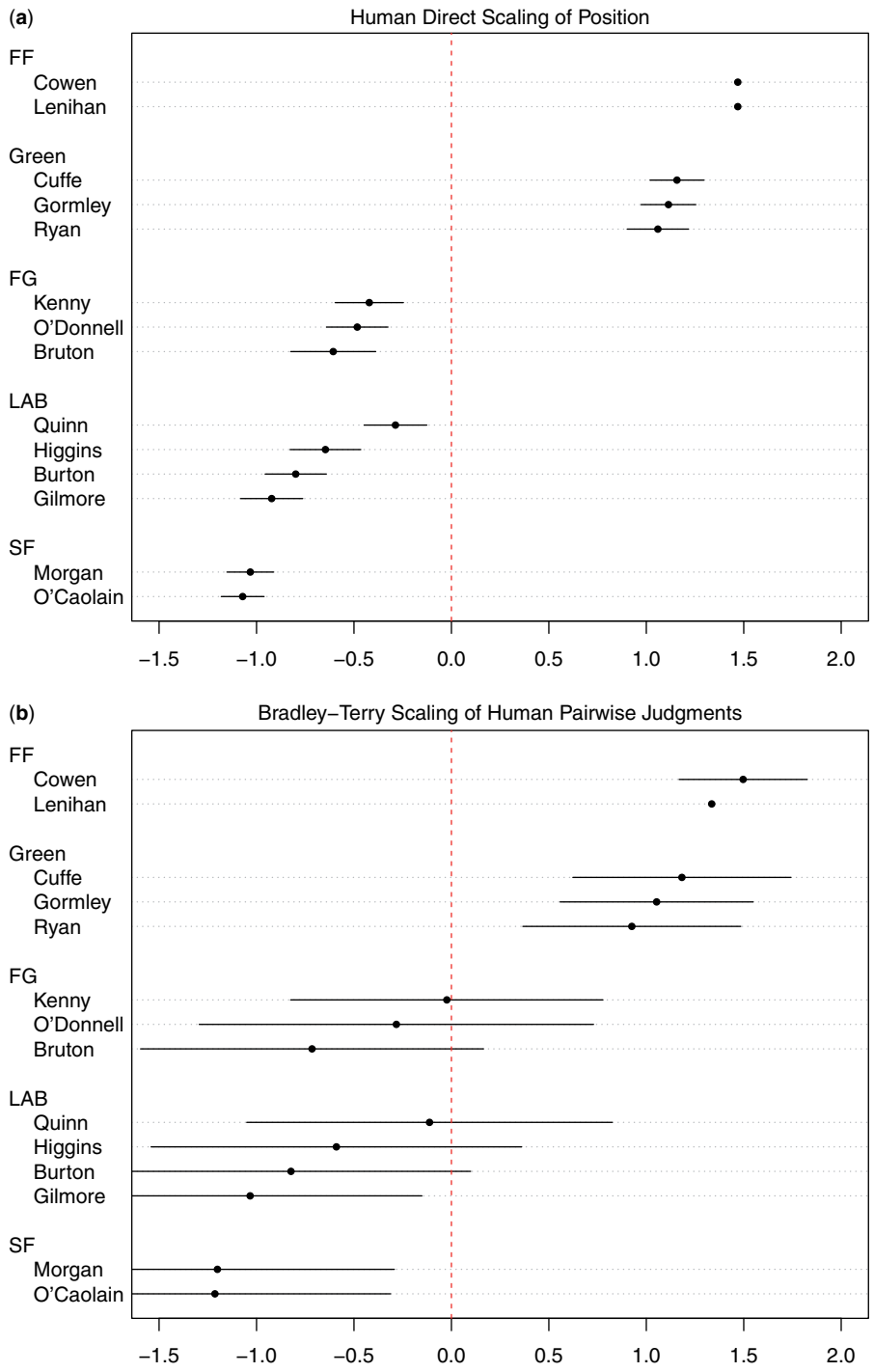

Fig. 1 Human placement results on anti- versus pro-budget scale, human direct scaling (a) and Bradley-Terry scaling from pairwise comparisons (b). Human scaling results are means of the 0–100 placements, with 95% confidence intervals assuming a standard normal sampling distribution of means. The Bradley-Terry results are rescaled to the unit normal for comparison, and were estimated using Lenihan as an anchor point (at zero, prior to rescaling).

5 Results: Validating the Quantitative Text Scaling Results

Having established the benchmark for assessing the placement of each speech on a pro- versus anti-budget divide, and having confirmed the validity of these placements, we now turn to the comparison of the quantitative scaling results against the human ratings to illustrate how validation of quantitative text scaling can be performed.

5.1 *Poisson Scaling Results*

Figure 2 presents the results of the Poisson scaling model on the unmodified speeches, along with 95% confidence intervals computed from the maximum likelihood estimation. (The details of this method for computing the confidence intervals, which rely on the model assumptions being correct, are discussed below.) As it turns out, the results correspond to the human placements very closely, correctly separating government and opposition (in the plot, by the dashed red line) and clustering each set of speakers by party. As in the human ratings, the opposition parties of Labour and Fine Gael opposed the budget in the same order as placed by the human raters. On the government side, the Greens and Fianna Fáil argued strongly for the government position. Slightly less pro-budget were the Greens, as expected. On the opposition side were Fine Gael and Labour, two would-be governing partners whose speakers' positions overlapped, although the median Labour speaker was more opposed to the budget than the median Fine Gael speaker, in line with expectations. Using "face validity" as a criterion for model assessment, the Poisson scaling gets at least a B+ for these results.

To compare the methods directly, we have plotted in Fig. 3 the statistical estimates against the human-scaled results, along with the standard parametric 95% confidence intervals from the Poisson scaling model and the standard errors of the mean from the human-scaled results. The dashed cross-hair lines (in red) divide government and opposition for both sets of estimates, as expected. Most parties cluster in similar groups with only minor differences. In the Poisson-scaled estimates, for example, Fine Gael opposition leader Enda Kenny was the most critical of the

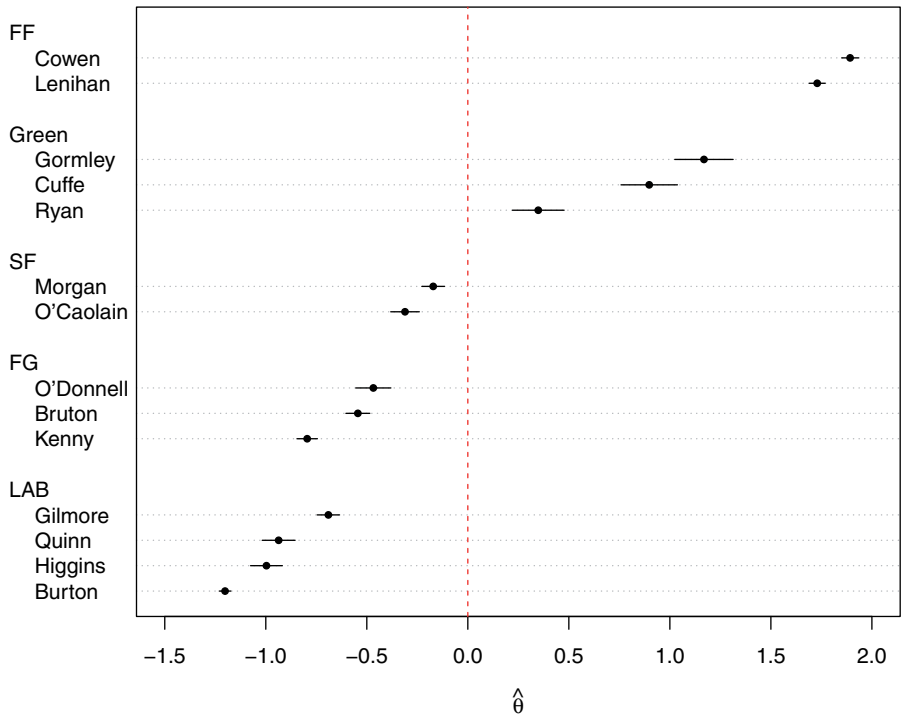

Fig. 2 Poisson scaling results.

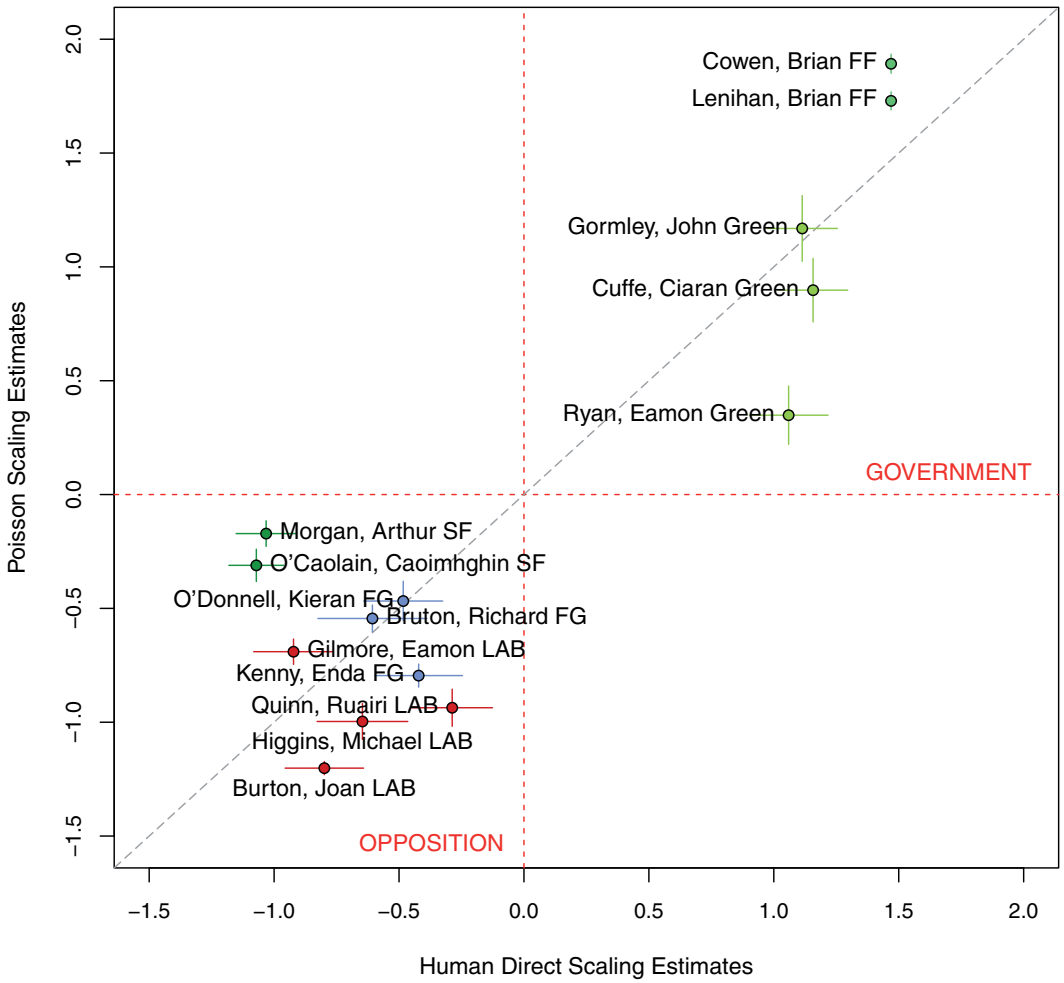

Fig. 3 Direct comparison of human versus Wordfish estimates.

budget, although in the human placements, the Fine Gael positions were indistinguishable. Among Labour speakers, party leader Eamon Gilmore's speech was the most pro-budget in the statistical results, but was rated by human readers as the most opposed to the budget, although they could not distinguish it from Labour Deputy Burton's. For the Green party, the statistical results found significant differences between the three Green ministers, whereas the qualitative placement found their positions indistinguishable.

All in all, most estimates of the speaker positions now lie along the 45° axis of agreement (shown by the long dashed gray lines), except for one party: Sinn Féin. The relatively middle position of Sinn Féin as on the opposition side of the budget, yet still between the main opposition and government party blocs, is incongruent with the human placements who clearly regarded the Sinn Féin speeches as the most extreme expressions of an anti-budget position. The Poisson scaling results for Sinn Féin thus results contradict our expectations based on the knowledge of Irish politics, as well as the results established from the human placements and pairwise comparisons. They suggest a model failure or that some deeper look at the textual data is needed.

5.2 *Explaining the Incongruent Position of Sinn Féin*

Among the main political parties in Irish politics, Sinn Féin is unusual in that one of its key political goals is to redefine the Irish state to reunite the Republic of Ireland with the northern counties currently part of the United Kingdom. As a left-wing party committed to democratic socialism, Sinn Féin considers itself the champion of the poor, minorities, migrants, and the working

class—all groups who would feel most keenly the bite of the austerity measures called for in the government's budget. Sinn Féin is also Euro-skeptic, having campaigned for a "No" vote in the Lisbon referendum held in 2008, and vociferously opposed both the creation of the National Asset Management Agency in 2009 to take over the bad loans of Ireland's failing banks, and the acceptance in the same year of an €85 billion rescue package from the European Union and the International Monetary Fund (IMF).

Prior to the budget debate, Sinn Féin publicized its own budget plan, an alternative to both the government budget announced by Minister Lenihan and the Fine Gael proposal articulated by Finance Spokesperson Richard Bruton. This plan called for a nearly €4 billion economic stimulus package, to be offset by raising €3.7 billion in revenue mainly through increased taxation, especially on the wealthiest and highest earners.[10] In the Dáil debates, Sinn Féin Deputy Arthur Morgan expressed not only his rejection of the budget in no uncertain terms but also his rejection of the Fine Gael–Labour alternative:

> Replacing Fianna Fáil and the Green Party with Fine Gael and the Labour Party will make no difference to economic recovery . . . Fine Gael and the Labour Party would implement the same policies in a different package, with the same bad results for the economy. While the establishment parties close ranks and display a disturbing uniformity in their policies . . . we are unique because we are the only party with an alternative analysis of the situation.

This statement suggests a possible second dimension to the budget debate, not captured here, in the one-dimensional scaling result. To the extent that the opposition to the budget is also opposition to the government, the speeches made by Sinn Féin Deputy Morgan and Dáil party leader Caoimhghín Ó Caoláin are expressing a rejection not only of the budget but also of both "system" party alternatives, including Fine Gael and Labour. The unidimensional scaling results, however, are unable to detect this difference, and estimate the Sinn Féin positions as lying in the middle of the axis of opposition to and support for the budget. Furthermore, there is no easy "fix" to this problem such as trimming a section of irrelevant text from the Sinn Féin speeches, because the anti-establishment language is interwoven with the Sinn Féin commentary on the budget.[11] Although there may be more sophisticated methods for "fixing" the Sinn Féin's speeches, these are neither simple nor generalizable.

6 Results: Evaluating the Uncertainty Estimates

As we saw from the confidence intervals in the direct comparison presented in Fig. 3, many differences *within* party suggested by the Poisson scaling results are judged indistinguishable by humans. This suggests either that the statistical scaling model is capable of detecting more nuanced differences than the human readers or—far more probably—that the standard errors from the statistical estimates are unrealistically small. Having used the human placements as a benchmark to validate the point estimates, we now draw on the aggregate pairwise judgments of difference to validate, or at least to pass some judgment on, the approximate validity of the standard errors produced by the Poisson scaling model.[12] For these differences to be meaningful for text analysis, we focus only on judgments of *difference*: substantively meaningful confidence intervals for positional estimates should assess differences in positions in a manner that corresponds to human perceptions of difference between two speeches, from a qualitative reading. If humans cannot detect a difference in the position of two texts, then a statistical model that declares them to different is underestimating the true level of uncertainty.

[10] *The Road to Recovery: Sinn Féin Pre-Budget 2010 Submission*, http://www.sinnfein.ie/files/2009/Pre-Budget2010_small.pdf.

[11] In supplementary materials to this article, we also show how Green Deputy John Gormley's position is wrongly estimated when a section of off-topic speech is included within his text.

[12] This approach holds strong similarities to the pairwise comparisons asked of human raters in Grimmer and King (2011), where a three-scale judgment of similarity from pairwise comparisons of documents was used to evaluate the performance of automated clustering methods on different texts.

The asymptotic maximum likelihood method we use to quantify uncertainty in the Poisson scaling model relies on three assumptions: that the model is correctly specified, that there are enough data available for the curvature of the log likelihood to be approximately quadratic, and that the word parameters ψ and β are sufficiently well estimated that they can be treated as known. By explicitly conditioning on document lengths, we remove the α parameters that decouple the likelihoods for each position parameter (see Lowe and Benoit 2011 for further details). If the first assumption is maintained, but we are less confident about the second and third, then we can instead use a bootstrap method (Davison and Hinkley 1997). Slapin and Proksch (2008) recommend a parametric bootstrap to maintain the assumption of a correctly specified model.

Both this and the asymptotic method share a key assumption: the model is correctly specified in all respects. Therefore, we attempt a less model-dependent method by *non*parametric bootstrapping directly from the data itself. This involves extending a method that is now well established for nontextual data to resampling from the text itself, *prior to* conversion into the quantitative matrix required for the application of the statistical model. Here, we make a deliberately simple first step, albeit a very significant one, away from the restrictive assumptions of the parametric Poisson scaling model, by resampling texts from their constituent words. More elaborate alternatives are possible, and while we have explored them in other work (Lowe and Benoit 2011), here we focus only on the simplest of these methods, by bootstrapping texts from their words.[13]

To assess the judgments of difference from the human ratings, we draw on the pairwise comparison questions following each text, in which a human reader compared the text just read to other texts previously read. In Table 2, we report the results of pairwise comparisons testing the "null hypothesis" of no difference between the pair of texts indicated in each row. In the "Qualitative/Pairwise" column, we report the results of the aggregated (modal) judgment of the human raters whether there was a detectable difference between the two texts. The "Qualitative/Scaling" column is based on a test of $H_0 : \theta_a = \theta_b$ for the pair of texts indicated, applying a t-test of the difference in sampling means of the human (0–100) placements. The "Poisson/Analytical" and "Poisson/Parametric Bootstrap" columns report similar tests using the estimated standard errors of each $\hat{\theta}_i$. In the "NP BS" (nonparametric bootstrap) column, we report the test that the middle 95% region of bootstrap replicates of $\hat{\theta}_a - \hat{\theta}_b$, from one hundred replicates, included zero.

Similar to the comparisons of errors, the most conservative measure (in terms of estimating uncertainty) was the nonparametric bootstrap method of Poisson scaling, which found only sixteen of the twenty-five pairwise comparisons to be different.

The asymptotic method and the parametric bootstrap lead to numerically very similar measures of uncertainty, although the bootstrap offers persistently slightly larger uncertainty measures for positions in this setting. This is presumably because for small numbers of documents, the word parameters are *not* particularly well estimated.[14] Nevertheless, both methods are clearly overcertain relative to our respondents' judgments.

Although it would hardly be novel or controversial to apply a nonparametric bootstrap to any other form of quantitative data, our use of this technique to avoid over reliance on unverifiable model assumptions for the Poisson scaling model has, as far as we are aware, not been applied to the quantitative analysis of textual data in this fashion. There is a tradition of using the bootstrap in correspondence analyses involving text, for example, as discussed in Greenacre (2007, chap. 25). But the focus there is on stability rather than estimating sampling variation.[15]

All the nonparametric bootstrap methods resample from the observed data matrix rather than reconstructing textual data itself, *prior to* converting these data into quantitative form. Our bootstrapping method thus simulates the process of stochastic production of text as in Benoit, Laver, and Mikhaylov (2009): it produces texts that might have been observed, by resampling and

[13] Alternatives to resampling texts from words would involve *block bootstrapping* methods. In block bootstrapping, consecutive blocks of observations of length K are resampled from the original time series, in either fixed blocks (Carlstein 1986) or overlapping blocks (Künsch 1989).

[14] Empirically, we find that with larger numbers of documents the two methods converge.

[15] Also there is no model reestimation—a single model is fitted and resampled rows of the data matrix are projected onto the space of positions. This effectively assumes known word parameters.

Table 2 Pairwise comparisons of difference

	Qualitative		*Quantitative (Poisson) scaling*		
	Pairwise	*Scaling*	*Analytical*	*Parametric bootstrap*	*NP BS word*
Party leader comparisons					
Gilmore LAB versus Cowen FF	DIFF	DIFF	DIFF	DIFF	DIFF
Gormley Green versus Cowen FF	DIFF	DIFF	DIFF	DIFF	DIFF
Gormley Green versus Gilmore LAB	DIFF	DIFF	DIFF	DIFF	DIFF
Gormley Green versus Kenny FG	DIFF	DIFF	DIFF	DIFF	DIFF
Gormley Green versus O'Caolain SF	DIFF	DIFF	DIFF	DIFF	DIFF
Kenny FG versus Cowen FF	DIFF	DIFF	DIFF	DIFF	DIFF
Kenny FG versus Gilmore LAB	DIFF	DIFF	DIFF	DIFF	–
O'Caolain SF versus Cowen FF	DIFF	DIFF	DIFF	DIFF	DIFF
O'Caolain SF versus Gilmore LAB	–	–	DIFF	DIFF	DIFF
O'Caolain SF versus Kenny FG	DIFF	DIFF	DIFF	DIFF	DIFF
Within-party comparisons					
Cowen FF versus Lenihan FF	–	–	DIFF	DIFF	DIFF
Bruton FG versus Kenny FG	DIFF	–	DIFF	DIFF	DIFF
Cuffe Green versus Gormley Green	–	–	DIFF	DIFF	–
Ryan Green versus Cuffe Green	–	–	DIFF	DIFF	DIFF
Ryan Green versus Gormley Green	DIFF	–	DIFF	DIFF	DIFF
Burton LAB versus Quinn LAB	DIFF	DIFF	DIFF	DIFF	–
Burton LAB versus Higgins LAB	DIFF	–	DIFF	DIFF	–
Higgins LAB versus Quinn LAB	DIFF	DIFF	–	–	–
O'Caolain SF versus Morgan SF	–	–	DIFF	DIFF	–
Other comparisons					
Bruton FG versus Gilmore LAB	DIFF	DIFF	DIFF	DIFF	–
O'Donnell FG versus Burton LAB	DIFF	DIFF	DIFF	DIFF	DIFF
Ryan Green versus Morgan SF	DIFF	DIFF	DIFF	DIFF	–
Burton LAB versus Bruton FG	–	–	DIFF	DIFF	DIFF
Quinn LAB versus Bruton FG	DIFF	DIFF	DIFF	DIFF	–
Morgan SF versus Gilmore LAB	DIFF	–	DIFF	DIFF	DIFF
Total observed differences (max 25)	19	15	24	24	16

reconstructing texts that could have been generated from the same data-generating process, and then converts these into quantitative data. The advantages of this method are that it can be easily adapted to preserve any essential features of the textual data-generating process, such as the sequence, grammar, and syntax, simply by redefining the resampling units. Lowe and Benoit (2011), for instance, tested different forms of block bootstrap resampling, although here (primarily to keep the discussion focused) we use only a word-level bootstrap. A second advantage is its generality: There is no method of analyzing text as data to which a text-level bootstrapping cannot be applied. Bootstrapping by resampling units from original texts offers a plausible means of approximating the sampling distributions for just about any form of quantitative text analysis.

7 Discussion

From our analysis, comparing qualitative judgments from a panel of nineteen human readers about pairwise differences, uncertainty regarding those differences, and placements of each speech overall on an "anti-/pro-budget" dimension to results to those from a statistical scaling model of relative word frequencies, emerge two novel contributions. First, we have demonstrated a complete research design to assess the semantic validity of unsupervised text scaling models by benchmarking the results to a human cognitive level. By using a panel of readers and aggregating their perceptions,

we have a high degree of confidence in the human judgments. This knowledge combined with direct qualitative judgments, we argue, provides the most *meaningful* benchmark against which to assess quantitative scaling results, in terms of both positional estimates and confidence about these estimates. Although our application has focused on a unidimensional latent trait model, the approach is quite general and could be applied more generally to almost any class of statistical model that treats text as quantitative data.

Second, our validation exercise has demonstrated how even "wrong" models can be useful. Despite numerous political and linguistic reasons why the strict assumptions of the Poisson scaling model are violated—a simplistic model of word counts as a multilevel Poisson process conditional on an unobserved θ_i—we have shown that unsupervised text scaling can produce semantically valid results, benchmarked according to independent human judgments of the texts. In comparing the positional estimates from the qualitative and quantitative placements of each speaker, we found a high degree of correspondence, with two exceptions. First, human readers found the positions of the Green party speakers to be indistinguishable, while all methods of Poisson scaling except the nonparametric bootstrapping method found their positions to be different. Second, and more significantly, the Poisson scaled results placed the two Sinn Féin speakers in a middle position, whereas the human readers judged these two speeches to be the most anti-budget of all. This difference may reflect an additional dimension of opposition to both the government and the system-party opposition of Fine Gael and Labour, a dimension not picked up by the unidimensional statistical scaling model.

Our comparison also attempted to benchmark the uncertainty estimates of the parametric scaling model through both direct comparison of the interval sizes and pairwise judgments from human rating and statistical decision making, and we also observed marked differences. Although the standard parametric approaches to the Poisson scaling model judged almost all pairwise speakers' positions to be significantly different from one another (twenty-four out of twenty-five compared), methods based on human reading produced far fewer perceptions of speakers as different: nineteen from direct questions of difference and fifteen from the aggregated scale placements. The bootstrap method produced results much closer to these perceptions of difference, suggesting that this method offers a far less assumption-reliant, and hence more meaningful, method for estimating confidence intervals from quantitative scaling models of textual data. Wrong model assumptions do a reasonable job of recovering estimates of position but will, unless corrected for, vastly overstate our confidence in them. This result offers a valuable lesson for using "wrong" models, since it shows that while positional estimates may be largely correct, an over-reliance on wrong assumptions will lead us to overstate our confidence in these estimates.

A third contribution lies in the demonstration of text-level bootstrapping. In applying nonparametric bootstrapping methods to textual data, we have taken the treatment of "text as data" to a new level. Bootstrapping methods have been applied previously to coded text units from qualitative content analysis (e.g., Benoit, Laver, and Mikhaylov 2009), but not for the purposes of purely quantitative approaches using text as data. Our nonparametric approach contrasts with both model-dependent asymptotic error computation and parametric bootstrapping methods, by avoiding assumptions that are unrealistic for natural language text. Here, we have applied text bootstrapping to the Poisson scaling model, but this approach is completely general and can be used for any scaling or statistical model drawing on text as data, including classifiers, other scaling methods, topic models, or even descriptive textual statistics, such as vocabulary diversity, readability, or keyword difference measures.

Funding

European Research Council (ERC-2011-StG 283794-QUANTESS).

References

Benoit, K., and M. Laver. 2006. *Party policy in modern democracies*. London: Routledge.

———. 2012. The dimensionality of political space: Epistemological and methodological considerations. *European Union Politics* 13:194–218.

Benoit, K., M. Laver, and S. Mikhaylov. 2009. Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science* 53(2):495–513.

Cameron, A. C., and P. K. Trivedi. 1998. *Regression analysis of count data*. Cambridge, UK: Cambridge University Press.

Carlstein, E. 1986. The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *Annals of Statistics* 14:1171–79.

Church, K., and W. Gale. 1995. Poisson mixtures. *Natural Language Engineering* 1:163–90.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.

Firth, D. 2005. Bradley-Terry models in R. *Journal of Statistical Software* 12:1–12.

Goodman, L. A. 1979. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* 74(367):537–52.

Greenacre, M. 2007. *Correspondence analysis in practice*, 2nd ed. London: Chapman and Hall.

Grimmer, J., and G. King. 2011. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences* 108(7):2643–50.

Grimmer, J., and B. M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3):267–97.

Jordan, M. I. 1995. *Why the logistic function? A tutorial discussion on probabilities and neural networks*. Computational Cognitive Science Report 9503, MIT.

Künsch, H. R. 1989. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* 17:1217–41.

Laver, M., K. Benoit, and J. Garry. 2003. Estimating the policy positions of political actors using words as data. *American Political Science Review* 97:311–31.

Lo, J., S.-O. Proksch, and J. B. Slapin. 2011. Party ideology and intra-party cohesion: A theory and measure of election manifestos. Paper presented at MPSA 2011.

Lowe, W., and K. R. Benoit. 2011. Practical issues in text scaling models: Estimating legislator ideal points in multi-party systems using speeches. Paper presented at MPSA 2011.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press.

Mikhaylov, S., M. Laver, and K. Benoit. 2012. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis* 20:78–91.

Monroe, B., and K. Maeda. 2004. *Talk's cheap: Text-based estimation of rhetorical ideal-points*. Working paper, Michigan State University.

Monroe, B. L., K. M. Quinn, and M. P. Colaresi. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16:372–403.

Proksch, S.-O., and J. B. Slapin. 2010. Position taking in the European Parliament speeches. *British Journal of Political Science* 40(3):587–611.

Slapin, J. B., and S.-O. Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52(3):705–22.

Wallach, H. M., I. Murray, R. Salakhutdinov, and D. Mimno. 2009. Evaluation methods for topic models. Proceedings of the 26th International Workshop on Machine Learning, New York, NY.

Zhang, H. 2004. The optimality of Naïve Bayes. In *FLAIRS Conference*, eds. V. Barr and Z. Markov. Menlo Park, CA: AAAI Press.