Inferring Concepts from Topics:
Towards Procedures for Validating Topics as Measures

Luwei Ying
Washington University in St. Louis

Jacob M. Montgomery
Washington University in St. Louis

Brandon M. Stewart
Princeton University

July 11, 2019

Abstract

Unsupervised topic models were developed as a clustering approach for exploring and summarizing the structure of large document sets. Expanding beyond this interpretation to viewing topic models as a method for measuring latent concepts requires validation to ensure that the measure captures the desired quantity of interest. Bespoke methods for validation are effective but appear in the literature with increasing rarity and are not available in all cases. To date, few general-purpose procedures have been proposed for validating the results of topic models.

We extend an existing method from computer science that proposes crowd-sourced tasks for model selection using human judgment (Chang et al., 2009). This prior work evaluates whether word sets learned by a topic model appear semantically related, but does not validate that the model captures the substantive quantity implied by the researchers' topic label. We close this gap by designing and testing a suite of task structures which provide more direct evaluation of label validity. We show that our tasks are easier for workers to complete and provide better discrimination across models and potential labels. We evaluate the reliability of worker judgments in these tasks and show how they can be completed easily and cheaply with human workers on Amazon Mechanical Turk (AMT). We illustrate our method with a novel analysis of Facebook posts by US Senators. While tailored, case-specific validation exercises may always be best, we aim to provide general tools to validate topics as measures.

Contents

1 Introduction

Many concepts in social science are not directly observable. If we wish to study democracy, culture, affect, or ideology, we need to develop a measurement strategy. How do we take observable data and use it to infer some unobserved trait of interest? Methods for handling this problem have varied markedly over time and across application areas. Congress scholars developed a number of tools for using roll-call behavior to infer member ideology, (e.g., Poole and Rosenthal, 2000; Poole, 2005; Clinton, Jackman and Rivers, 2004), survey researchers rely on tools such as Guttman scaling or factor analysis to infer latent traits such as 'tolerance' from survey responses (Gibson and Bingham, 1982), while network scholars use block structures, latent space models to infer communities from network connections (Goldberg et al., 2013; Minhas, Hoff and Ward, 2019) and pairwise competition models (Renshon and Spirling, 2015).

Recently scholars have increasingly turned towards text-as-data methods as a way to measure the contents of document sets, supplementing a long tradition of manual content analysis. Unsupervised probabilistic topic models have emerged as a particularly popular strategy for analysis since their introduction to political science in Quinn et al. (2010). Topic models are attractive because they both discover a set of themes in the text and annotate documents with these themes. Due to their ease-of-use and scalability, these models have quickly become a common way of measuring key explanatory and outcome variables. Recent examples include, inferring the degree to which candidates discuss particularistic policies (Catalinac, 2016), inferring how international organizations allocate regulatory effort (Pratt, 2018), and inferring the policy emphasis of media outlets (Barnes and Hicks, 2018).

However, utilizing topic models as a tool for measurement breaks sharply with their initial intended use case of language modeling and dimension reduction.[1] They were not designed to provide measures of pre-specified latent traits within texts. Instead, they are a clustering

[1] In the original article outlining LDA, Blei, Ng and Jordan (2003) primarily focus on information retrieval, document classification and collaborative filtering applications.

approach for exploring and usefully summarizing the structure of large document sets. In reaction, scholars in this tradition have emphasized the necessity of robust validation of model results before their use (Quinn et al., 2010; Denny and Spirling, 2018), with Grimmer and Stewart (2013) naming a key principle for text methods, "validate, validate, validate." But at the same time, the literature on topic model validation is extremely sparse. Early work was excruciatingly careful to validate the substantive meaning of the topics (e.g., Quinn et al., 2010; Grimmer, 2010), including carefully constructed *application-specific* criteria.[2] These bespoke methods for validation are effective but appear in the literature with increasing rarity.[3] To date, few general-purpose procedures have been proposed for validating the results of topic models in a transparent way.

This situation is unsatisfactory. On the one hand, we have cutting-edge models applied to immense collections of documents that previous generations could neither have collected nor analyzed. On the other hand, the validity of the findings from these studies rests entirely on our confidence in the authors' qualitative interpretations of the model outputs, most of which cannot be (or at least are not) reported. In many cases, no validations are reported to readers beyond providing the most probable words for each topic and statements implying that the resulting topics appeared coherent and meaningful under inspection. Going forward, if the discipline is to rely on topic modeling for rigorous scientific inference, we need a more rigorous set of standards for validation.

In this paper, we design and test a suite of validation exercises and provide tools to make them easy to run. Specifically, we extend the prior work of Chang et al. (2009) in computer science which proposes crowd-sourced tasks for model selection using human judgment. Chang et al. (2009) evaluates whether word sets learned by a topic model appear

[2]For example, Grimmer (2010) shows in an analysis of U.S. Senate press releases that senators talk more frequently about issues related to committees they chair. This is an intuitive evaluation that the topic model is able to detect something we are *ex ante* confident is true, but it does not straightforwardly generalize to other settings.

[3]We speculate that this is a consequence of the increased routinization of text analysis methods. With early innovations there is both a greater need to demonstrate validity and more space in which to do so. As tools become a regular part of the toolkit, they become less a focus of the article and thus don't permit the space or time to do extensive validation.

semantically related, but does not validate that the model captures the substantive quantity implied by the researchers' topic label. In addition to improving the original task design, we bridge this gap by proposing additional validation tasks for the concept labels themselves.

In the next section we discuss the use of topic models in social science, review the general problem of measurement validation, and consider standard approaches for topic model validation in both the social sciences and computer science. We then offer some principles for designing an off-the-shelf validation exercises using crowdsourced coding from non-experts. Next, we describe and test three tasks for evaluating semantic model fit and two tasks for label validation using topic models fit to Facebook posts from US Senators. Our results show that our tasks are easy for workers to complete and provide discrimination across models and potential labels. We also show that the evaluations are reliable and can be completed quickly and cheaply with human workers on Amazon's Mechanical Turk. We conclude with a discussion of the limitations of this kind of crowd-sourced validation design and the need for more research on topic validation.

In the end, we doubt that any complete, general-purpose validation exists, and our approach is not intended to be exhaustive and global. Instead, we hope that our method can be the first of many general-purpose approaches for more directly evaluating and communicating semantic model fit and label validity.

2 The use of topics and prior strategies of evaluation

As text-as-data methods have grown in popularity, researchers have increasingly used topic models to capture latent concepts.[4] This is not surprising since topic models are a method of data reduction that were specifically designed to make interpretation if not easy, then at least possible. In introducing latent Dirichlet allocation models, Blei, Ng and Jordan

[4]We do not offer a complete enumeration of articles employing this practice, but examples are not difficult to find (e.g., Al-Saggaf, 2016; Bagozzi, 2015; Bagozzi and Berliner, 2018; Bauer et al., 2017; Hayden et al., 2017; Lucas et al., 2015; DiMaggio, Nag and Blei, 2013; Roberts et al., 2014; Ryoo and Bendle, 2017; Nowlin, 2016; Terman, 2017; Velden, Schumacher and Vis, 2018).

(2003, p. 993) state, "The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks."

In the social sciences, researchers quickly uncovered the potential of topic models for measuring explanatory and dependent variables. The last decade has witnessed important work in all sub-fields where topic models measure latent traits including: senators' home styles in press releases (Grimmer, 2013), freedom of expression in human rights reports (Bagozzi and Berliner, 2018), religion in political discourse (Blaydes, Grimmer and McQueen, 2018), styles of radical rhetoric (Karell and Freedman, 2019) and more. In these works, the topic models identified useful latent traits anticipated by the authors, but in other cases, the models directed scholars towards new conceptualizations (Airoldi and Bischof, 2016; Grimmer and King, 2011; Velden, Schumacher and Vis, 2018).

This trend is promising in that this approach opens up important new lines of inquiry — especially in the context of the explosion of new textual data sources online. At the same time it is worrying in the sense that we may be running ahead of ourselves. Do these topics measure what they are supposed to measure? How would we know? We lack an established standard for affirming that a topic measures a particular concept.[5]

2.1 The problem of topic validation

The strength and weakness of topic models is that topics are simultaneously learned and assigned to documents. Thus, the researchers must, first, infer whether or not there are *any* coherent topics, second, place a conceptual label on those topics, and only *then* assess whether that concept is measured well. In this more open-ended process the potential for creative interpretation is vastly expanded — with all of the advantages and disadvantages

[5]These issues are complicated by the explicitly confirmatory, hypothesis-testing style of most quantitative work in the social sciences and the relative undervaluing of exploratory or descriptive work. That is, the problem isn't with using text analysis to measure properties of text but rather that in published work we erode the difference between confirming an *ex-ante* hypothesis and a data-driven discovery. See, for example, the discussion in Egami et al. (2018).

that brings. These concerns are magnified because people are extremely good at seeing patterns even where none exist (see e.g. Kalish, Griffiths and Lewandowsky, 2007).

Validity is a concern for any measurement in the social sciences. Yet, the potential problem for topic models is perhaps biting due to the *post hoc* nature of the process. A useful analogy here is with confirmatory (CFA) and exploratory factor analysis (EFA) in a survey setting. When assessing the validity of, for example, a newly proposed survey scale via CFA, we at least know that the survey items were developed with a specific concept in mind and can impose structure onto the model. Our problem is simply assessing whether the concept was measured well relative to our pre-determined target. To establish validity, researchers need to show that items load strongly on the underlying factor as expected and that the proposed model fits the data adequately.

Conceptually, CFA is closer to supervised learning, where the analyst designs a coding scheme (documenting it in a codebook). She then codes a large sample of documents according to that coding scheme, trains the learner, and then the algorithm annotates the remainder of the corpus. In this setting, the analyst has designed the measurement in order to fit with their particular argument and our primary validity concern is that the algorithm can annotate the documents with sufficiently high accuracy that the error is negligible.[6] Because we designed the measurement device, there is substantially more control over exactly what the latent categories are capturing. And if we believe in the originally proposed measurement scheme, the validity of the measurement is justified by high-predictive accuracy.

In contrast, topic models are more analogous to exploratory factor analysis (EFA), a method that has itself undergone significant criticism in the past.[7] Like EFA, topic models make specific assumptions about how the data is generated to estimate interpretable latent

[6]This glosses over a number of problems common to supervised learning in practice. Notably, we have the same concerns about discovery and estimation being done on the same document set if the coding scheme is developed with the same corpus where the coding is done (see e.g. Egami et al., 2018) for a formalization. In practice, it also isn't immediately clear what to do with non-trivial classification error although proposals based on resampling (Stewart and Zhukov, 2009; Benoit, Laver and Mikhaylov, 2009) and analytic corrections are available.

[7]For a particularly sharp, if perhaps over-broad, criticism of EFA, see Armstrong (1967).

traits from the observed data. However, the interpretation and adequacy of the various "factors" or "topics" are not justified by the model fitting process. Since the researcher has no ability to direct the model towards specific latent concepts, the quality of the model fit is at best indirectly related to the quality of the measure.

Indeed, the meaning of topics is not a function of the model itself but a function of the validation exercises that come *after* the model is fit (Grimmer and Stewart, 2013). The topics must be interpreted and linked to important latent concepts by examination of the documents and model outputs (e.g., high probability words). However, the qualitative nature of this interpretation makes it difficult for researchers to (a) complete this task in a replicable fashion or to (b) justify their decisions and conclusions in a way that can be assessed by readers. In the language of King, Keohane, and Verba (1994), the procedure is not public.

2.2 Validation practices for topic models

Since the procedure of turning a topic into a measure is not fully public, the need for validating the resulting numerical summary becomes even more important. Validating measures of latent concepts is a common task in the social sciences. Here, we borrow from extant work in the measurement literature to classify validation exercises into three inter-related categories (Adcock and Collier, 2001; Bollen, 2014):

1. Content validity: Does the measure include indicators we would expect it to include?
2. Criterion validity: Does the measure relate to external events in ways we would expect?
3. Construct validity: Does the measure capture what it claims to measure?
    - Convergent validity: If various measures theoretically should be related, is this expectation met?
    - Divergent validity: If various measures theoretically should *not* be related, is this expectation met?

We can use this measurement perspective to evaluate common practices of validation in the applied social scientific topic modeling literature.[8] In our reading, current practices are

[8]Note this isn't a novel perspective. In one of the earliest applications of topic modeling in the social

strongest with respect to content validity. Most research based on topic models provide word clouds or top words that allow us to assess (if imperfectly) whether or not expected words are correlated with the concept as we might expect. If a topic is supposed to represent the European debt crisis, it is comforting to see that top words for the topic include word stems like: "eurozone", "bank", "crisi", "currenc", and "greec" (Barnes and Hicks, 2018). However, it is important to recognize that this is far from perfect or complete. Many articles utilizing topics as measures provide *nothing* but word clouds or top words. However, this evidence is rarely clear-cut. So, for instance, we also see in the Euro/Debt crisis topic words like "year", "last", "auster", and "deficit". The first two words are at best ambiguous and the last two seem more associated with other topic labels (*Austerity Trade-Offs* and *Macro/Fiscal*) in the article (Barnes and Hicks, 2018).

Some scholars show that topic frequencies vary as expected in the face of external events (e.g. Quinn et al., 2010), a practice that be conceptualized as establishing criterion validity. Thus, Greene and Cross (2017) show that the topics prevalances in speeches by members of the European Parliament respond in expected ways to exogenous external events such as the collapse of the Lehman Brothers bank (p. 88). This is useful, but do researchers start with external events and confirm expected trends, or do they observe trends and then identify external events that offer plausible explanations? Moreover, given how context-dependent this process is, it is not available for all models or even all topics in the same model.

Finally, some scholars validate topics by using research assistants to categorize documents based on pre-specified coding schemes and comparing the results (e.g. Grimmer and Stewart, 2013) or developing new coding schemes to evaluate a newly discovered concept (Grimmer and King, 2011). This practice aims to test construct validity since measures intended to capture the same underlying concept should be highly correlated.

Validation for topic models in computer science focuses much more heavily on predictive accuracy, usually assessed through some measure of held-out log likelihood (Wallach et al.,

sciences, Quinn et al. (2010) frame the exercise through the lens of measurement theory similarly to how we have above.

2009). This provides a sense of whether or not the model is over-fitting but provides little direct evidence that is capturing something of interest for making a particular argument.

Pushing beyond predictive accuracy in computer science, Mimno et al. (2011) and Newman et al. (2010) introduced surrogate measures for coherent and interpretable topics based on point-wise mutual information. These measures reward topics which have high probability words which frequently occur together in documents. Mimno et al. (2011) show that this metric, which they call 'semantic coherence', correlate well with expert human judgments in a task analyzing grants submitted to the National Institute of Health. Roberts et al. (2014) extend this idea by suggesting that the model should jointly maximize exclusivity of words to individual topics as well as semantic coherence.

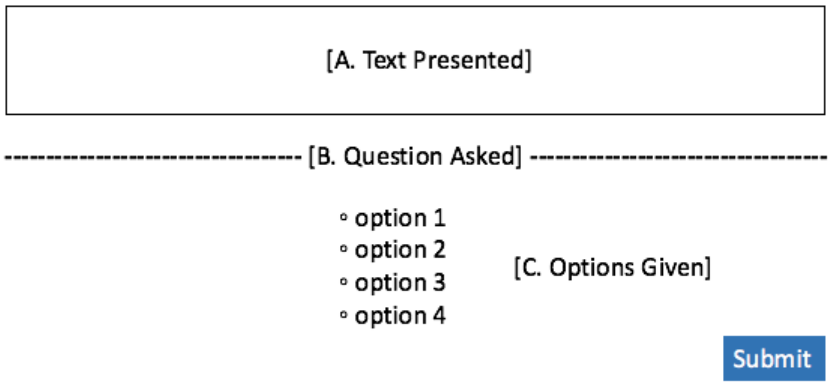
While maximizing model fit through the held-out log likelihood or surrogate criteria like semantic coherence may have correlated well with human judgment in the past, it is difficult to know how this criterion will generalize to future settings. Some researchers have moved to involve external coders to obtain indicators of topic model qualities at the aggregated level. Airoldi and Bischof (2016) and Newman et al. (2010), for example, have directly asked people to rate the coherence of the learned topics. Others have focused on finding better ways communicate the results of topic models graphically so readers might judge for themselves (Chuang, Manning and Heer, 2012; Chaney and Blei, 2012; Sievert and Shirley, 2014; Freeman et al., 2015). The challenge is that expert judgment does not scale and it is not obvious how to make use of these methods with non-expert judges.

2.3 Using the wisdom of the crowds

In an agenda-setting piece of work, Chang et al. (2009) introduced a set of crowd-sourced tasks for evaluating topic models.[9] The core idea is to transform the validation task into a set of two games which, if they can be completed with high accuracy, would imply the

[9]This has been followed up in Lau et al. (2011) and Lund et al. (2019). In political science, Lowe and Benoit (2013) used an innovative crowd-sourcing task design for assessing the validity of a scaling measure which can address other unsupervised text analysis designs like those based on Wordfish (Slapin and Proksch, 2008) or more advanced approaches (Spirling, 2012).

Figure 1: A Diagram for the Common Structure of Crowd-Sourced Validation Tasks



quality of the evaluated topic model.

The common structure for both tasks is shown in Figure 1. In each, a specific question is a presented to the coders (B) and they must choose from several options (C). For some tasks, additional information is provided above the question (A).

The first task, *Word Intrusion*, is designed to detect topics which are semantically cohesive in the sense that they identify a well-defined concept. Workers are presented with five words[10] and asked to identify the "most irrelevant" word — the intruder. Four words are chosen randomly from the highest probability words related to a topic and the remaining "intruder" is chosen from high probability words from a different topic (so it is not distinguishable from rarity alone).

The second task, what we call *Top 8 Word Set Intrusion* (T8WSI), is more complex.[11] The coder is presented with an actual document (or snippet from the document) and asked to identify a word set that does not belong. The coder must choose from four word sets where each set is the eight highest probability words for a topic, three of the word sets represent the high probability topics for the shown document, and one of the word sets is from a low probability topic (that is, a topic not associated with the the document).

In their study, Chang et al. (2009) find that the models with the best performance are

[10]Chang et al. (2009) presented tasks with six words, but we found that five words resulted in more reliable coding.

[11]Chang et al. (2009) call this Topic Intrusion but we have given it a more descriptive name.

not consistently those with the best model fit statistics (such as held-out log likelihood). This crowd-sourced approach to validation has become popular, now accumulating over 1,300 Google Scholar citations, and spawning attempts at automation (Lau, Newman and Baldwin, 2014). In the language of measurement models, we think of these tasks as testing convergent validity (words constructing the same concept are supposed to correlate) and discriminant validity (words constructing the different concepts are supposed to diverge) of the proposed latent concepts.

This crowd-sourced design is a way of showing that the topics make sense on their face but social scientists also need to establish whether the topics correspond to stated latent concepts. If we wish to use the outputs from topic models as a measure of some concept, it is not enough to show that "there is a *there* there." We must also show that the conceptual labels are themselves valid. In the next sections, therefore, we improve and extend these evaluations for measurement validation.

3 Designing and assessing an off-the-shelf evaluation

In this paper, we pursue the goal of designing an off-the-shelf evaluation design for topic models. We develop two classes of designs: one oriented towards model selection which extends the word intrusion task of Chang et al. (2009) to evaluate the semantic coherence of a given topic model, and a second oriented towards validating that a given topic corresponds to its label. Before we present our method, however, in Section 3.1 we offer some basic design principles and in Section 3.2 we describe the data that we use to assess the designs.

3.1 Principles

In developing the task, we want designs which are (1) generalizable, (2) discriminative, (3) easy-to-use, (4) reliable. In this paper we use the structural topic model (Roberts et al., 2013) to produce our topic models, but we want to find designs that are *generalizable* to

a wide variety of mixed-membership and single-membership topic models. The designs we present below should work for any mixed-membership topic model that uses a multinomial distribution over words to represent a topic and several will also work for single-membership models. The generalizable principle also reflects our desire to have evaluations that work in a variety of different substantive settings, with different size document collections, document lengths, and numbers of topics.

The tasks also need to be *discriminative*. Our task structures follow the general design of Chang et al. (2009) which have the form of games. These games need to be of medium difficulty because if they are too easy or hard, ceiling and floor effects (respectively) will hinder our ability to discriminate across different models. Indeed, our proposed extensions to the Chang et al. (2009) designs are motivated by a desire to make the tasks more possible to do and thus more discriminative. Better discrimination in turn leads to better information about model selection.

To be deployed in practice, the tasks need to be *easy-to-use*. All the tasks are designed to be completed by Mechanical Turk workers quickly. Along with the paper we will provide implementations using the SentimentIt platform of Carlson and Montgomery (2017) so that the tasks can be run quickly and cheaply.

Finally, we demonstrate that our tasks are *reliable*. Despite the fact that these tasks involve inherently subjective documents, we show across a variety of tasks and topics that the results are surprisingly stable under replication.

We divide the validation tasks into two parts. Model selection (Section 4) establishes that the model is semantically coherent in the sense of the prior work by Chang et al. (2009). A model should pass this task if the topics are immediately recognizable to a human evaluator as distinct concepts. Label validation (Section 5) goes further to assess whether or not a set of analyst-provided labels correspond with the contents of the model and the documents. We envision the model selection tasks as being most useful at early stages when making choices among competing models and the label validation tasks as helping to verify that the topic

model is measuring what it is intending to measure (as expressed by the label).

3.2 Empirical Illustration

To illustrate and assess our method, we rely on topic models fit to a novel dataset relevant to political science. The corpus comes from U.S. senators' Facebook pages from the 115th Congress.[12] We scraped every individual post from April 2018 back to when each page was initially created.[13] For text pre-processing, we removed all numbers, punctuation marks, and stopwords in the SMART stopword list. Additionally, we made a customized stopword list with state names (full or partial), state abbreviations, and the words such as "sen" and "senator" which are ubiquitous in senators' public pages. We converted all words to lower cases but did not stem them, making them easier to read for the online coders. Finally, we removed those posts about life events (e.g., "XX added a life event.") and those shorter than 10 words. We fit four structural topic models using the remaining 169,076 documents:

1. a model with 100 topics;
2. a model with 20 topics and adding the senators' names as a covariate;
3. a model with 10 topics and no covariates, and;
4. a 20 topic model (same as Model 2) limited to one EM iteration (which keeps the model from properly converging).

The first three models provide different feasible options for analyzing this corpus that we might want to consider in practice and we have no *ex ante* preference between them. Model 4, which is a topic model that has not been allowed to converge, is a baseline that we use to assess whether or not the tasks can clearly identify a flawed model. Note that even this model appears reasonable on first glance because of the initialization procedure in STM.[14]

[12]This was the Senate as composed in July 2017. Three Senators did not have public Facebook pages.

[13]The end date occurs when Facebook implemented changes to their Graph API. The earliest date of a post is September 2007.

[14]By the default the `stm` package (Roberts, Stewart and Tingley, 2015) uses a spectral method of moments (Arora et al., 2013) initialization strategy. Roberts, Stewart and Tingley (2016) show that it is a highly effective initialization strategy, but Arora et al. (2013) show that it has strong performance in its own right. Thus this is a relatively strong flawed baseline that could have proven challenging to detect.

Additional information about these topic model fits are provided in the rest of the paper and some additional information is listed in the Appendix.

4 Model selection using coherence evaluations

We considered three task structures for evaluating the semantic coherence of topic model fits. Topic models that perform well on these tasks are those where the topics pick out sharply defined topics which are distinctive from each other. The task structures are summarized in Table 1, where the column names correspond to the annotated slots in Figure 1. The first two, the Word Intrusion (WI) and the Top 8 Word Set Intrusion (T8WSI) tasks are slight alterations from the original methods from Chang et al. (2009) described above. Example tasks are shown in shown in Panel (a) and (b) of Figure 2.

4.1 Novel task structure

We also designed and tested one additional task to overcome what we perceived to be shortcomings from the Chang et al. (2009) task structures. Specifically, initial work indicated that coders found the WI and T8WSI tasks to be so difficult that the results could be uninformative. Further, we worried that the T8WSI results were too sensitive to the specific words included in the top 8 word sets for each topic, making the results somewhat arbitrary and again less informative.[15] Our aim was to produce a task structure that would make the task easier for coders to complete and would be less sensitive to the words that happen to fall in the top eight in order to increase our ability to disciminate between models.[16] This new task is summarized in the final rows of Table 1.[17]

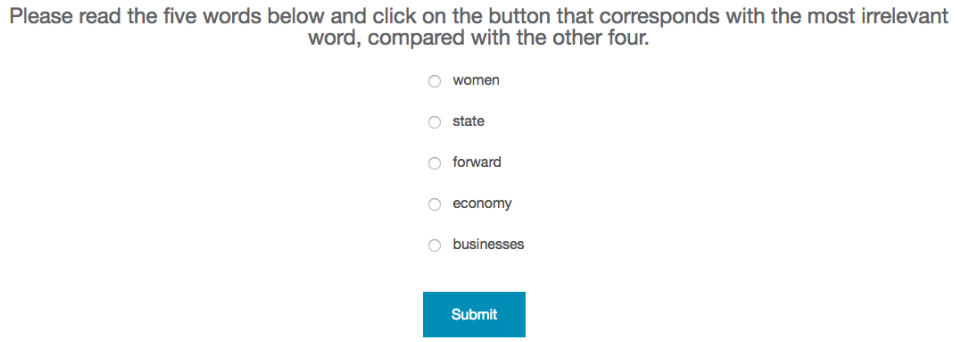
This *Random 4 Word Set Intrusion* (R4WSI) task chooses a random document from the

[15] As we will discuss in Section 6 this also makes the task particularly sensitive to the pre-processing choices making it difficult to adjudicate between different specifications (see also Denny and Spirling (2018)).

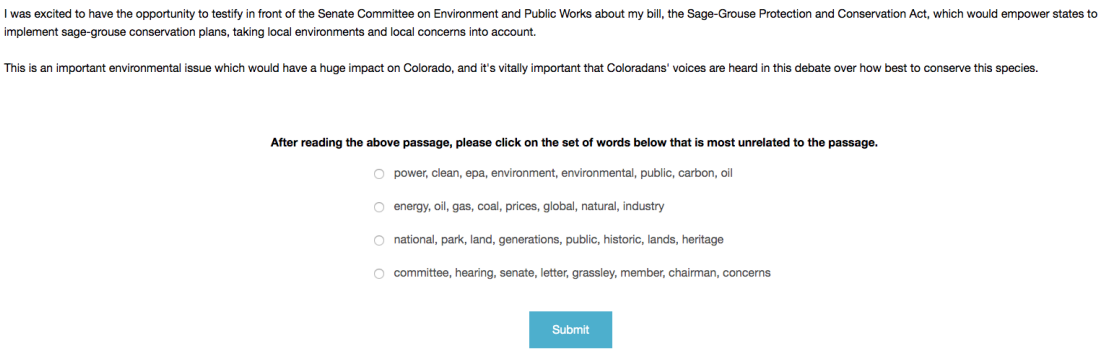
[16] The Appendix describes another variant that proved too easy in that we observed ceiling effects where nearly all plausible models scored near the top.

[17] In all tasks but T8WSI, we ensure that each topic from a given model is represented equally.

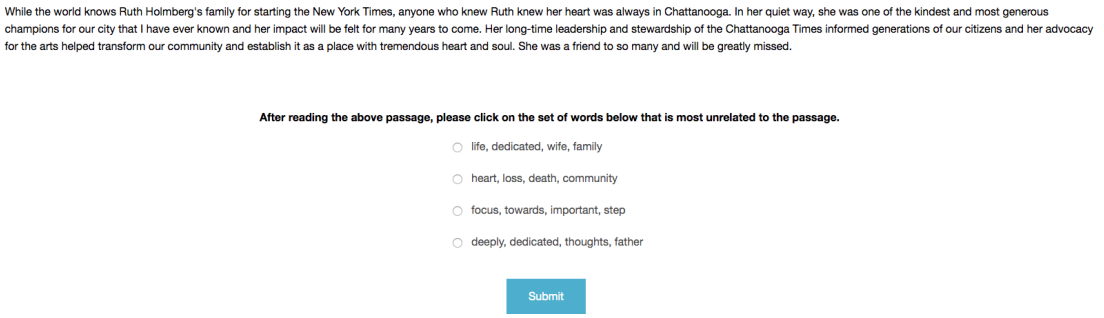
Figure 2: Example HITs for Model Selection Tasks



(a) Word intrusion



(b) Top 8 word set intrusion



(c) Random 4 word set intrusion

Table 1: Task Structures for Coherence Evaluations

	A. Text Presented	B. Question Asked	C. Options Given
Word intrusion	Nothing	Please read the five words below and click on the button that corresponds with the most irrelevant word, compared with the other four.	Five words, four randomly drawn from the highest-probability topic and the intruder from any other low-probability topics in the model
Top 8 word set intrusion	A document randomly selected from the corpus	After reading the above passage, please click on the set of words below that is most unrelated to the passage.	Four word sets (each with eight highest-probability words), three from the top three highest-probability topics and the intruder from any other low-probability topics in the model
Random 4 word set intrusion	A document randomly selected from the corpus	After reading the above passage, please click on the set of words below that is most unrelated to the passage.	Four word sets (each with four words), three are randomly drawn from the highest-probability topic and the intruder from any other low-probability topics in the model

corpus to display. Workers are then asked to identify the word set that does not belong. They must then choose from four different word sets. Structurally, this task is similar to the T8WSI task. However, there are two key differences. First, the three non-intruder word sets are all chosen from the same topic (the topic most associated with the document). Second, in the T8WSI task, the word sets presented to workers are always the eight top words for each topic. In this new task structure, we randomly chose four words from the top 20 words of the topic most associated with the document. An example of this general structure is shown in Panel (c) of Figure 2.

One advantage of the new R4WSI task is that it allows for coders to interact with more than just the eight most probable words for each topic. We found this to be helpful because for some topics the highest eight words were hard to interpret while their semantic relationship was perfectly clear when looking at 12 of the the top 20 words. Further, R4WSI builds on the intuition behind both of the other task structures to provide coders with more clues about the concepts the topics reflect. Coders can infer the semantic content of a topic by looking at all of the word sets and picking out the underlying concept that unites three out of four sets. In this way it mimics the logic of the WI task. Coders can also look at the document and see how the word sets relate, mimicking the logic of the T8WSI task. And, of course, they may do both at once, making the task easier to complete successfully. The end result, we argue, is coder decisions that are more informative about the quality of the underlying model leading to better discrimination.[18]

4.2 Results

We tested these three task structures using workers from Amazon's Mechanical Turk (AMT) from February, 2018 to April, 2019. To qualify to complete tasks, workers had to complete an online training module designed in Qualtrics.[19] These modules were designed to explain

[18] An additional advantage is that this structure is valid for topic models that assign documents to only one topic, making it more generalizable.

[19] Additional details are in the Appendix. An example training module can be seen here: `https://wuslpolysci.co1.qualtrics.com/jfe/form/SV_aVHM7GqHx7e41Nz`

the task, provide some background about the document set, and walk workers through examples to ensure they understood the goal. Worker qualifications were all handled using the SentimentIt.com API (Carlson and Montgomery, 2017). Workers were paid \$0.02/task for the WI task and \$0.04 for the other tasks since they required reading documents. These tasks were often completed in less than a day and always completed in less than three days. Costs for a single run of 500 tasks ranged from \$17-\$29, including fees to Amazon and replacements for low quality work.[20]

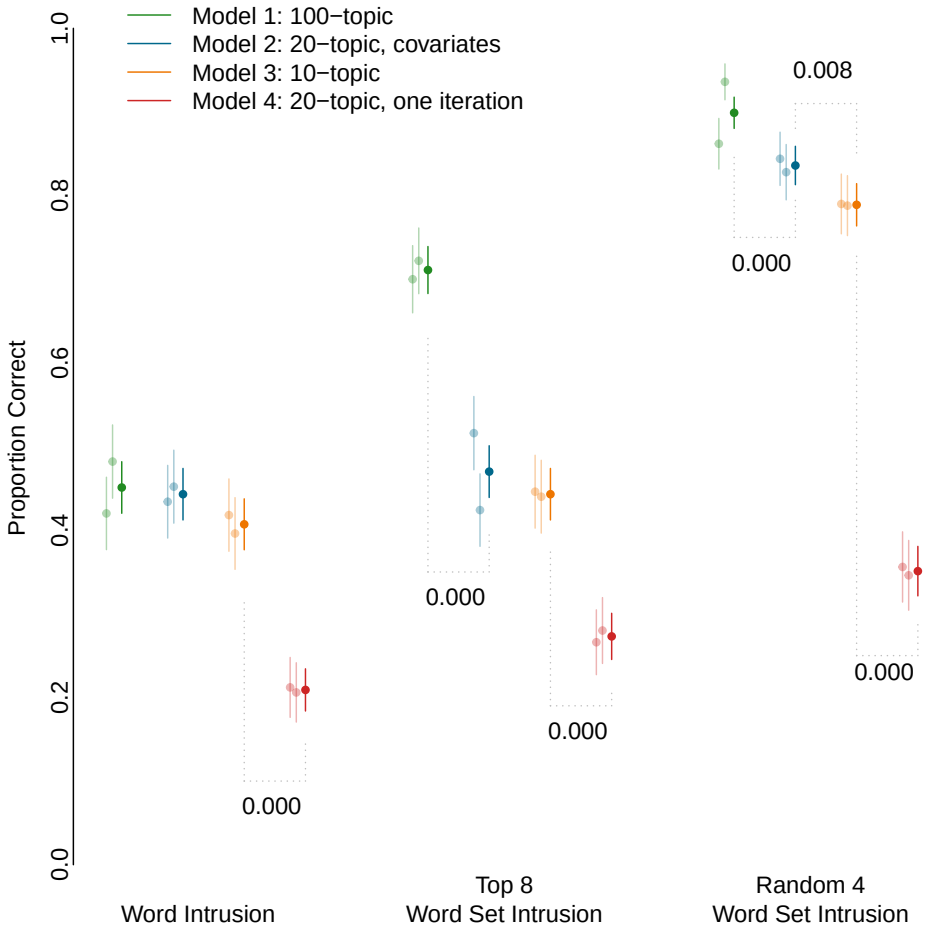
For each task structure we posted 500 tasks via the SentimentIt.com API for all four models. To assess the reliability of the workers and task structures, we then posted these *exact same tasks* again. Thus, in total workers completed 12,000 tasks. To monitor the quality of the work, we compared the answers from the two rounds for each task-model combination. Workers who agreed with others at a lower rate than average for that task-model by more than 7% were banned from all future tasks and their answers were replaced using new workers.[21]

There are three main conclusions that we draw from the results in Figure 3. First, all three task structures agreed on the rank ordering of the semantic coherence of the four models, although with different degrees of discrimination. All task structures easily identified the non-converged baseline model as the worst, which provides a check that the test has the ability to identify a model known to be a relatively poor fit. Of the remaining models, all three task structures identified the rank ordering as Model 1 as best (100 topics), Model 3 as worst, and Model 2 as intermediate. However, whether these differences were distinguishable from random noise at our given sample size depended on the task structure. None of these three are distinguishable based on the WI task. Models 2 and 3 are not distinguishable using the T8WSI task. Only the R4WSI clearly discriminates between all models.

[20]We note that these payment rates are actually somewhat higher than most tasks available on AMT for the amount of time they take. Further, the SentimentIt platform automatically approves HITS for all workers regardless of whether the data was deemed to be high quality. While it is not possible to ensure that workers are paid a sufficient minimum wage (different workers complete tasks at different rates), we encourage researchers to evaluate the time each task takes and offer ethical levels of compensation to workers.

[21]Across all task settings in the main text and appendix, 14% of tasks had to be replaced.

Figure 3: Results for Coherence Evaluations



Note: The 95% confidence intervals are presented. The two transparent bars represent two identical trails (500 HITs each). The solid bar represents the pooled result (1000 HITs). When two models yield significantly different results, the p-value is noted. (Significance tests are difference in proportions as calculated by the `prop.test` function in R.) Among the 12 pairs of identical trails, 2 pairs have yielded significantly different results, i.e., the two transparent bars are significantly different.

Second, there are stark differences in how difficult workers found these tasks and the quality of information gained. Setting aside our non-converged baseline, we can see from the proportion of tasks answered correctly, that the WI and T8WSI tasks are quite difficult for workers to correctly interpret and code. In seven out of eight cases, workers could not even complete 50% of these tasks correctly. This contrasts starkly with the R4WSI task where all task-model combinations result in more than 78% correct codings. We interpret this as evidence that it is easier for coders to glean the underlying semantic logic of the topic.

Third, although there is obviously some variation, the proportion estimates across runs appear to be fairly stable. Indeed, the estimated proportions are only statistically different in two of the 12 task-model combinations. Moreover, the relative rankings of the models are relatively consistent across repetitions. The only exceptions (e.g., the T8WSI where the 10-topic model scored higher than the 20-topic model on the second repetition) are in cases where the differences are not statistically reliable.[22]

To the extent that all three approaches resulted in similar rankings, we feel confident that this is a useful exercise for evaluating the semantic coherence of models. Further, the model clearly identified the non-converged model as inferior, consistent with *ex ante* expectations. However, the evidence suggests that some task structures are more useful than others. Based on these results, we would recommend that researchers focus primarily on the R4WSI and (to a lesser extent) the T8WSI tasks as providing the best combination in terms of the reliability of the estimates and the level of differentiation they offer across models.[23]

[22] Additional information about comparing the coding decisions across runs is provided in the Appendix. However, these should be interpreted with care since mere replication is not necessarily a helpful metric. Finding that coders chose the same *wrong* answer is not necessarily a good sign. In general, replication rates corresponded with the rates of correct response associated with the model/task. For instance, in tasks where workers got 95% of tasks correct, agreement rates can theoretically only go as low as 90%.

[23] We note that all of these task structures are evaluating whether or not the sets of high probability words are easily recognizable as concepts and distinguishable from each other. This actually has a strong connection to prior ways of evaluating topic models. The semantic coherence metric used by Mimno et al. (2011) is maximized when the high probability words in a topic co-occur primarily with each other (the model-based surrogate for a recognizable concept) and Roberts et al. (2014) add the requirement that these word sets have low overlap with other topics (ensuring that they are distinguishable). As such, we would expect these task validations to broadly agree with the semantic coherence and exclusivity ways of choosing topic models but to have more discriminating power (due to the added benefit of human contextual awareness). In future versions of this paper we intend to empirically explore this connection by comparing these tasks to various

5 Label Validation

Using our semantic coherence evaluations above, we selected Model 1 (estimated with 100 topics) as having the best fit. In practice a researcher would now need to place conceptual labels on the topics, essentially committing to what they are claiming the topics each measure. This process is inherently qualitative, requiring researchers to consider top words, read representative documents, and understand the general context of the document set. Once the researcher assigns a label to a topic, e.g. 'Income Inequality', the reader needs a way of assessing whether or not the topic is measuring the concept implied by the label. Unfortunately the best assessment, carefully reading many documents, is not easy to translate into an article and so in this section we try to design alternative checks on label quality. None of the previous methods (such as those in Chang et al. (2009)) use human generated labels and thus they are unable to assess this component of measurement validity.[24]

Ideally, our task will allow us to discriminate between good and bad labels. Unfortunately, to assess whether or not our task structure is working, we need an example of what "good" and "bad" labels are. That is, we need some sort of ground truth to see if our proposed methods work as expected. To address this, we asked two coders to, independently, label each of the 100 topics. Each coder carefully read the high-probability words and frequent & exclusive words (FREX) (Roberts, Stewart and Airoldi, 2016), as well as ten to fifteen representative documents per topic (Grimmer and Stewart, 2013). From the topics that both coders deemed as coherent, we picked ten economy-related topics where the coder labels were most consistent. The final labels and the high probability words for each are shown in Table 2.[25] We refer to these labels as the "Careful Coder" labels.

To provide a contrast, we also created our own labels based solely on the high probability

model statistic based approaches.

[24]Nielsen (2017) provides an excellent example of how to build confidence in a set of text-based measurements by triangulating many pieces of evidence. However, this is only possible as he is able to devote large chunks of a book to this task. Our goal here is to provide a piece of evidence that can be presented concisely.

[25]High probability documents for each of these ten topics are also provided in the Appendix.

and FREX word output from the STM package. These labels, which we refer to as "Word Coder" labels, are shown in the second column of Table 2. Note that conceptually these two label sets are not dramatically different (e.g., renewable energy vs. environment). However, we feel far more confident in the Careful Coder labels since they were independently identified by two experts who spent extensive time with the model and underlying data.

5.1 Novel task structures

We imagine that a researcher wishes to validate *only* the ten economy-related topics rather than the complete set of topic labels in the model. Thus, we focus only on the construct validity of these ten. However, the tasks structures we present here could easily be extended to include all labels of interest depending on the substantive question.[26]

To validate topic labels, we designed two alternative task structures summarized in Table 3: *Label Intrusion* and *Optimal Label*. Note that for this second task, we explored two alternative approaches for generating intruding labels (discussed below). These tasks, intended to provide evidence of construct validity, were again designed to be generalizable, discrimative, easy-to-use, and reliable.

First, we considered a *Label Intrusion* (LI) task where the coder is shown a text and asked to identify a label that does not apply. Three of the labels come from topics highly associated with the document (the top three topics for that document) and one is selected from the remaining seven labels. An example of this layout is shown in Figure 4. This structure intentionally mimics the word set intrusion design above.

The second task has the same basic layout, but the goal is not to find the intruder but rather the "best" label for the document.[27] The *Optimal Label - Economy* (OL-Econ) presents a document and four economy related labels. One label is for the the highest proba-

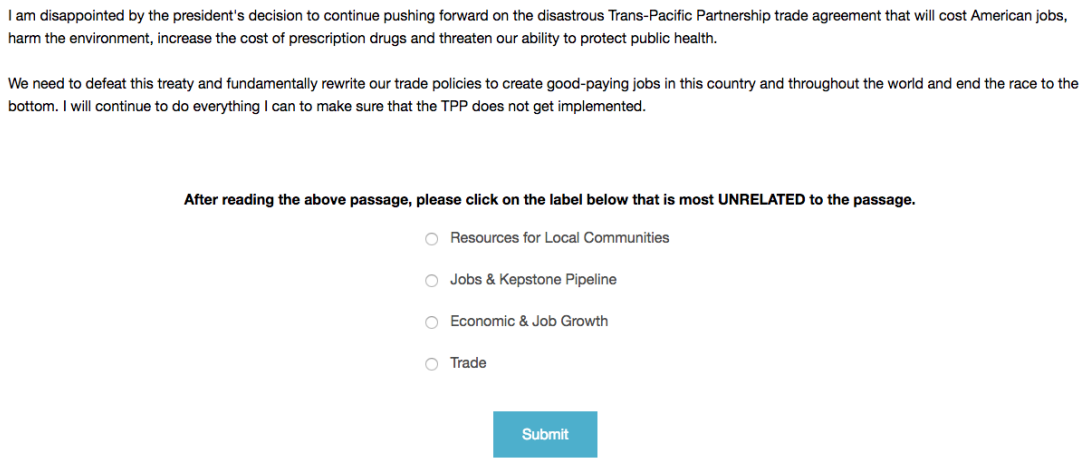
[26]The Appendix includes additional information about two additional task structures that were focused on word sets rather than documents. We found these task structures to be unsuitable for differentiating between strong and weak label sets.

[27]Specifically, the prompt is replaced with, "After reading the above passage, please click on the label below that BEST summarizes the passage."

Table 2: Labels Validated

Careful Coder	Word Coder	High Probability Words
Raising the Debt Ceiling	Deficits	spending, debt, money, dollars, fiscal, taxpayers, problem, crisis, solution, voting, trillion, reid, problems, limit, avoid, billions, spend, vote, harry, voters
Income Inequality	Unemployment	todays, times, percent, press, wall, street, unemployment, says, journal, poverty, according, conference, income, editorial, column, top, study, poor, americans, million
Equal Pay for Women	Middle-class Income	pay, workers, fair, class, families, middle, labor, paid, leave, women, equal, employees, hardworking, work, working, retirement, fairness, gap, wages, wage
Resources for Local Communities	Provide Resources	need, help, continue, communities, sure, ensure, provide, needs, resources, make, needed, possible, necessary, tools, pushing, reach, enable, smart, successfully, pushed
Jobs & Keystone Pipeline	Training Workers	jobs, job, create, build, opportunities, pipeline, companies, training, workers, keystone, workforce, expand, help, boost, invest, new, skills, thousands, century, bring
Trade	International Investment	including, trade, potential, international, largest, benefit, involved, china, supports, track, include, partnership, promote, worlds, faces, highlighting, experts, value, products, tremendous
Renewable Energy	Environment	energy, oil, gas, coal, prices, global, natural, industry, fuel, ben, production, wind, efficiency, renewable, resources, cardin, policy, reduce, fossil, solar
Economic & Job Growth	Manufacturing	economy, economic, growth, grow, development, technology, industry, manufacturing, innovation, creating, new, creation, growing, innovative, policies, sector, investment, private, plant, investments
Celebrate Small Businesses Messages	Local Economy	businesses, business, small, company, owners, employees, drive, entrepreneurs, inc, local, products, sales, craft, owner, store, beer, started, like, entrepreneurship, start
Student Loan Rates & Minimum Wage	Minimum Wage	increase, billion, cost, rates, costs, student, loan, rate, save, americans, interest, raise, lower, minimum, loans, higher, million, afford, average, struggling

Figure 4: Example HITs for Label Validation



bility topic and the other three labels are chosen randomly from the remaining nine economy topics. The *Optimal Label-ALL* (OL-All) is identical, except that the three randomly chosen labels come from the complete set of 83 valid topics.[28]

We were attracted to the "optimal" label task structure because it is similar to the validation exercises already common in the literature where research assistants are asked to divide documents into predefined categories to assess topic quality (Grimmer, 2013). Further we expected this task to generally be easier for coders to complete. Finally, this task structure has the advantage of being the most directly interpretable since it essentially asks coders to confirm or refute the conceptual labels assigned to the documents.

In addition, we anticipated that discriminating between the ten economic topics would be harder than discriminating between all 83 possible topics. That is, discriminating between conceptually similar topics (e.g., trade vs. income inequality) is understandably a "harder test" than discriminating between clearly distinct topics (e.g., trade vs. congratulating local sports teams). Which approach is better will depend on specific researcher needs, but we tested both variants of the OL task structure to confirm this intuition.

[28] Among the 100 topics, 17 were identified by the coders as too vague to be labeled.

Table 3: Task Structures for Label Validation

	A. Text Presented	B. Question Asked	C. Options Given
Label intrusion	An economy related document randomly selected from the corpus[a]	After reading the above passage, please click on the label below that is most UNRELATED to the passage.	Four labels, three for the top three predicted topics and one for another random economy-related topic
Optimal label (economy related labels)	An economy related document randomly selected from the corpus[b]	After reading the above passage, please click on the label below that BEST summarizes the passage.	Four labels, one for the presented topic and three for other random economy-related topics
Optimal label (all labels)	An economy related document randomly selected from the corpus	After reading the above passage, please click on the label below that BEST summarizes the passage.	Four labels, one for the presented topic and three for other random topics

[a]Top three predicted topics are all economy-related.
[b]Top one predicted topics is economy-related.

5.2 Results

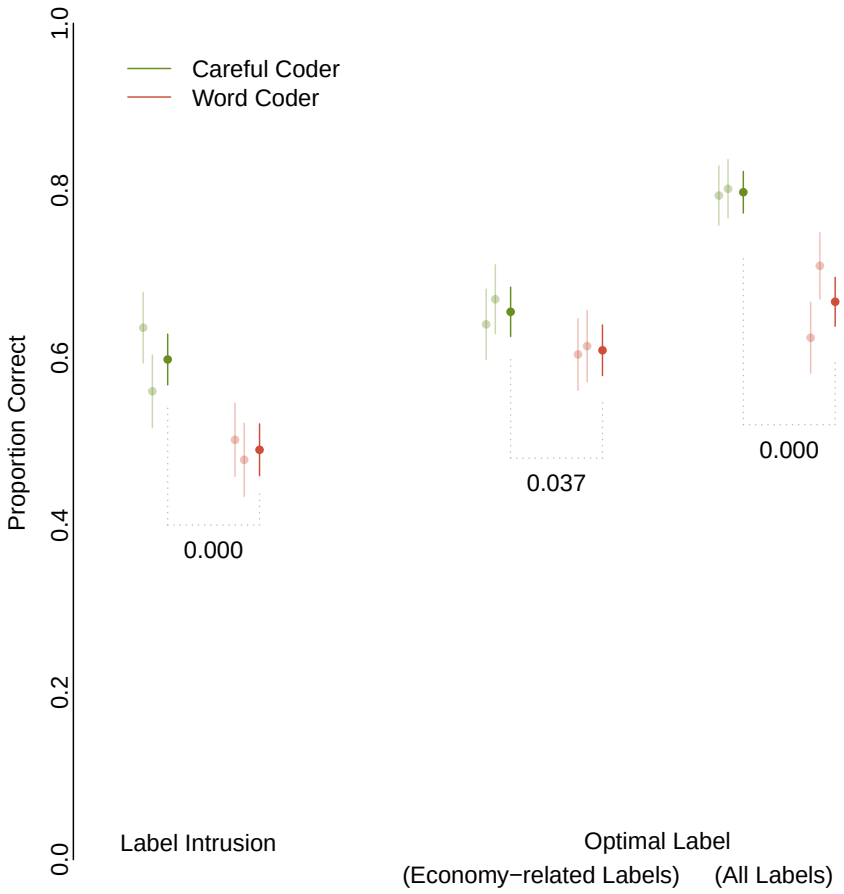
To test these task structures we followed the same basic procedures discussed above. For each task/coder combination we created 500 tasks that were coded by trained workers on AMT.[29] These were then repeated so that we could assess worker quality and replace work from low-quality workers. In total, workers completed 6,000 high quality HITS. The results are shown in Figure 5.

First, the results across runs again seem fairly reliable with rank orderings of the label sets mostly being indistinguishable across repetitions. Two out of six task/coder evaluations were statistically distinguishable across runs. This is again not perfect, but seems reasonable given uncontrolled difference across worker populations between repetitions. However, it does suggest that researches may wish to increase the number of HITs for label validation to improve reliability.

Second, the results are again consistent across task structures in identifying the Careful

[29]Workers were paid 0.02/HIT for the OLW task, 0.03/HIT for the WSI task, and 0.04 for the other tasks since they required reading documents.

Figure 5: Results for Label Validation



Note: The 95% confidence intervals are presented, where the two transparent bars represent two identical trails (500 HITs each) and the solid bar represents the pooled result (1000 HITs). P-values are again based on the pooled set of tasks based on a difference in proportions test. Among the 6 pairs of identical trails, 2 pairs have yielded significantly different results, i.e., the two transparent bars are significantly different.

Coder labels as being superior. However, there is again some heterogeneity in how difficult the tasks were for coders with the optimal label task generally having higher rates of correct coding, especially when intruders were drawn from the broader label set.[30]

In all, based on these results we would recommend that labels be evaluated using the Label Intrusion task and some form of the Optimal Label task. Whether it is better to draw intruder labels from the entire set of topics or a set of highly trusted and conceptually related topics will depend on the purposes of the researcher. The closely related topics represent a harder test, but this in turn may artificially lower the number of correct responses and make fine-grained distinctions more difficult.

6 Limitations and Future Directions

The task designs we offer are not a replacement for other validation exercises such as careful reading of the texts or demonstrating agreement with alternative supervised or hand-coding measures. Their central advantage is that they are low cost, reliable, and easy to communicate to a reader. The difficulty of alternate strategies, particularly careful reading of the texts, is that the way they appear in the final article amounts to a statement that says "trust me, I checked this." For any given application, custom designed solutions will likely be superior, but our tasks provide something that researchers can reach for in most circumstances.

As we stated at the outset our goal here is not to present the final word on this methodological question, but rather to begin a dialogue about how and when it is appropriate to make inferences about latent concepts from topic models. Towards that end, we conclude with a discussion of the limitations of this approach and specific areas where future research may make improvements. We begin this discussion by reconsidering our four principles of generalizability, discrimination, ease-of-use and reliability.

[30]See the Appendix for two additional task structures where this pattern did not hold. These tasks were based on word sets instead of documents, which significantly improved the perceived performance of the "Word Coder" labels.

Generalizability The validations we consider have several built-in assumptions that limit generalizability. First, the documents have to be *accessible* to the workers who are completing the tasks. This means that documents have to be in a language the workers can read,[31] short enough to be readable[32], and require little background to understand. The analyst must also be allowed to post them in a semi-public way.[33] For some these concerns will inhibit the use of our designs, but they would not preclude most designs that have been published in the literature thus far.

A more subtle consideration is that basing the representation on a fixed number (e.g. 20) of the most probable words can present challenges in certain model fits. Topic models can have very sparse distributions over the vocabulary, particularly with large number of topics, large vocabularies or when fit with collapsed Gibbs sampling. If the topic is too sparse, the later words in the top twenty might have close to zero probability, making the words essentially random. If stop words are not removed, the vocabulary can include high frequency words which are probable under all topics and thus also not informative of the topic.[34] This is another instance of text pre-processing decisions playing a consequential role in unsupervised learning (Denny and Spirling, 2018). In our setting, it is straightforward to apply these steps after the model has been run just for the purposes of the validation tasks.

Discrimination The tasks were largely able to discriminate between different options (either models or label sets) but the challenge throughout this paper is that we don't have access to a ground truth. That is, we can see that the tasks discriminate among options

[31]See Benoit et al. (2016) for a discussion of crowdsourced coding using international coders in multiple languages via the CrowdFlower system. AMT relies primarily on a US-based workforce since workers must have a US bank account to participate.

[32]We can use excerpts for long documents, but this similarly implies that a short summary can capture the gist of the document.

[33]See e.g. Romney, Stewart and Tingley (2015) on data access issues and intellectual property restrictions as limitations to transparency in statistical text analysis.

[34]There are also some concerns that may arise when not stemming or lemmatization as some word lists will be uninformative if they include many variants on the same word (e.g. `love`, `loves` and `loved`). This can also make the word set intrusion task trivially easy in some cases if multiple versions of the same word appear across different word sets (thus ruling them out as the intruder). In future versions of the paper, we will investigate the extent of these issues and find ways to address them.

but we have only circumstantial evidence that they discriminate *correctly*. The nature of measurement validity is that there is likely no way to actually do this kind of discrimination in the abstract, but we believe that these tasks are a useful part of a broader assessment. The coherence evaluations help to ensure that the topics convey a clear concept and are distinguishable from each other while the label validation exercises ensure that the researcher-assigned labels are at least somewhat accurate.[35]

Ease-of-Use The tasks are easy and relatively cheap to deploy using the SentimentIt platform. While not as simple as statistics which can be easily calculated from the model, they are about as straightforward to implement as a human task is likely to be. These evaluations were all completed in less than three days and sometimes in only a few hours. Further, while certainly not free, the 500 task runs we used here are fairly affordable with costs ranging between $17 and $29.

An important ease-of-use limitation that we have not yet addressed is the difficulty of interpreting the results in isolation. Above, we focus on the relative accuracy of the tasks across models or label sets in large part because it is not clear exactly what the accuracy levels themselves mean. For example, Model 1 scores 70% on the top 8 word set intrusion task. Is this good or bad? Is it comparable to performance on a completely different data set? Documents which involve more complex material or technical vocabularies may lead to poorer scores not because the models are worse, but simply because the task is inherently harder. Readers may naturally want to assess some cut-off heuristic where models or labels that score below a particular threshold are not acceptable for publication. We note that this would be problematic and would fall into many of the traps that bedevil the debate

[35]The tension arising from the lack of a ground truth is present in early parts of the literature as well. Chang et al. (2009) simply assert that their task designs select the most 'semantically meaningful' topic models, but do not have any empirical evidence for that claim. More problematically, it isn't clear what empirical evidence for this claim could look like. Probably the closest analog would be using the judgment of subject matter experts as in Grimmer and King (2011) (two teams of political scientists) and Mimno et al. (2011) (NIH staff members). This kind of evidence is very costly to collect and the experience in specific applications does not necessarily generalize. The design as presented rests on the argument that being able to pass these tests is a reasonable consequence of a semantically coherent model.

over p-values. Finding the right way to compare evidence across datasets remains an open challenge. Authors will need to provide readers with context for evaluating and interpreting these numbers, preferably by evaluating multiple models using multiple validation methods. With that said, random guessing would lead to a 25% correct rate for tasks with four choices. A minimal standard would be that coders should be able to substantially exceed this number. As we accumulate more evidence about such validation exercises, however, it may become possible to get a better sense of what an "adequate" score will be.

Reliability The task designs replicate across runs using the same population of Mechanical Turkers.[36] The SentimentIt platform helps to ensure that future iterations of the task would appear in the same way. With that said, researchers must take care to adequately train and screen workers, monitor data quality, and watch out for low quality workers who might flood researchers with low quality tasks.[37]

One final limitation is worth emphasizing. These tasks will not evaluate all properties we would need to see in a measurement. For example, many researchers use topics as outcomes in a regression. When estimating a conditional expectation, we want to know not only that the label is associated with the topic loadings but that they are proper interval scales (so that the mean is meaningful). These validation designs do nothing to assess these properties, and further work is needed to establish under what circumstances topic probabilities can be used as interval estimates of latent traits.

7 Conclusion

One reason the text as data movement is exciting is because it comes with a rapidly expanding evidence base in the social sciences (King, 2009). The conventional sources of evidence such

[36] Information on agreement rates for the same prompt across two different batches of workers are included in the appendix and are generally quite high.

[37] Anecdotally, we have found that worker quality is higher during normal work hours in US time zones.

as large surveys, summaries of voting records or economic data are giving way to individual study-specific datasets collected using text analysis techniques. This means that increasingly individual scholars are taking on the role of designing unique measurements for their study. Because we can't assume that these measurements will be used, examined and tested across many studies (e.g. as has happened with NOMINATE scores), it becomes imperative to develop best practices for regular diagnostic tests that researchers can run quickly on their own measurements and convey quickly.

We have offered a first step in this direction by improving upon the existing crowd-sourced designs of Chang et al. (2009) and extending them to create new designs that assess how well a label represents a corresponding topic. We tested these task structures using a novel topic model fit to Facebook posts by US Senators, and provided evidence that the method is reliable and allows for discrimination between models, based on semantic coherence, and labels, based on their conceptual appropriateness for specific documents. These kinds of crowd-sourced judgments allow us to capture human judgment and the linguistic knowledge that comes with it to examine our models, without experiencing the scale issues of relying on experts. The tasks are quick to complete and relatively inexpensive. They have relatively little cost in terms of the analyst's time when implemented through the SentimentIt system. These designs will offer a fruitful space for innovation, however, and our collective work on validating topics as measures is just getting started.

References

Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.

Airoldi, Edoardo M. and Jonathan M. Bischof. 2016. "Improving and Evaluating Topic Models and Other Models of Text." *Journal of the American Statistical Association* 111(516):1381–1403.

Al-Saggaf, Yeslam. 2016. "Understanding Online Radicalisation Using Data Science." *International Journal of Cyber Warfare and Terrorism* 6(4):13–27.

Armstrong, J. Scott. 1967. "Derivation of Theory by Means of Factor Analysis or Tom Swift and His Electric Factor Analysis Machine." *The American Statistician* 21(5):17–21.

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu and Michael Zhu. 2013. A Practical Algorithm for Topic Modeling with Provable Guarantees. In *International Conference on Machine Learning*. pp. 280–288.

Bagozzi, Benjamin E. 2015. "The Multifaceted Nature of Global Climate Change Negotiations." *Review of International Organizations* 10(4):439–464.

Bagozzi, Benjamin E. and Daniel Berliner. 2018. "The Politics of Scrutiny in Human Rights Monitoring: Evidence from Structural Topic Models of US State Department Human Rights Reports." *Political Science Research and Methods* 6(4):661–677. WOS:000446846200004.

Barnes, Lucy and Timothy Hicks. 2018. "Making Austerity Popular: The Media and Mass Attitudes toward Fiscal Policy." *American Journal of Political Science* 62(2):340–354.

Bauer, Paul C., Pablo Barberá, Kathrin Ackermann and Aaron Venetz. 2017. "Is the Left-Right Scale a Valid Measure of Ideology?: Individual-Level Variation in Associations with "Left" and "Right" and Left-Right Self-Placement." *Political Behavior* 39(3):553–583.

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

Benoit, Kenneth, Michael Laver and Slava Mikhaylov. 2009. "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53(2):495–513.

Blaydes, Lisa, Justin Grimmer and Alison McQueen. 2018. "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds." *Journal of Politics* 80(4):1150–1167.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan):993–1022.

Bollen, Kenneth A. 2014. Measurement Models: The Relation between Latent and Observed Variables. In *Structural Equations with Latent Variables*. John Wiley & Sons, Ltd pp. 179–225.

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

Catalinac, Amy. 2016. "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections." *Journal of Politics* 78(1):1–18.

Chaney, Allison June-Barlow and David M. Blei. 2012. Visualizing Topic Models. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*. pp. 288–296.

Chuang, Jason, Christopher D. Manning and Jeffrey Heer. 2012. Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM pp. 74–77.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.

Denny, Matthew J. and Arthur Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It." *Political Analysis* 26(2):168–189.

DiMaggio, Paul, Manish Nag and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of US Government Arts Funding." *Poetics* 41(6):570–606.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts and Brandon M. Stewart. 2018. "How to Make Causal Inferences Using Texts." *arXiv preprint arXiv:1802.02163* .

Freeman, M. K., J. Chuang, M. E. Roberts, B. M. Stewart and D. Tingley. 2015. "stmBrowser: Structural Topic Model Browser.".

Gibson, James L. and Richard D. Bingham. 1982. "On the Conceptualization and Measurement of Political Tolerance." *The American Political Science Review* 76(3):603–620.

Goldberg, Mitchell D., Heather Kilcoyne, Harry Cikanek and Ajay Mehta. 2013. "Joint Polar Satellite System: The United States next Generation Civilian Polar-Orbiting Environmental Satellite System." *Journal of Geophysical Research: Atmospheres* 118(24):13–463.

Greene, Derek and James P. Cross. 2017. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." *Political Analysis* 25(1):77–94.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1):1–35.

Grimmer, Justin. 2013. "Appropriators Not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science* 57(3):624–642.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297.

Grimmer, Justin and Gary King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization." *Proceedings of the National Academy of Sciences* 108(7):2643–2650.

Hayden, Jessica M., Matthew J. Geras, Nathan M. Gerth and Michael H. Crespin. 2017. "Land, Wood, Water, and Space: Senator Robert S. Kerr, Congress, and Selling the Space Race to the American Public." *Social Science Quarterly* 98(4):1189–1203.

Kalish, Michael L., Thomas L. Griffiths and Stephan Lewandowsky. 2007. "Iterated Learning: Intergenerational Knowledge Transmission Reveals Inductive Biases." *Psychonomic Bulletin & Review* 14(2):288–294.

Karell, Daniel and Michael Raphael Freedman. 2019. "Rhetorics of Radicalism." *American Sociological Review* .

King, Gary. 2009. The Changing Evidence Base of Social Science Research. Routedge pp. 91–93.

Lau, Jey Han, David Newman and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 530–539.

Lau, Jey Han, Karl Grieser, David Newman and Timothy Baldwin. 2011. Automatic Labelling of Topic Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics pp. 1536–1545.

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.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23(2):254–277.

Lund, Jeffrey, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber and Kevin Seppi. 2019. "Automatic Evaluation of Local Topic Quality.".

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics pp. 262–272.

Minhas, Shahryar, Peter D. Hoff and Michael D. Ward. 2019. "Inferential Approaches for Network Analysis: AMEN for Latent Factor Models." *Political Analysis* 27(2):208–222.

Newman, David, Jey Han Lau, Karl Grieser and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics pp. 100–108.

Nielsen, Richard A. 2017. *Deadly Clerics: Blocked Ambition and the Paths to Jihad*. Cambridge University Press.

Nowlin, Matthew C. 2016. "Modeling Issue Definitions Using Quantitative Text Analysis." *Policy Studies Journal* 44(3):309–331.

Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting*. Cambridge University Press.

Poole, Keith T. and Howard Rosenthal. 2000. *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press on Demand.

Pratt, Tyler. 2018. "Deference and Hierarchy in International Regime Complexes." *International Organization* 72(3):561–590.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1):209–228.

Renshon, Jonathan and Arthur Spirling. 2015. "Modeling "effectiveness" in international relations." *Journal of conflict resolution* 59(2):207–238.

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2015. "STM: R Package for Structural Topic Models.".

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2016. "Navigating the Local Modes of Big Data." *Computational Social Science* 51.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley and Edoardo M. Airoldi. 2013. The Structural Topic Model and Applied Social Science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. Harrahs and Harveys, Lake Tahoe pp. 1–20.

Roberts, Margaret E., Brandon M. Stewart and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111(515):988–1003.

Romney, David, B. Stewart and Dustin Tingley. 2015. "Plain Text? Transparency in Computer-Assisted Text Analysis." *Qualitative & Multi-Method Research* .

Ryoo, Joseph and Neil Bendle. 2017. "Understanding the Social Media Strategies of U.S. Primary Candidates." *Journal of Political Marketing* 16(3-4):244–266.

Sievert, Carson and Kenneth Shirley. 2014. LDAvis: A Method for Visualizing and Interpreting Topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. pp. 63–70.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

Spirling, Arthur. 2012. "US Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911." *American Journal of Political Science* 56(1):84–97.

Stewart, Brandon M. and Yuri M. Zhukov. 2009. "Use of Force and Civil–Military Relations in Russia: An Automated Content Analysis." *Small Wars & Insurgencies* 20(2):319–343.

Terman, Rochelle. 2017. "Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage." *International Studies Quarterly* 61(3):489–502.

Velden, Mariken Van Der, Gijs Schumacher and Barbara Vis. 2018. "Living in the Past or Living in the Future? Analyzing Parties' Platform Change In Between Elections,The Netherlands 1997–2014." *Political Communication* 35(3):393–412.

Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov and David Mimno. 2009. Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM pp. 1105–1112.