Topics, Concepts, and Measurement:
A Crowdsourced Procedure 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

August 29, 2021

Abstract

Topic models, as developed in computer science, are effective tools for exploring and summarizing large document collections. When applied in social science research, however, they are commonly used for measurement, a task that requires careful validation to ensure that the model outputs actually capture the desired concept of interest. In this paper, we review current practices for topic validation in the field and show that extensive model validation is increasingly rare, or at least not systematically reported in papers and appendices. To supplement current practices, we refine an existing crowd-sourcing method for validating topic quality (Chang et al., 2009) and go on to create new procedures for validating conceptual labels provided by the researcher. We illustrate our method with an analysis of Facebook posts by U.S. Senators and provide software and guidance for researchers wishing to validate their own topic models. While tailored, case-specific validation exercises will always be best, we aim to improve standard practices by providing a general-purpose tool to validate topics as measures.

[*] Corresponding author, `luwei.ying@wustl.edu`

1 Introduction

Many core concepts in the social sciences are not directly observable. To study democracy, culture, or ideology, we must first build a measure and make inferences about unobservable concepts from observed data. Methods for handling this problem have varied markedly over time and across fields. Congress scholars developed multiple tools to infer member ideology from roll-call behavior (e.g., Poole and Rosenthal, 1985; Clinton et al., 2004) while survey researchers rely on tools such as factor analysis to infer traits such as 'tolerance' from survey responses (e.g., Gibson and Bingham, 1982).

Recently, social scientists have turned towards text-as-data methods as a way to derive measures from written text, supplementing a long tradition of manual content analysis with computer-assisted techniques. Unsupervised probabilistic topic models have emerged as a particularly popular strategy for analysis since their introduction to political science by Quinn et al. (2010). TMs 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, TMs have become a standard method for measuring concepts in text.

Yet, TMs were not originally designed for the measurement use-case. Blei et al. (2003) present latent Dirichlet Allocation (LDA) as a tool for information retrieval, document classification, and collaborative filtering. Given this shift in focus, the scholars who introduced the "topics as measures" tradition to political science emphasized the necessity of robust validation (Quinn et al., 2010; Grimmer, 2010), with Grimmer and Stewart (2013) naming a key principle for text methods, "validate, validate, validate." Early work was excruciatingly careful to validate the substantive meaning of the topics through carefully constructed *application-specific* criteria and bespoke evaluations. Yet as we have routinized TMs, validation has received less emphasis and less space on the page. In our review of recent practice in top political science journals below, we show that over half of articles using TMs report only a list of words associated with the topic and only a handful of articles report fit statistics.[1]

[1] More details of the review are in Section 2.2. This includes only validations of meaning that authors

Meanwhile extensive, application-specific validations are more rare.

This *status quo* presents a challenge. On the one hand, we have the ability to measure important concepts using immense collections of documents that previous generations could neither have collected nor analyzed. On the other hand, the value of these findings increasingly rests entirely on our confidence in the authors' qualitative interpretations, which cannot be succinctly reported.[2] The most important step for addressing this challenge is renewed attention to validation, but by their very nature customized, application-specific validations are difficult to formalize and routinize.

In this article, we take a different approach. We design and test a suite of validation exercises designed to capture human judgment which can be used in a wide range of settings. Our procedure refines a prior crowdsourcing method for validating *topic quality* (Chang et al., 2009) and presents a new design for validating the researcher-assigned *topic labels*. We provide software tools and practical guidance that make all our validations straightforward to run. Crucially, our goal is not to *supplant* bespoke validation exercises but to *supplement* them. While no single method can validate TMs for all settings, our aim is to re-emphasize the importance of validation and begin a dialogue between methodologists and applied researchers on improving best practices for validating topics as measures.

In the next section, we review how TMs are validated in the social sciences, drawing on a new survey of articles in top political science journals. Section 3 lays out our principles in designing new crowdsourced tasks and introduces our running example. We then outline and evaluate our designs for validating topic coherence (Section 4) and label quality (Section 5). We conclude by discussing limitations of our designs and future directions for what we hope

report in main papers or appendices and excludes authors' statements about reading the documents. We focus on validations reported to the reader, although authors likely conduct more extensive validation exercises on their own (and indeed we see some evidence of this in replication archives).

[2] In some cases, e.g. Nielsen (2020), extensive replication archives are available which contain all the documents necessary for readers to explore the work themselves. Barberá et al. (2019) provides a custom website which shows all the topics over time and with sample documents and illustrates how this can be used to check against external events for one of the topics. Both of these approaches are fantastic and allow the interested reader to deeply explore the validity of the measurement. However, we argue that there is also a need for a simple measure that provides an approximate summary of model quality that does not require extensive reader expertise or investment of time.

is only the first of many new methods for validating topics as measures.

2 How topic models are used and validated

In the social sciences, researchers quickly uncovered the potential of TMs for measuring key concepts in textual data. Political Science in particular has witnessed important work in all sub-fields where TMs 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 et al., 2018), styles of radical rhetoric (Karell and Freedman, 2019) and more. In other works, the models are used to explore new conceptualizations which may in turn be measured using a different sample or a different approach (Grimmer and King, 2011; Pan and Chen, 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 the move towards measurement is worrying if we are 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.[3] In this section, we describe why TM validation is an essential task. We then briefly characterize early approaches to validation and conclude with a review of recent empirical practices.

2.1 The importance of topic validation

The strength and weakness of TMs 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 that brings. The

[3] By "standard" we mean that the scholarly community has not reached anything like a consensus as to whether and how validations should be reported to readers and reviewers.

interpretation and adequacy of the topics are not justified by the model fitting process—those motivating assumptions were simply *conveniences* not structural assumptions about the world to which we are committed (Grimmer et al., 2021). Instead, our confidence in the topics as measures comes from the validation that comes *after* the model is fit (Grimmer and Stewart, 2013). This places a heavy burden on the validation exercises because they provide our primary way of assessing whether the topics measure the concept well relative to an externally determined definition.

A further complication is that TMs are typically fit, validated, and analyzed in a single manuscript. By contrast, NOMINATE was extensively validated before widespread adoption (e.g., Poole and Rosenthal, 1985) and subsequently used in thousands of studies. Novel psychological batteries are often reported in a stand-alone publications (e.g., Cacioppo and Petty, 1982; Pratto et al., 1994), or at the very least subjected to common reporting standards. In other words, the common practice of one-time-use TMs means that research teams are typically going about this process alone.

The inherent difficulty of validation is critical for how readers and researchers alike understand downstream inferences. Subtle differences in topic meanings can matter, and outputs like the most probable words under a topic are, in our experience, rarely unambiguous. Whether a topic relates to "reproductive rights" or "healthcare," for instance, can be difficult for a reader to ascertain based these kinds of model outputs.[4] Yet showing that, for instance, female legislators are more likely to discuss "healthcare" has very different substantive implications than finding they are more likely to discuss "reproductive rights."[5]

Understanding when validation is needed is complicated by the ostensibly confirmatory, hypothesis-testing style of most quantitative work in the social sciences. Published work often

[4] In our example below the top ten words for the "healthcare/reproductive rights" topic are: health, care, access, affordable, services, coverage, healthcare, medicaid, mental, medicare.

[5] One consequence of this ambiguity is related to "researcher degrees of freedom" in both labeling and model fitting. On the modeling side this may include pre-processing steps (Denny and Spirling, 2018), selection of solutions across initializations (Roberts et al., 2016), hyperparameter selection, and more. This flexibility may inadvertently lead researchers down "the garden of forking paths" towards theory confirmation (John et al., 2012; Gelman and Loken, 2013).

erodes the difference between confirming an *ex-ante* hypothesis and a data-driven discovery (Egami et al., 2018)—settings that require different kinds of validation. Of course, this tension is not unique to TMs and, in fact, echoes debates about exploratory and confirmatory factor analysis of a previous era (see Armstrong, 1967).

Early approaches to validation. The early TM literature in political science followed a common pattern for validation (Quinn et al., 2010; Grimmer, 2010; Grimmer and Stewart, 2013). First, estimate a variety of models, examine word lists, and carefully read documents which are highly associated with each topic. Then, in combination with theory, evaluate predictive validity of topics by checking that topics are responsive to external events, convergent validity by showing that it aligns with other measures, and hypothesis validity by showing that it can useful test theoretically interesting hypotheses. These latter steps are what we call *bespoke validations* and are highly-specific to the study under consideration. For example, Grimmer (2010) shows in an analysis of US Senate press releases that senators talk more frequently about issues related to committees they chair. This is an intuitive evaluation that the model is detecting something we are *ex ante* confident is true, but that expectation is specific to this setting. In short, this approach is heavy on “shoe-leather” effort and involves a great deal of customization—but it is also the gold standard of validation.

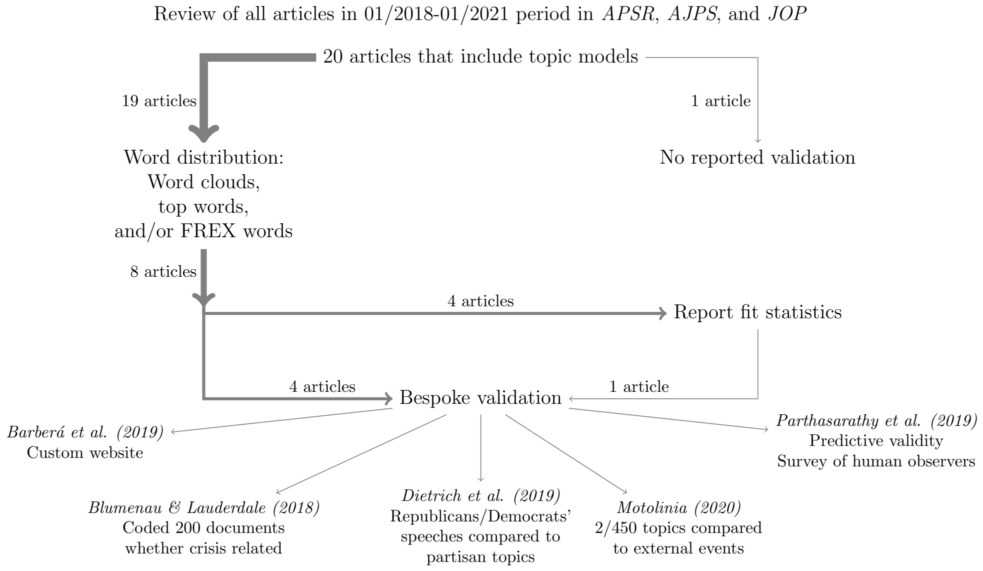
2.2 A review of recent practices

How are TMs validated in more recent articles published in top journals? To assess current practices in the field, we identified all articles published in the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics* from January 1, 2018 to January 2021[6] that included the phrase “topic model.” Out of the 20 articles, the topic serves as an outcome variable in 13 and as a predictor in 8.[7]

[6] This includes all articles published online at the time of our search.

[7] In some cases, the TMs are not central to the core analysis (e.g., Rozenas and Stukal, 2019, which uses them as a validation) or measurement was not a primary goal (e.g., Blumenau and Lauderdale, 2018, which uses them for prediction).

Figure 1: Survey of practices in topic model analysis in top political science journals


Review of all articles in 01/2018-01/2021 period in *APSR*, *AJPS*, and *JOP*
20 articles that include topic models
19 articles
1 article
Word distribution: Word clouds, top words, and/or FREX words
No reported validation
8 articles
4 articles
Report fit statistics
4 articles
Bespoke validation
1 article
Barberá et al. (2019) Custom website
Parthasarathy et al. (2019) Predictive validity Survey of human observers
Blumenau & Lauderdale (2018) Coded 200 documents whether crisis related
Dietrich et al. (2019) Republicans/Democrats' speeches compared to partisan topics
Motolinia (2020) 2/450 topics compared to external events

We created three dichotomous variables reflecting the most common classes of validation strategies reported: topic-specific word lists, fit statistics, and bespoke validation of individual topic meanings.[8] Notably, we have omitted "authors reading the text" which—while an essential form of validation—cannot be clearly demonstrated to the reader and thus is not fully public in the sense of King et al. (1994).[9] The results of our analysis are summarized in Figure 1. We did not explicitly exclude articles who used TMs for non-measurement purposes because we found it too difficult to reliably assess and thus the 20 articles should be taken as the size of our sample, but not necessarily the number of articles which would ideally have used validations of meaning.

Topic-specific word lists. The most common form of validation—used in 19 of the 20 articles—is presenting word lists for at least some subset of topics.[10] These could be either the most probable words in the topic under the model or alternative criteria such as frequency

[8] Two authors coded each article independently and all three authors discussed cases where there was disagreement to arrive at a consensus. The full set of articles and our codings are shown in Appendix SI1.

[9] "If the method and logic of a researcher's observations and inferences are left implicitly, the scholarly community has no way of judging the validity of what was done" (King et al., 1994, p 8).

[10] Even the 20th article included a list in the replication materials although this was not mentioned in the appendix.

and exclusivity (FREX) (Roberts et al., 2016) and are sometimes reported in word clouds. In practice such lists help to establish content validity (e.g., does the measure include indicators we would expect it to include?; does the measure exclude indicators that are extraneous or ambiguous?).[11] The word lists allow the reader to assess (if imperfectly) whether or not words are correlated with the assigned topic label as they might expect. If a topic is supposed to represent the European debt crisis, for instance, 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).

11 of those 19 articles provide *only* word lists. These are often short and rarely provide numerical information about the probability under the model. While lists can be intuitive, they are rarely unambiguous. In the European debt crisis topic above we also see "year", "last", "auster", and "deficit". The first two words are 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). Stripped of their context, word lists are hard to assess making it hard for the reader to make their own judgment.

Fit statistics. Beyond word sets, 4 of the 20 articles also reported fit statistics such as held-out log likelihood (Wallach et al., 2009) or surrogate scores such as "semantic coherence" (Mimno et al., 2011; Roberts et al., 2014).[12] This provides a sense of whether or not the model is over-fitting, and some previous research shows that surrogates correlate with human judgements.

[11] For a more comprehensive discussion of measurement validation, see Adcock and Collier (2001).

[12] Traditional held-out log likelihood statistics provide a measure of fit to the data under the model. Mimno et al. (2011) introduce the use of a pointwise mutual information metric which they call *semantic coherence*. This metric checks how often the most probable words in a topic are to actually appear together in the same document. They show that this evaluation metric correlates well with expert human annotators in an analysis of grants from the U.S. National Institutes of Health. Generally speaking we refer to measures like this as *surrogate scores* because instead of measuring model fit they measure what we hope is a surrogate for human judgment.

Bespoke approaches. Five articles reported additional validations of topic meaning designed especially for their case to establish construct validity (does the measure relate to the claimed concept?). Blumenau and Lauderdale (2018) coded 200 documents as to whether the document was related to the Euro crisis with the goal of finding topics that maximized predictions of crisis-related votes. In a supplemental analysis, Dietrich et al. (2019) qualitatively identify partisan topics and show Republicans/Democrats speak more about their topics. Motolinia (2020) fit a TM with 450 topics and reported validations for two relative theoretical expectations (see their Figure 2). Barberá et al. (2019) provided considerable information about topics including a custom website[13] showing high frequency words and example documents, and reported a validation against external events for one of the 46 topics. Arguably, the most thorough reported validation was in Parthasarathy et al. (2019), which validates topics against theoretical predictions and survey responses from human observers of public deliberations in India. What counts as a bespoke validation is unavoidably subjective, but we emphasize here that we are considering bespoke validation of individual topic meanings which excludes many other valuable analyses.[14]

Summary of findings. We emphasize that our analysis is limited to validations *reported* to readers. In many cases, the topics were validated in additional ways that could not be (or at least *were* not) reported. For instance, Blaydes et al. (2018, p.1155) write, "Our research team also evaluated the model qualitatively . . . , selecting the specification and final model that provided the most substantive clarity." This is an essential part of the process, but isn't easily visible to the reader. The reader can see the reported high probability words (Table 1) and qualitative descriptions of topics (Appendix C). Careful qualitative evaluation is arguably the most important validation, but it is not easily communicated.

[13] http://pablobarbera.com/congress-lda/

[14] For example, Nielsen (2020) provides extensive evidence that results are robust to TMs of different sizes, Roberts et al. (2020) provides a variety of balancing checks for their text matching procedure, and Pan and Chen (2018) uses TMs for exploration and a supervised learning approach for the eventual inference. None of these are counted as bespoke validations because they don't directly evaluate the meaning of the topics or the labels put on them. They do however explicitly validate key part of the analysis which are most important to the argument.

Our point is not to call into the question of any of these findings, but merely to characterize common approaches to validation. Articles coded with bespoke validation are not necessarily validated well, and articles without using bespoke validation well are not necessarily validated poorly. Our results do show that there is limited agreement on what kinds of validations of topic meaning should be shown to the reader. Twelve of twenty articles report only key words. Four of twenty report fit statistics. Five report external validation of topic meaning. Just one article reports all three forms of validation we coded (Barberá et al., 2019) and only one engages in the kinds of extensive bespoke validations described above.[15]

Thus, our overall finding is that aside from word lists, which are near universal, there are few consistently-used validation practices. Not surprisingly, extensive customized validations appear relatively rarely. This suggests the need for more validations that can be customized to the measurement task at hand, but can also be quickly and precisely conveyed to readers. Towards this end, we present an approach based on crowdsourced coding of word sets, documents, and topic labels. We emphasize again that this should not be seen as a *substitute* for theory-driven custom validation exercises or extensive reading, but rather as an *additional* tool.

3 Designing and assessing an off-the-shelf evaluation

In this article, we pursue the goal of designing an off-the-shelf evaluation design for TMs that leverage human ability to assess words and documents in context, can be easily and transparently communicated to readers, and is less burdensome than alternative such as training expert coders or machine learning classifiers. We develop two classes of designs: one extends the intrusion tasks of Chang et al. (2009) to evaluate the semantic coherence

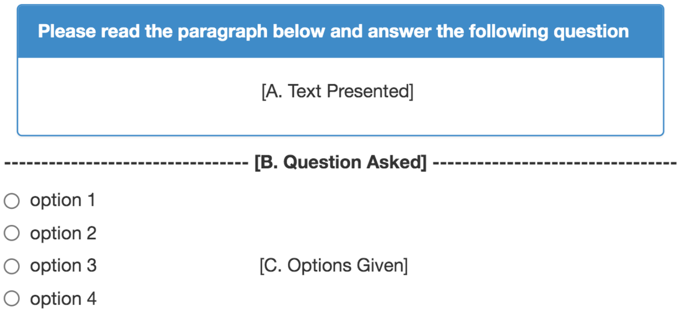
[15] This may be in part because the routinization of TMs has allowed researchers to use them in an increasing variety of settings—we observed cases of TM analysis as a form of exploration, as robustness checks for the main analysis, or as a validation of an alternative measurement strategy. In these settings, extensive validation may not be as necessary.

of a given TM (Section 4), and a second oriented towards validating that a set of topics corresponds to their researcher-assigned labels (Section 5). Before we present our method, we review the Chang et al. (2009) approach in Section 3.1, introduce our design principles in Section 3.2, and describe the data we will use for evaluation in Section 3.3.

3.1 Using the wisdom of the crowds

In an agenda-setting article, Chang et al. (2009) introduced a set of crowd-sourced tasks for evaluating TMs.[16] The core idea is to transform the validation task into short games which—if they are completed with high accuracy—imply a high quality model. The common structure for the two original tasks is shown in Figure 2. In each, a question (B) is presented to the coders and they must choose from options (C). Section (A) provides additional context for some tasks such as a document.

Figure 2: A Diagram for the Common Structure of Crowd-Sourced Validation Tasks



Please read the paragraph below and answer the following question

[A. Text Presented]

-------------------------------- **[B. Question Asked]** --------------------------------

○ option 1
○ option 2
○ option 3
○ option 4

[C. Options Given]

The first task in Chang et al. (2009), *Word Intrusion* (WI), is designed to detect topics which are semantically cohesive. We present workers with five words such as: `tax`, `payment`, `gun`, `spending`, `debt`. Four words of these words are chosen randomly from the high probability words from a given topic and an "intruder" word is chosen from the high probability words from a different topic. The human is then asked to identify the "most irrelevant" of the words—the intruder—which in the case above is `gun`. If the topic is semantically

[16] 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.

coherent the words from the topic should have clear relevance to each other and the intruder stands out. An example for each task structure is shown in Appendix SI2.

The second task, *Top 8 Word Set Intrusion* (T8WSI), detects coherence of topics within a document.[17] We present the coders with an actual document (or snippet from the document) and four sets of eight words such as,

```
(jobs, business, energy, new, economy, create, state, economic)
(work, project, forward, need, american, legislation, support, make)
(oil, energy, security, pipeline, administration, states, strategy, must)
(day, family, holiday, summer, beach, play, sunshine, vacation)
```

Each of the four word sets contains the eight highest probability words for a topic. Three of these topics correspond to the highest probability topics for the displayed document, while one is a low probability for that document. The human is asked to identify the *word set* that does not belong—which in this case is `(day...vacation)`. Here the worker has both the cue from the document itself and from the pattern of co-occurence across topics.

When these tasks can be completed with high-accuracy by workers, it demonstrates that words within a topic are coherent (Word Intrusion) and that the topics that co-occur within a document are coherent (Top 8 Word Set Intrusion). Yet, they do not include the research-assigned labels for the topics and thus cannot demonstrate that topics represent what the researcher describes them as measuring. In Section 4 we will improve on these existing design for evaluating coherence and in Section 5 we introduce new designs for validating the labels.

3.2 Principles

We design the tasks to be generalizable, discriminative, reliable and easy to use. All tasks we present are *generalizable* to any mixed-membership model that represents a topic as a distribution over words and two of our designs also work with single-membership models. The

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

approach also generalizes to different substantive settings, varying document collection sizes, lengths of documents, and number of topics.We design the tasks to *discriminate* based on model quality, which involves ensuring that successful completion is correlated with higher quality models, but also that the tasks are of medium difficulty to avoid ceiling or floor effects. Further, even though these tasks involve subjective judgments, we demonstrate that they are *reliable* by showing that results are stable under replication.

Finally, this innovation is only helpful if scholars actually employ these techniques. Despite being highly cited, the approach in Chang et al. (2009) is rarely used in the academic literature and, as we already demonstrated in our review, extensive validations of TMs are rare. Thus, we prioritize *ease of use* and develop software to help users implement our methods.[18] Using workers on Amazon Mechanical Turk (MTurk) we were able to get results quickly and cheaply (usually in an afternoon and for less than $50 per task/model). For the researcher there is a fixed cost in getting set up on MTurk and building training modules for the workers and creating a set of gold standard HITs. But it does not require additional specialized skills and it is less arduous than alternatives such as establishing coding procedures for research assistants and/or training supervised classification algorithms. In addition to the software, we provide additional guidance and directions in the Appendix.

3.3 Empirical Illustration

As an empirical testbed, we collected US senators' Facebook pages from the 115th Congress and applied a series of common preprocessing steps.[19] We fit five structural topic models

[18] The R package, `validateIt` is currently available on github at `https://github.com/Luwei-Ying/validateIt` and can be installed easily using the `devtools` (Wickham et al., 2020) function `install_github()`.

[19] Three senators did not have public Facebook pages. We scraped every individual post from April 2018 back to when each page was initially created. The earliest date of a post is September 2007. We removed all numbers, punctuation marks, and stopwords (in the SMART stopword list). We also removed state names (full or partial), state abbreviations, and common titles such as "sen" and "senator." We converted all words to lower cases, but did not stem. Finally, we removed non-English posts, life events (e.g., "XX added a life event."), and those shorter than 10 words.

(STM; Roberts et al., 2013, 2016) using 163,642 documents.[20] In order to establish a clear benchmark for a flawed model, we estimate Model 1, a 10-topic STM run for only a single iteration of the expectation maximization algorithm. (Even this model appears reasonable at first glance because of the initialization procedure in STM, thus making for a strong test.[21]) We then fit three standard STM models with 10, 50, and 100 topics (Models 2-4). We do not have prior expectations of the quality ordering of these models. Finally, in order to provide a model which is almost certainly overfit given the length of the documents, we fit a 500 topic model (Model 5).

4 Coherence evaluations

We present three task structures designed to pick out distinctive and coherent topics. This aligns closely with the stated goals of analysts in the social sciences. For instance, Kim (2018, Appendix, p. 39) justifies the choice of 25 topics stating, "models with the lower number of topics do not capture distinct topics, while the model with 30 topics does not provide additional categories that are meaningful for interpretation." Similarly, Barnes and Hicks (2018, p. 346, footnote 13) say they chose the number of topics, "at which the topics content could be interpreted as substantively meaningful and distinct."

Table 1 summarizes all three task structures, where column names correspond to the annotations in the sample diagram from Figure 2. The Word Intrusion (WI) and the Top 8 Word Set Intrusion (T8WSI) tasks are slight alterations from the methods in Chang et al. (2009) (discussed above). The primary difference is that we combine the probability mass for words with a common root and randomly draw words according to their mass (in contrast with drawing words uniformly). The term "probability mass" here refers to the topic-specific probability assigned to a given token (remembering that topics are represented as word

[20] We randomly select 10% of the documents (16,364) and held out 50% of the tokens in these documents so later we will be able to compare the results from our methods with held-out log likelihood.

[21] The `stm` package (Roberts et al., 2019) uses a spectral method of moments (Arora et al., 2013) initialization strategy. Roberts et al. (2016) show that it is an effective initialization strategy for the main estimation routine, but Arora et al. (2013) show that it provides good solutions alone.

Table 1: Task Structures for Coherence Evaluations

	A. Text Presented	B. Question Asked	C. Options Given
WI	NA	Please read the five words below, and choose one that is most IRRELEVANT to the other four.	Four words mass-based selected from the top twenty high-probability words of one topic and one word (the intruder) mass-based selected from the top twenty high-probability words of another topic
T8WSI	A randomly selected document	After reading the above passage, please click on the set of words below that is most UNRELATED to passage.	Three word sets (each containing the top eight high-probability words) from the top three high-probability topics and one word set (the intruder) from another topic
R4WSI	NA	Please click on the word set below that is most UNRELATED to the other three.	Three word sets (each containing four mass-based selected words) from the top twenty high-probability words of one topic and one word set (the intruder) mass-based selected from the top twenty high-probability words of another topic

distributions). Combining the probabilities in this way is a bit like stemming the word after the modeling is complete. This allows us to show complete words to the human coders while also preventing multiple words with a common root from appearing in the same task.

In our initial testing, we found that the WI and T8WSI tasks were often too difficult for coders, reducing their power to discriminate. Further, T8WSI is sensitive to the words included in the "top eight," making the results more arbitrary and again less informative. To address these concerns, we designed a new task, *Random 4 Word Set Intrusion* (R4WSI) which we summarize in the final row of Table 1.

In R4WSI, we present the coder with four different sets of four words such as,

```
(voting, nominee, court, confirmation)

(judge, supreme, rights, legal)

(citizens, nomination, decision, jury)

(serve, veterans, overseas, fight)
```

Similar to WI, three of these sets of words are chosen from the same topic, while an intruder word set comes from a different topic.[22] The coder's goal is identify the intruder *word set* (here `serve...fight`). In this new design, coders have access to 12 words from the non-intruder topic and thus more context to identify a common theme resulting in more informative decisions.

We tested these three task structures using workers with master certifications from Amazon's Mechanical Turk (AMT) from March to July, 2020. To qualify, workers had to complete an online training module described in Appendix SI6. The training explains the task, provides background about the document set, and walks workers through examples to ensure they understand their goals. In Appendix SI4, we emphasize that these training modules are critical for screening workers with the requisite skills and knowledge and putting the tasks in context for the coders.

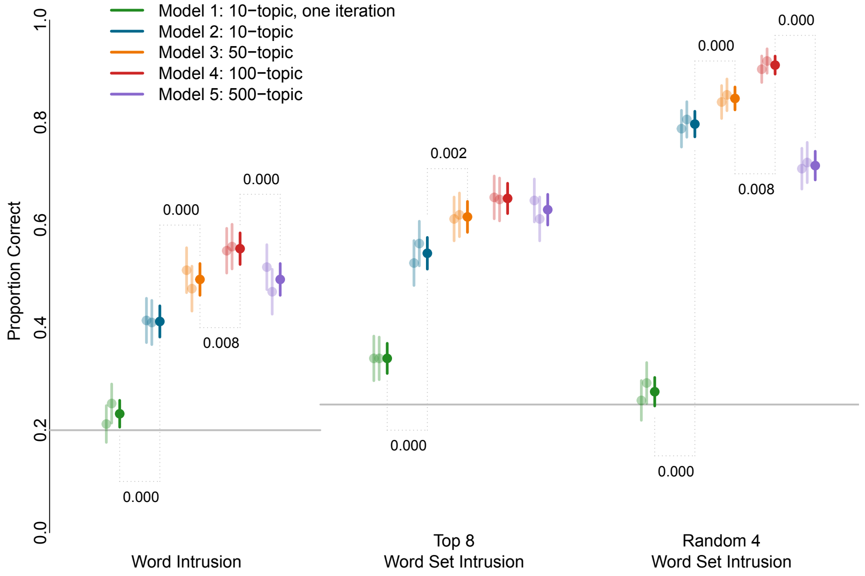
We paid $0.04/task for WI, $0.08/task for T8WSI, and $0.06 for R4WSI (which corresponds to roughly $15 per hour on average). For each task structure we posted 500 tasks, which Amazon calls human intelligence tasks (HITs) for all five models. To assess the consistency of task structures, we then posted these *exact same tasks* again. To monitor the quality of the work, we randomly mixed in a gold-standard HIT every ten HITs.[23] In total, workers completed 16,500 tasks. However, a single batch of 500 HITs—a typical case for an applied researcher—takes only a few hours with total costs in the range of $25-$60.

Figure 3 shows the results for all five of our models on each of the three tasks. The first two light color bars indicate the two identical runs and the third darker line indicates the pooled results of those runs. We also indicate when the the difference in means is significant across model pairs with connecting dotted lines, where the numbers represent p-values for a difference in proportions test (n=2000). We make three observations. First, all task structures easily identified the non-converged baseline (Model 1) as the worst, which

[22] The four words are chosen at random from the top 20 words associated with a topic with the restriction that no word stems should be repeated across word sets.

[23] We suppressed the qualification of workers who have missed more than 2 gold-standard HITs or who have done a relatively large number of HITs of a specific task structure. This operation has no negative impact on their Mturk records. We have rejected and replaced work from two workers (267 HITs in total) who missed more than 4 HITs each.

Figure 3: Results for Coherence Evaluations



Note: The 95% confidence intervals are presented. The two light bars represent two identical trials (500 HITs each). The dark 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.) No identical trails (two light bars) are significantly different from each other. The grey horizontal line represents the correct rates from random guessing.

provides a check that this approach has the ability to identify a model known to be a relatively poor fit. Second, all of them are able to identify over-fitting as the 500-topic model (Model 5) appears to be worse than the 100-topic model (Model 4) in all task structures. Third, all of the task structures are reliable in that they provide nearly indistinguishable estimates across runs when we include 500 tasks.

Overall, these results provide evidence of several advantages of the R4WSI task structure. The estimated held-out log likelihood for Models 1-5, respectively, are -8.316, -7.981, -7.767, -7.705, and -7.984 (higher is better). This rank ordering (with the 500-topic scoring lower than the versions with 10 or 50 topics) is consistent with R4WSI but not WI and T8WSI. R4WSI also more clearly distinguishes the unconverged Model 1 as inferior. The higher accuracy rates suggest that R4WSI task is indeed easier for workers to understand and complete with workers identifying the intruder nearly 85% of the time for Model 4. This suggests that the model has identified meaningful and coherent patterns in the document set that humans can reliably recognize. While all the tasks appear reasonably effective, we recommend the R4WSI task for applied use.

5 Label Validation

In social science research, scholars typically place conceptual labels on topics that indicate the concept they are measuring. The accuracy of these labels may have relatively low stakes if topics are only used for prediction (e.g., Blumenau and Lauderdale, 2018). However, in the majority of applications we reviewed, the stakes are high as the label communicates to the reader the nature of the evidence that the text provides about a theoretical claim of interest (e.g., Barnes and Hicks, 2018; Horowitz et al., 2019; Magaloni and Rodriguez, 2020; Gilardi et al., 2021). In many cases, the individual labels may be important, but play a less central role in the analysis than the label assigned to a cluster of topics which share a common trait of interest (e.g., Barberá et al., 2019; Dietrich et al., 2019; Lacombe, 2019; Martin and McCrain, 2019; Motolinia, 2020). Reflecting the differences in social science usage of TMs, these concerns of label validity are largely unaddressed by the designs that originated in computer science (Chang et al., 2009).

We develop label validations for these use cases and test them on the 100-topic model (Model 4). First, we ask, "Are the conceptual labels sufficiently precise and non-overlapping to allow us to distinguish between closely related topics?" Specifically, we identified ten topics related to domestic policies and focus our analysis on only these topics. Second, we ask, "Can we usefully distinguish two broad conceptual categories of discussion from each other?" Specifically, we identified ten topics related to the military and foreign affairs and focus on coders' ability to distinguish between these topics and the "domestic" policy topics.[24]

[24] A different strategy might be to generate a list of potential labels and use crowdsourcing to choose the "best" option. We show an example of this procedure in Appendix SI-3.3. However, there is a danger on

A problem for validating any new validation method is that we lack an unambiguous ground truth—many possible labels would accurately describe the contents of a topic and many labels would not. Ideally, our task will allow us to discriminate between higher and lower quality labels. In our empirical case, we need to produce a set of labels for which we have strong *a priori* expectations.

Members of our research team independently labeled each of the 100 topics. Each of us carefully read the high-probability words and frequent and exclusive words (FREX) (Roberts et al., 2016), as well as 50 representative documents per topic (Grimmer and Stewart, 2013). From the topics that all of us deemed as coherent, we picked ten domestic topics and ten international/military topics where the labels were most consistent. The final labels for each are shown in Table 2 with additional details in the appendix. We refer to these as "careful coder" labels. To provide a contrast, we asked research assistants to create their own set of labels based only on the high probability and FREX words (i.e., without looking at the documents). These labels, which we refer to as "cursory coder" labels, are shown in the second column of Table 2. Our expectation is that the careful coder labels are better labels (and thus should score more highly on the tasks) but that the cursory coder provides a reasonably strong baseline.[25]

5.1 Novel task structures

We designed two task structures to evaluate label quality which are summarized in Table 3: *Label Intrusion* and *Optimal Label.*[26] In the *Label Intrusion* (LI) task the coder is shown a text and four possible topic labels. Three of the labels come from the three topics most

relying on the crowd to *choose* the topic labels in isolation rather than to validate the topic sets proposed by researchers. As we show in Appendix SI-3.3, crowd workers can easily miss basic facts about the topics. Specifically, we show that workers may tend to favor more specific labels for a given document even when the actual topic is much broader.

[25] We also present the labels in random order to yet another coder along with high probability words, FREX words, and 50 documents associated with each topic. This final coder was given the alternative labels in a random order and asked to pick the superior label reflecting the underlying concept. The coder picked 19 out of 20 labels developed using our "careful coder" procedure as being the most appropriate.

[26] We evaluated two additional tasks structures reported in Appendix SI3.

Table 2: Labels to Validate

Careful Coder	**Cursory Coder**
Domestic Topics	
Equal Pay for Women	Working Class
Healthcare/Reproductive Rights	Planned Parenthood
Agriculture	Farm Bill
Student Loan/Debt	Economy
Drug Abuse	Prescription Medicine
Higher Education/Job Training	Grants for Colleges
Wall Street/Financial Sector	Banking
Government Shutdown/Congressional Budget	Government Spending
Obamacare/Tax Policy	Healthcare
Deficits/Debt/Budget	Debt Ceiling
International/Military Topics	
International Trade	Manufacturing
Praising Active Military/Military Units	"Welcome Home" Messages
Terrorism	Islamic Extremists
Military Sexual Assault	Military Affairs
Nuclear Deterrence/International Security	Foreign Affairs
Air Force	Military
Honoring Specific Veterans	Military Service
Honoring Veterans/Heroes	"Thank you" Messages
Military Operations/Armed Conflicts	Counter-terrorism
Veterans Affairs/Veterans Healthcare	Veterans

associated with the document and one is selected from the remaining seven labels ("Within Category") or seven plus the ten international labels ("Across Category"). The coder is asked to identify the intruder, mimicking the word set intrusion design.

The second task, *Optimal Label* (OL), presents a document and four labels. One label is for the highest probability topic and the other three labels are chosen randomly from the remaining nine domestic labels ("Within Category") or nine plus the ten international labels ("Across Category"). The coder is asked to identify the best label. This optimal label task structure 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). 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 and measures their accuracy in doing so.

Table 3: Task Structures for Label Validation

	A. Text Presented	B. Question Asked	C. Options Given
LI	A randomly selected document[a]	Please read the four labels below and click on the label that is most UNRELATED to the passage.	*Within Category:* Three labels for the top three high-probability topics and one label for other domestic topics; *Across Categories:* Three labels for the top three high-probability topics and one label for other domestic or international/military topics
OL	A randomly selected document[b]	Please read the four labels below and click on the label that BEST summarizes the passage.	*Within Category:* One label for the highest-probability topic and three labels for other domestic topics; *Across Categories:* One label for the highest-probability topic and three labels for other domestic or international/military topics

[a] Top three predicted topics among the ten domestic topics.

[b] Top one predicted topic among the ten domestic topics.

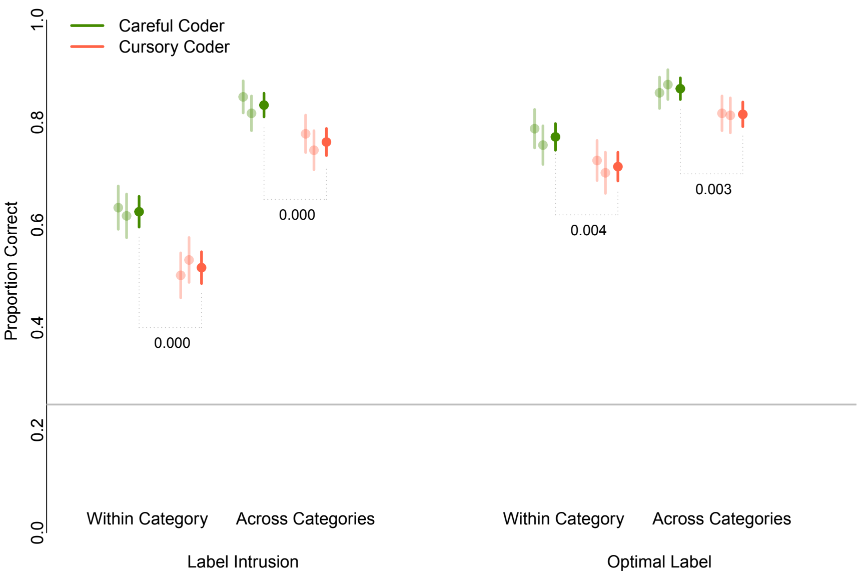
In addition, we anticipated that discriminating between only domestic topics would be harder than discriminating between topics where intruders could be either domestic or military/international topics. That is, discriminating between conceptually similar topics (e.g., Drug Abuse vs. Healthcare/Reproductive Rights) is understandably a “harder test” than discriminating between clearly distinct topics (e.g., Drug Abuse vs. Terrorism).

5.2 Results

For each task/coder combination we created 500 tasks (plus 50 gold-standard HITs for evaluation purpose) that were coded by trained workers on AMT for $0.08 per HIT. These were then repeated so that we could assess worker quality and replace work from low-quality workers. In total, workers completed 8,800 HITs and the results are shown in Figure 4.

The results are positive for both tasks. With 500 HITs the results across runs are reliable with rank orderings of the label sets being indistinguishable across repetitions. Second, the results are consistent across task structures in identifying the careful coder labels as being superior. Finally, in Table 4 we show that workers achieve much higher correct rates when the goal is to distinguish across the broader conceptual categories (domestic vs. in-

Figure 4: Results for Label Validation



Note: The 95% confidence intervals are presented, where the two light bars represent two identical trials (500 HITs each) and the dark bar represents the pooled result (1000 HITs). p-values are based on the pooled set of tasks based on a difference-in-proportions test. No identical trials are significantly different from each other. The grey horizontal line represents the correct rate from random guessing.

ternational/military policies). For instance, when all three intruders crossed this conceptual boundary, coders were able to choose the correct optimal label 96.4% of the time for the careful coder labels while that figure falls to 78.8% when intruders were limited to other domestic topics.

Both the LI and OL tasks are reasonable choices for applied researchers. The LI task only works for mixed membership models and will be most effective when most documents strongly express multiple topics (and capturing more than the top label is particularly important). The OL task is more easily interpretable and can work for both single- and mixed-membership models, but relies on the ability of the coder to pick out the single best label which can be difficult in documents that are best represented by a mixture over many topics. In both designs, the researcher must also choose whether to draw the comparison topics from a set of conceptually-related topics or from across broad categories. Closely related topics represent a harder test, but when the primary research claim is about the broader category,

Table 4: Disaggregation of Figure 4 Accuracy Rates for ”Across” Condition by Intruder Categories.

	Intruder from a Different Category	
Label Intrusion	No	Yes
Careful Coder	0.703	0.928
Cursory Coder	0.490	0.939

		Number of Incorrect Labels from a Different Category		
Optimal Label	Zero	One	Two	Three
Careful Coder	0.788	0.816	0.896	0.964
Cursory Coder	0.717	0.788	0.835	0.918

Note: All documents included are about domestic policy, so a cross-category option is any international label. It is possible to have zero international labels (as in the "within" condition) because in the "across" category condition we are randomly selecting labels from both categories.

the task of making fine-grained distinctions may be unnecessarily difficult.

In our particular application, the results suggest that coders can easily make distinctions between broader policy categories (e.g., domestic and international policy debates). When looking only within a narrow set of topics, however, our results indicate a need for caution. When considering only the ten domestic policy topics, the coders could identify an intruder only 70.3% of the time for the "careful coder" labels and less than half (49.0%) of the time for the cursory coders. This suggests that the careful coder labels are substantially better, but depending on the downstream task, even 70% might be concerning. The corresponding numbers for the OL task (78.8% and 71.7%) corroborate this finding, indicating that the careful coder lables are better, but we should put less faith in the fine-grained distinctions.

6 Limitations

Our goal is not to present the final word on this methodological question, but rather to begin a dialogue. Our collective work on validating topics as measures is just getting started. With this in mind, we highlight three limitations.

Limitation 1: These Designs Should Not Replace Bespoke Validation When it comes to validation, there is no substitute for testing a measure against substantive, theoretically driven expectations in a bespoke validation. As Brady (2010, p. 78) writes, "Measurement ... is not the same as quantification, and it must be guided by theories that emphasize the relationship of one measure to another." Yet, as we noted in our review, bespoke validations appear so infrequently in the published literature that it may be helpful to extend the toolbox with new options.

The central advantage of these new tasks is that they are low cost, reliable, and easy to communicate to a reader. For any given application, there is likely a custom-designed solution which will be superior, but our tasks provide an approach that researchers can reach for in most circumstances. In the best case scenario, our proposed tasks would offer a complement to essential but difficult to convey validation methods such as close reading of the underlying text.

The ongoing need for bespoke validation is inextricably connected to the fact that we do not have access to a ground truth to benchmark validations against and thus we cannot guarantee that they will be accurate in general. Our 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 sufficiently accurate to be distinguished among themselves. Importantly by using human judgments, our validations occupy a space between expert assessments and statistical metrics which lack any human judgment at all.[27]

[27] 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.

Limitation 2: These Designs Have Limited Scope While a major advantage of our designs is that they are more general than a given bespoke strategy, there are nonetheless some limitations in scope arising from the simplification inherent in the tasks. To begin, the documents have to be *accessible* to the workers. At a minimum documents have to be in a language the workers can read. Mechanical Turk relies primarily on a US-based workforce, but Pavlick et al. (2014) shows that it is possible to find workers with specific language skills and our experience shows that only a small number of workers are needed to complete these coding tasks. There are also alternative crowdsourcing platforms with more international workers (Benoit et al., 2016).[28] Still, future research is needed to show that this approach is feasible for non-English texts. In addition, several of the task structures require coders to read documents or excerpts. This is reasonable for social media posts and other short texts that are the basis of most applications of TMs to date. Our document set is particularly well-positioned to use this technique, but that in turn makes it a comparatively easy case. Future work might explore how to best handle excerpting long documents or training workers for specialized texts (e.g., Blaydes et al., 2018).

A more subtle limitation is that the representation of topics using a fixed number (e.g., 20) of the most probable words can present challenges in certain model fits. TMs 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.[29] This is another instance of text pre-processing decisions may play a consequential role in unsupervised learning (Denny and Spirling, 2018). Because these concerns will arise in the creation of the training

[28] Eshima et al. (2020) build on our task structures using international workers with a custom-built Qualtrics module.

[29] 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).

module for the workers, researchers will know in practice when this issue is arising and can adjust accordingly (e.g., by considering a smaller number of words).

We also emphasize that these designs cannot evaluate all properties necessary for accurate 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.

Limitation 3: Results Are Difficult to Interpret in Isolation A final limitation 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, but in practice applied researchers may only be evaluating a single model. If Model 3 scores 61.6% on the T8WSI 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 over p-values. Thus, finding the right way to compare evidence across datasets remains an open challenge although one that exists for any kind of validation metric (including model fit statistics and bespoke evaluations). Authors will need to provide readers with context for evaluating and interpreting these numbers, perhaps by evaluating multiple models or using multiple validation methods. At a minimum, as readers we should expect to see that coders substantially exceed the threshold for random guessing (which is marked in all our plots). Still, as we accumulate more evidence about such validation exercises, it may become possible to get a better sense of what an “adequate” score will be.

7 Conclusion

The text-as-data movement is exciting, in part, because it comes with a rapidly expanding evidence base in the social sciences (King, 2009). The conventional sources of data such as surveys or voting records are giving way study-specific, text-based datasets collected from the Internet or other digital sources. This means that individual scholars are increasingly taking on the role of designing unique measurements for each study built from messy, unstructured, textual records. While greatly extending the scope of the social sciences, this expansion places new burdens of validation on researchers which must be met with new, widely-applicable tools.

We have taken a step in this direction by improving upon the existing crowd-sourced tasks of Chang et al. (2009) and extending them to create new designs that assess how well a set of labels represent corresponding topics. 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 leverage the ability of humans to understand natural language without experiencing the scale issues of relying on experts.

Recognizing that such advancements are only helpful if they are straightforward enough for researchers to apply in their own work, we have built an `R` package which automates much of the work of launching these tasks. While they do require a fixed cost in time and effort to set up, they are a straightforward way to include external human judgement. Our 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 $20 and $60. Nonetheless, there are still improvements to be made in terms of best practices for worker recruitment, training, and task structure. This is particularly true as the workforce and platforms are moving targets and future work might discover new challenges or new ways to ensure data reliability.

The social sciences have reimagined topic models for a purpose very different from the original goals of information retrieval in computer science. Yet these new ambitions bring with them new responsibilities to validate topic models with same high standards we apply to other measures in the social sciences. Early topic modeling work handled this with extensive bespoke validations, but as the topic model fitting routinized, the validations have not followed suit. In short, there is no free lunch: any method used for measurement—unsupervised topic models, supervised document classification, or any non-text approach—requires validation to ensure that the learned measurement is valid. This paper makes what is hopefully only one of many efforts to give renewed attention to measurement validation for text-as-data methods in the social sciences.

8 Funding

Funding for this project was provided by the Weidenbaum Center on the Economy, Government, and Public Policy at Washington University in St. Louis, the National Science Foundation RIDIR program award number 1738288 and The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879.

9 Acknowledgements

We are grateful for helpful comments from Kevin Quinn, Arthur Spirling and audiences at the 2019 Summer meeting of the Society of Political Methodology, IV Latin American Political Methods Meeting, The International Methods Colloquium, Washington University in St. Louis, New York University, the University of California, Davis and the University of California, San Diego. A number of scholars whose work we reviewed in Section 2 also graciously provided comments including Benjamin E. Bagozzi, Pablo Barberá, Lisa Blaydes, Kaiping Chen, Germán Feierherd, Matthew Hayes, Tim Hicks, Michael Horowitz, Matthew J. Lacombe, Benjamin Lauderdale, Kenneth Lowande, Lucia Motolinia, Richard A. Nielsen, Jennifer Pan, Margaret Roberts, Arturas Rozenas, and Dustin Tingley. We thank Ryden Butler for providing the data.

10 Data Availability Statement

Replication code for this article is available at Ying et al. (2021) at `https://doi.org/10.7910/DVN/S02EBF`.

References

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

Armstrong, J. S. (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, S., R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu (2013). A Practical Algorithm for Topic Modeling with Provable Guarantees. In *International Conference on Machine Learning*, pp. 280–288.

Bagozzi, B. E. and D. Berliner (2018, October). 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.

Barberá, P., A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review 113*(4), 883–901.

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

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review 110*(2), 278–295.

Blaydes, L., J. Grimmer, and A. 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, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022.

Blumenau, J. and B. E. Lauderdale (2018, January). Never Let a Good Crisis Go to Waste: Agenda Setting and Legislative Voting in Response to the EU Crisis. *The Journal of Politics 80*(2), 462–478.

Brady, H. E. (2010). Doing Good and Doing Better: How Far Does the Quantitative Template Get Us. In H. E. Brady and D. Collier (Eds.), *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (Second ed.).

Cacioppo, J. T. and R. E. Petty (1982). The need for cognition. *Journal of Personality & Social Psychology 42*(1), 116–131.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, pp. 288–296.

Clinton, J., S. Jackman, and D. Rivers (2004). The Statistical Analysis of Roll Call Data. *American Political Science Review 98*(2), 355–370.

Denny, M. J. and A. 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.

Dietrich, B. J., M. Hayes, and D. Z. O'Brien (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech. *American Political Science Review 113*(4), 941–962.

Egami, N., C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart (2018). How to Make Causal Inferences Using Texts. *arXiv preprint arXiv:1802.02163*.

Eshima, S., K. Imai, and T. Sasaki (2020). Keyword assisted topic models. *arXiv preprint arXiv:2004.05964*.

Gelman, A. and E. Loken (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University 348*.

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

Gilardi, F., C. R. Shipan, and B. Wüest (2021). Policy diffusion: The issue-definition stage. *American Journal of Political Science 65*(1), 21–35.

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

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

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

Grimmer, J., M. E. Roberts, and B. M. Stewart (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science 24*.

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), 267–297.

Horowitz, M., B. M. Stewart, D. Tingley, M. Bishop, L. Resnick Samotin, M. Roberts, W. Chang, B. Mellers, and P. Tetlock (2019). What makes foreign policy teams tick: Explaining variation in group performance at geopolitical forecasting. *The Journal of Politics 81*(4), 1388–1404.

John, L. K., G. Loewenstein, and D. Prelec (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science 23*(5), 524–532.

Karell, D. and M. R. Freedman (2019). Rhetorics of radicalism. *American Sociological Review*.

Kim, S. E. (2018). Media bias against foreign firms as a veiled trade barrier: Evidence from Chinese newspapers. *American Political Science Review 112*(4), 954–970.

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

King, G., R. O. Keohane, and S. Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press.

Lacombe, M. J. (2019, July). The Political Weaponization of Gun Owners: The National Rifle Association's Cultivation, Dissemination, and Use of a Group Social Identity. *The Journal of Politics 81*(4), 1342–1356.

Lau, J. H., K. Grieser, D. Newman, and T. Baldwin (2011). Automatic Labelling of Topic Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1536–1545. Association for Computational Linguistics.

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

Lund, J., P. Armstrong, W. Fearn, S. Cowley, C. Byun, J. Boyd-Graber, and K. Seppi (2019, May). Automatic Evaluation of Local Topic Quality.

Magaloni, B. and L. Rodriguez (2020). Institutionalized Police Brutality: Torture, the Militarization of Security, and the Reform of Inquisitorial Criminal Justice in Mexico. *American Political Science Review 114*(4), 1013–1034.

Martin, G. J. and J. McCrain (2019). Local news and national politics. *American Political Science Review 113*(2), 372–384.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum (2011). Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272. Association for Computational Linguistics.

Motolinia, L. (2020). Electoral Accountability and Particularistic Legislation: Evidence from an Electoral Reform in Mexico. *American Political Science Review*, 1–17.

Nielsen, R. A. (2020). Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers. *American Journal of Political Science 64*(1), 52–66.

Pan, J. and K. Chen (2018). Concealing corruption: How chinese officials distort upward reporting of online grievances. *American Political Science Review 112*(3), 602–620.

Parthasarathy, R., V. Rao, and N. Palaniswamy (2019). Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies. *American Political Science Review 113*(3), 623–640.

Pavlick, E., M. Post, A. Irvine, D. Kachaev, and C. Callison-Burch (2014). The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics 2*, 79–92.

Poole, K. T. and H. Rosenthal (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, 357–384.

Pratto, F., J. Sidanius, L. Stallworth, and B. Malle (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology 67*(4), 741–741.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010, January). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science 54*(1), 209–228.

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

Roberts, M. E., B. M. Stewart, and R. A. Nielsen (2020). Adjusting for confounding with text matching. *American Journal of Political Science 64*(4), 887–903.

Roberts, M. E., B. M. Stewart, and D. Tingley (2016). Navigating the Local Modes of Big Data. *Computational Social Science 51.*

Roberts, M. E., B. M. Stewart, and D. Tingley (2019). stm: An R package for structural topic models. *Journal of Statistical Software 91*(2), 1–40.

Roberts, M. E., B. M. Stewart, D. Tingley, and E. 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*, pp. 1–20. Harrahs and Harveys, Lake Tahoe.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand (2014). Structural Topic Models for Open-ended Survey Responses. *American Journal of Political Science 58*(4), 1064–1082.

Rozenas, A. and D. Stukal (2019, June). How Autocrats Manipulate Economic News: Evidence from Russia's State-Controlled Television. *The Journal of Politics 81*(3), 982–996.

Wallach, H. M., I. Murray, R. Salakhutdinov, and D. Mimno (2009). Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1105–1112. ACM.

Wickham, H., J. Hester, and W. Chang (2020). *devtools: Tools to Make Developing R Packages Easier*. R package version 2.3.1.

Ying, L., J. M. Montgomery, and B. M. Stewart (2021). Replication Data for: Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures. `https://doi.org/10.7910/DVN/SO2EBF`, Harvard Dataverse, V1.