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

June 2, 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.

[*]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. 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. Replication data available in Ying et al. (2021).

[†]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.

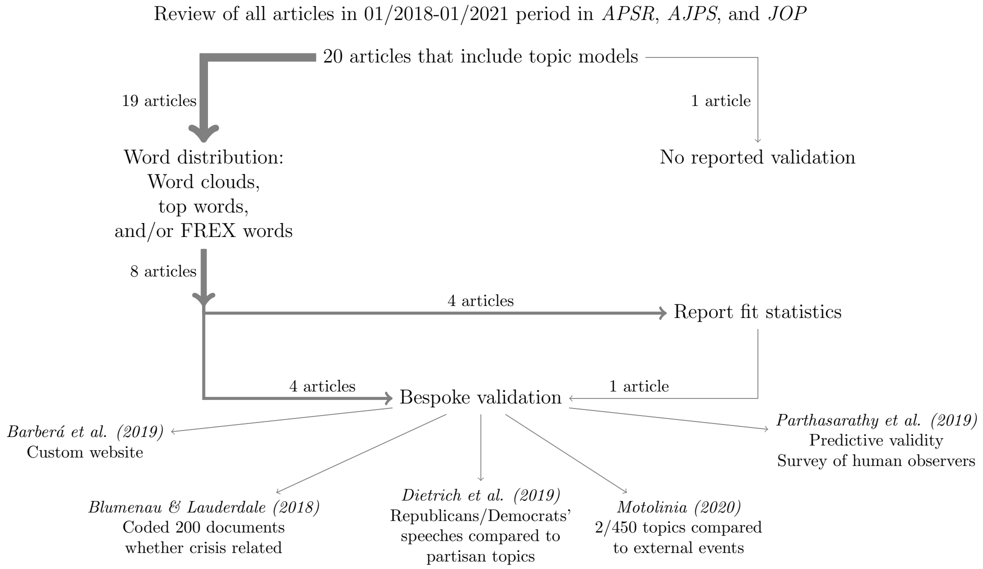
2.2 A review of recent practices

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



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

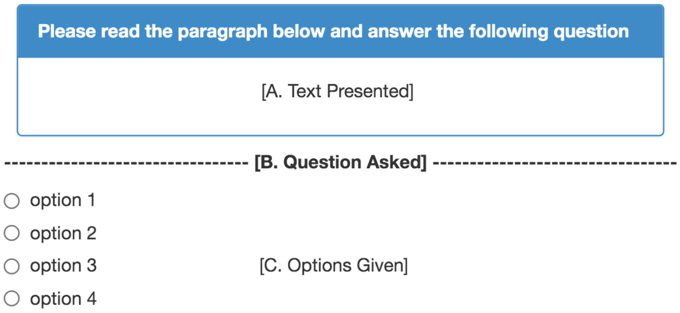
[15] This may be in part because the routinization of TMs has allowed researchers to use them in an increasing variety of settings—we observed cases of TM analysis as a form of exploration, as robustness checks for the main analysis, or as a validation of an alternative measurement strategy. In these 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.

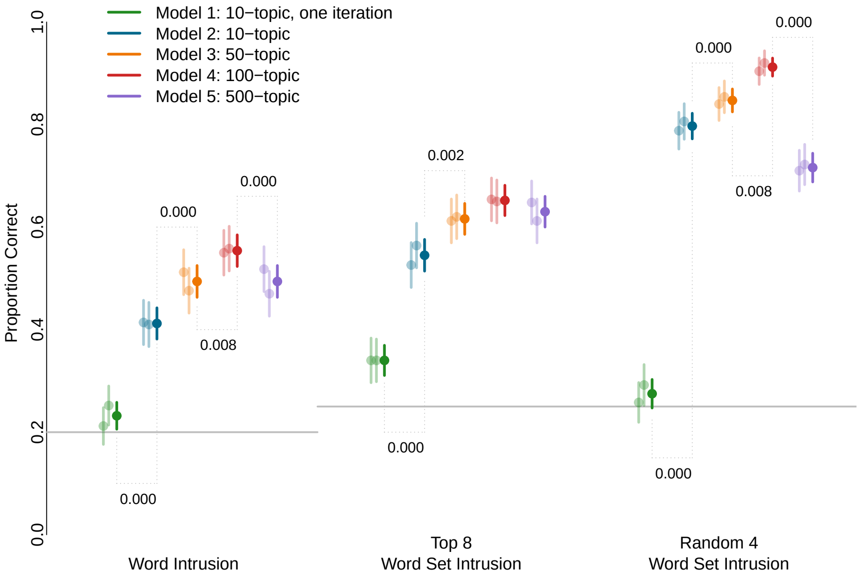
We paid $0.04/task for WI, $0.08/task for T8WSI, and $0.06 for R4WSI (which corresponds to roughly $15 per hour on average). For each task structure we posted 500 tasks, which Amazon calls human intelligence tasks (HITs) for all five models. To assess the consistency of task structures, we then posted 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.

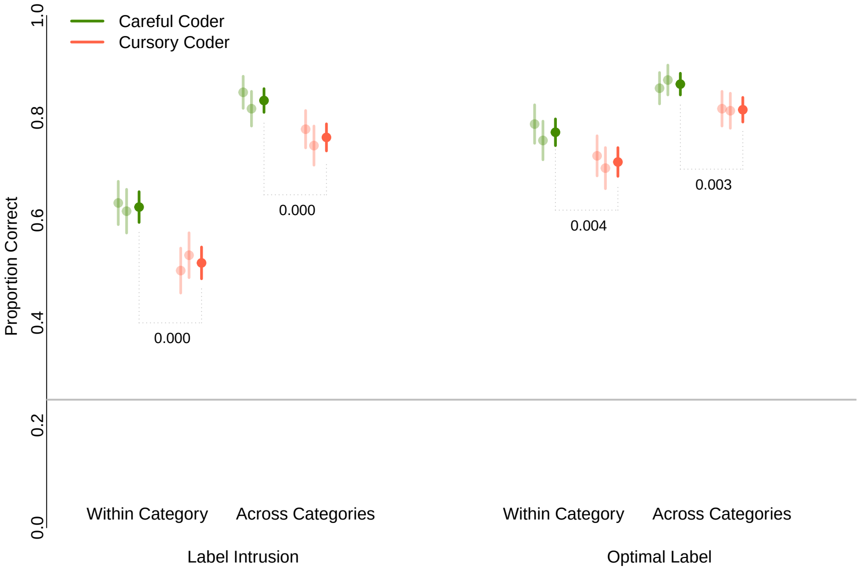
In addition, we anticipated that discriminating between only domestic topics would be harder than discriminating between topics where intruders could be either domestic or military/international topics. That is, discriminating between conceptually similar topics (e.g., Drug Abuse vs. Healthcare/Reproductive 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: Accuracy within and across Two Broader Categories

Label Intrusion	Within	Across[a]		
Careful Coder	0.703	0.928		
Cursory Coder	0.490	0.939		
Optimal Label	Within	1 Across[b]	2 Across	3 Across
Careful Coder	0.788	0.816	0.896	0.964
Cursory Coder	0.717	0.788	0.835	0.918

[a] The intruder is one of the ten international labels

[b] One of the three intruders is one of the ten international labels

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]

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

[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.

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 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.

[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).

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.

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.

Supplementary Information

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 Stewart
Princeton University

June 2, 2021

The supplementary information to "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures" includes seven sections. The first section expands on the current validation practices in the literature. The second section shows example HITs for each task as they appear on the online workers' screen including an example HIT for a non-English text. The third section presents and discusses several "less ideal" validation task designs that we have tried but decided not to recommend for future researchers. This includes an approach to using crowdsourcing to choose between alternative labels. The fourth section summarizes all of our recommendations for validating a topic model and topic labels including practical tips on training and monitoring workers. The fifth section introduces our R package, `validateIt`, and provides a manual walking people through the procedures of posting their own tasks. The sixth section provides an example of the training module we used and discusses workers' performance. The last section contains additional information on the corpus and model fit.

Contents

1 Current Practices in Political Science: A Survey SI: 1

2 Example HITs SI: 6
2.1 Example HITs for Topic Validation . . . . SI: 6
2.2 Example HITs for Label Validation . . . . SI: 7
2.3 Example HITs in non-English (Chinese) . . . . SI: 8

3 Alternative Task Designs and Results SI: 9
3.1 R4WSI Tasks with Documents . . . . SI: 9
3.2 Two Alternative Label Validation Tasks . . . . SI: 11
3.3 Selecting among Alternative Labels? . . . . SI: 13

4 A Checklist for Crowdsourced Topic Validation SI: 17

5 Software and Working Example SI: 19
5.1 R Package: `validateIt` . . . . SI: 19
5.2 Fitting Candidate Topic Models and Labeling . . . . SI: 19
5.3 Training and Certification . . . . SI: 21
5.4 Preparing Tasks . . . . SI: 23
5.4.1 Preparing Tasks for Topic Validation . . . . SI: 23
5.4.2 Preparing Tasks for Label Validation . . . . SI: 26
5.5 Posting Tasks and Maintaining . . . . SI: 28
5.6 Retrieving and Evaluating Results . . . . SI: 32

6 Workers' Training and Performance SI: 35
6.1 Training Module for Word Intrusion . . . . SI: 35
6.2 Training Module for Top 8 Word Set Intrusion . . . . SI: 39
6.3 Training Module for Random 4 Word Set Intrusion . . . . SI: 46
6.4 Training Module for Label Intrusion . . . . SI: 50
6.5 Training Module for Optimal Label . . . . SI: 56
6.6 Workers' Performance . . . . SI: 62

7 More on the Corpus, Topic Model Fit, and Labeling SI: 63
7.1 Word Mass Distribution . . . . SI: 63
7.2 Word Clouds . . . . SI: 64
7.3 Representative Documents for Labeling . . . . SI: 68

1 Current Practices in Political Science: A Survey

To construct the review of recent practice in Section 2.2 of the main paper, we identified all papers published in the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics* from January 1, 2018 to January 2021 (all articles published online at the time of our search) that included the phrase "topic model." This resulted in twenty publications. Table SI1-SI3 present the full set of articles and our detailed coding for each of them. Two authors coded each article independently and all three authors discussed cases where there was disagreement to arrive at a consensus coding.

As explained in the main text, we have omitted a number of different types of validations, particularly authors reporting that they have carefully read the documents. While this is doubtless the most important validation strategy, we are explicitly interested here only in (1) validations of topic meaning that are (2) reported in the main body or appendix of the paper, and (3) where the reader can observe the result of the work done. This leaves out a lot of interesting validations but corresponds to our setting of interest.

While coding these articles' validation practices, we created three dichotomous variables reflecting the most common classes of validation strategies reported (in either the main paper or the appendix only) which met our criteria: topic-specific word lists, fit statistics, and bespoke validation of individual topic meanings. Topic-specific word lists include either the most probable words in the topic under the model or alternative criteria such as frequency and exclusivity (FREX) (Roberts et al., 2016) words. We included both cases where the words were reported in lists or visualized in word clouds. Model fit statistics include held-out log likelihood (Wallach et al., 2009) or surrogate scores such as "semantic coherence" (Mimno et al., 2011; Roberts et al., 2014). These provide a sense of whether or not the model is over-fitting, and some previous research shows that surrogates correlate with human judgements. We also coded the bespoke approaches—additional validations of topic meaning designed especially for their case to establish construct validity. These were more varied and unavoidably subjective. We describe many of the individual procedures below.

In Table SI1-SI3, we also provide more information with regard to the corpus, model choice, and whether or not the measure from topic models is used as a predictor or outcome variable. In addition to the three approaches to validation we coded, several papers took creative and unique approaches to validation that do not fit easily into this coding scheme. We have marked them in Table SI3 and follow the table with short summary notes.

We emphasize that the result of this coding should not be used to critique any given study—articles used topic models for a wide-variety of purposes and to varying degrees as the foundation of their evidence base. For example, Pan and Chen (2018) uses the topic model for exploration (which does not itself need to be validated) and Rozenas and Stukal (2019) uses it as a validation of another measure (where a call for validation could quickly turn into an infinite regress). We estimate in at least 1/4 of the cases, the validations we code arguably aren't directly relevant because the topic model was not used for measurement or was not a central part of the analysis. However, making this assessment at an article-level proved too difficult to do consistently. Thus we encourage readers to consider these primarily as aggregate statements about the approaches being used in the literature as we have presented them in the main paper.

Table SI1: Topic model Usage in Three Political Science Journals

No.	Author Date	Journal	Title
1	Barberá et al. (2019)	APSR	Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data
2	Barnes and Hicks (2018)	AJPS	Making Austerity Popular: The Media and Mass Attitudes toward Fiscal Policy
3	Berliner et al. (2020)	JOP	The Political Logic of Government Disclosure: Evidence from Information Requests in Mexico
4	Blaydes, Grimmer and McQueen (2018)	JOP	Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds
5	Blumenau and Lauderdale (2018)	JOP	Never Let a Good Crisis Go to Waste: Agenda Setting and Legislative Voting in Response to the EU Crisis
6	Dietrich, Hayes and O'Brien (2019)	APSR	Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech
7	Feierherd (forthcoming)	AJPS	Courting Informal Workers: Exclusion, Forbearance, and the Left
8	Horowitz et al. (2019)	JOP	What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance at Geopolitical Forecasting
9	Jiang and Zeng (2019)	JOP	Countering Capture: Elite Networks and Government Responsiveness in China's Land Market Reform
10	Kim (2018)	APSR	Media Bias against Foreign Firms as a Veiled Trade Barrier: Evidence from Chinese Newspapers
11	Lacombe (2019)	JOP	The Political Weaponization of Gun Owners: The National Rifle Association's Cultivation, Dissemination, and Use of a Group Social Identity
12	Lowande (2018)	JOP	Politicization and Responsiveness in Executive Agencies
13	Magaloni and Rodriguez (2020)	APSR	Institutionalized Police Brutality: Torture, the Militarization of Security, and the Reform of Inquisitorial Criminal Justice in Mexico
14	Martin and McCrain (2019)	APSR	Local News and National Politics
15	Motolinia (2020)	APSR	Electoral Accountability and Particularistic Legislation: Evidence from an Electoral Reform in Mexico
16	Nielsen (2020)	AJPS	Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers
17	Pan and Chen (2018)	APSR	Concealing Corruption: How Chinese Officials Distort Upward Reporting of Online Grievances
18	Parthasarathy, Rao and Palaniswamy (2019)	APSR	Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies
19	Roberts, Stewart and Nielsen (2020)	AJPS	Adjusting for Confounding with Text Matching
20	Rozenas and Stukal (2019)	JOP	How Autocrats Manipulate Economic News: Evidence from Russia's State-Controlled Television

Table SI2: Topic model Usage in Three Political Science Journals (Continued)

No.	Corpus	Model	K	Outcome	Predictor
1	Tweets sent by members of Congress, different samples of the public, and media outlets	LDA	100	yes	yes
2	Articles from the *Guardian* and the *Telegraph*	STM	6	yes	no
3	Access-to-information (ATI) requests filed with the Mexican federal government agencies	LDA	20	no	yes
4	Medieval Muslim and Christian books	vMF; HTM	4 broad/60 sub	yes	no
5	Legislative summaries from European Parliaments	LDA	29	no	yes
6	Transcripts from the US members of Congress's speeches	STM	30	yes	yes
7	Parcerias (i.e., cooperation agreements) signed by PT mayors and the federal Ministry of Labor in Brazil	STM	5 topics highlighted	no	no
8	Forecast Explanations	STM	45	yes	yes
9	Petitions filed on the Local Leader Message Board in China	LDA	30	no	no
10	News articles on automakers by *Beijing Daily*, *Beijing Youth Daily*, *Beijing Morning Post*, and *Beijing Evening News*	STM	25	yes	no
11	Editorials of the *Rifleman*	STM	6	no	no
12	Contact descriptions provided in agency correspondence logs	STM	12	yes	no
13	Mexican jurisprudential theses	STM	5	no	no
14	Transcripts from TV news broadcast	LDA	15	yes	no
15	Transcripts from Mexican legislative sessions	CTM	450	yes	no
16	Text from saaid.net	STM	10	yes	yes
17	Citizen complaints in Online Sentiment Monitoring Reports in China	STM	40	yes	yes
18	Transcribed and translated proceedings of assemblies in Indian villages	STM	15	yes	no
19	Publications (expanded on Maliniak, Powers and Walter 2013) and Weibo posts (Fu, Chan, and Chau 2013)	STM	15 and 100	no	yes
20	Daily news reports from Russia's state-owned television network	STM	80	yes	no

Table SI3: Topic model Usage in Three Political Science Journals (Continued)

No.	Word Distributions	Fit Statistics	Bespoke Meaning Validation	Bespoke Validation Details	See Notes
1	yes	yes	yes	Custom website	yes
2	yes	no	no		no
3	yes	yes	no		no
4	yes	no	no		yes
5	yes	no	yes	Coded 200 documents whether crisis related	no
6	yes	no	yes	Republicans/Democrats' speeches compared to partisan topics	yes
7	yes	no	no		no
8	yes	no	no		no
9	yes	no	no		no
10	yes	no	no		no
11	yes	no	no		no
12	yes	yes	no		no
13	yes	no	no		no
14	yes	yes	no		no
15	yes	no	yes	2 topics compared to external events	no
16	yes	no	no		yes
17	yes	no	no		yes
18	yes	no	yes	Predictive validity; Survey of human observers	no
19	yes	no	no		yes
20	no	no	no		yes

Supplemental Notes

- Barberá et al. (2019)
  They provide full time-series plots on their custom website for all of their topics that can be used by the reader to compare to important external events. The site also includes sample documents for each topic so that readers can read along themselves. Finally, Barberá et al. (2019) compare the coverage of their topics to key votes in the United States Congress.

- Blaydes et al. (2018)
  The replication archive includes paragraph-long descriptions of each topic and subtopic informed by their reading of the texts.

- Dietrich et al. (2019)
  In an analysis of a topic chosen to represent women, Dietrich et al. (2019) compare their results based on the topic model to results based on a dictionary method designed to measure the same concept. By our coding rules this does not validate the meaning of the topic which is why in the table above we focused on the partisan topics analysis which is more ancillary in the original paper.

- Nielsen (2020)
  The appendix includes extensive evidence of the robustness of the results to different number of topics. The extensive replication archives also allow for the interested reader to explore more deeply.

- Pan and Chen (2018)
  As noted in a footnote of the original paper, topic models are used here primarily as an exploratory device. The eventual evidence is based on a supervised learning method. Thus, the topic model itself isn't directly evaluated for meaning, but it doesn't need to be because it served its purpose.

- Roberts et al. (2020)
  This paper presents text matching, where the goal of the topic model is to adjust for confounding by conditioning on the high-dimensional text. Roberts et al. (2020) present approaches to balance checking, conduct a hand-coding comparison of matched pairs and report example matched pairs. These are relevant validations for this particular context, but are not validations of the topic meaning in the sense we are exploring here.

- Rozenas and Stukal (2019)
  In this case, the topic model is being used as a validation check on a measure using supervised learning and thus is described more minimally (since to validate the validation would create a problem of infinite regress). In some sense this approximately fits our description of bespoke validation where the 'model is detecting something we are ex ante confident is true' (where the true thing is what is found earlier in the analysis). We ultimately decided not to consider this a case of bespoke validation because the expectation isn't fixed prior to the study. Based on the appendix, our coding for word

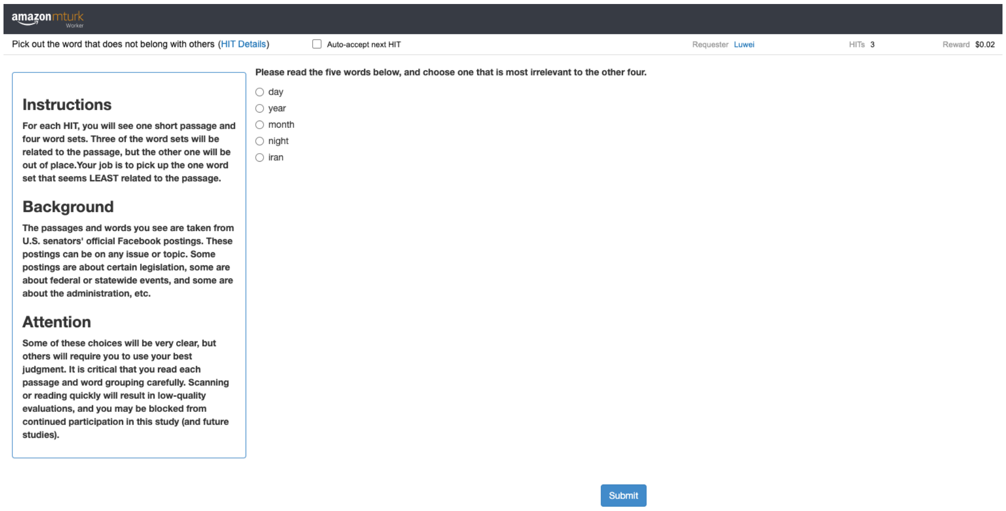
distributions was 'no', but here it is worth noting that the replication archive (which we did not consider in our coding) contained a list of FREX words for each topic. In total, this was a difficult edge case.

2 Example HITs

This section provides example HITs for The tasks we proposed in the paper, showing one screenshot for each of these tasks from the workers' view on the Mturk platform.

2.1 Example HITs for Topic Validation

Figure SI1: An Example HIT for Word Intrusion (WI)



Pick out the word that does not belong with others (HIT Details) □ Auto-accept next HIT Requester Luwei HITs 3 Reward $0.02

Instructions

For each HIT, you will see one short passage and four word sets. Three of the word sets will be related to the passage, but the other one will be out of place.Your job is to pick up the one word set that seems LEAST related to the passage.

Background

The passages and words you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

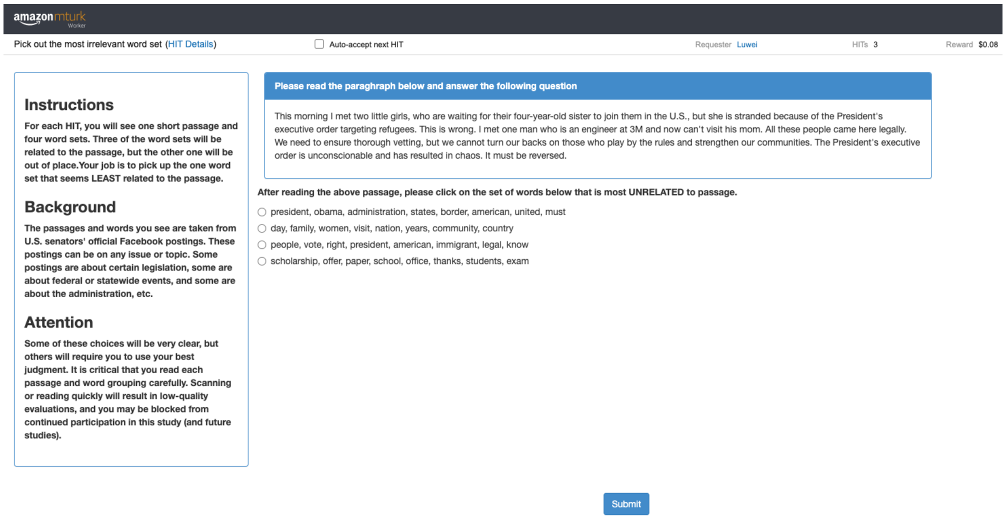
Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each passage and word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Please read the five words below, and choose one that is most irrelevant to the other four.

○ day
○ year
○ month
○ night
○ iran

Figure SI2: An Example HIT for Topc 8 Word Set Intrusion (T8WSI)



Pick out the most irrelevant word set (HIT Details) □ Auto-accept next HIT Requester Luwei HITs 3 Reward $0.08

Instructions

For each HIT, you will see one short passage and four word sets. Three of the word sets will be related to the passage, but the other one will be out of place.Your job is to pick up the one word set that seems LEAST related to the passage.

Background

The passages and words you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each passage and word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

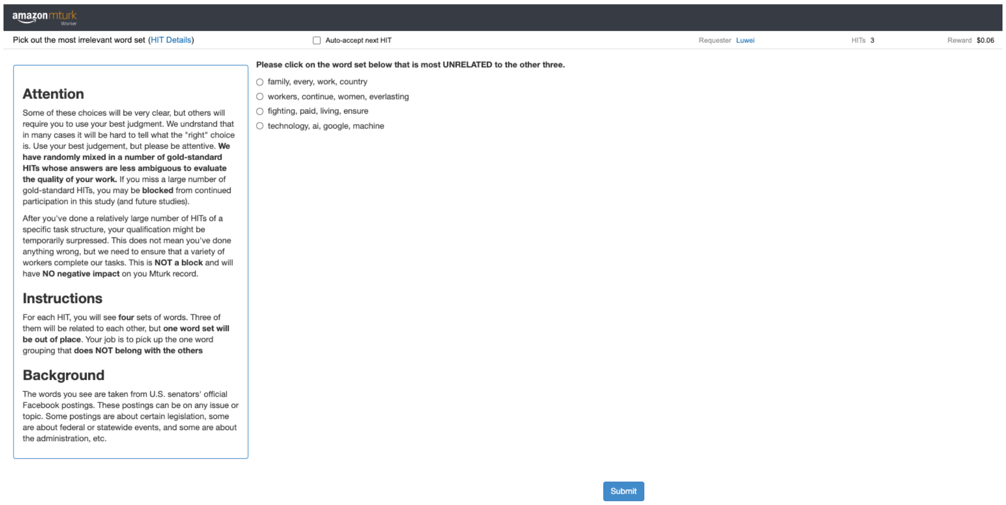
Please read the paraghraph below and answer the following question

This morning I met two little girls, who are waiting for their four-year-old sister to join them in the U.S., but she is stranded because of the President's executive order targeting refugees. This is wrong. I met one man who is an engineer at 3M and now can't visit his mom. All these people came here legally. We need to ensure thorough vetting, but we cannot turn our backs on those who play by the rules and strengthen our communities. The President's executive order is unconscionable and has resulted in chaos. It must be reversed.

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

○ president, obama, administration, states, border, american, united, must
○ day, family, women, visit, nation, years, community, country
○ people, vote, right, president, american, immigrant, legal, know
○ scholarship, offer, paper, school, office, thanks, students, exam

Figure SI3: An Example HIT for Random 4 Word Set Intrusion (R4WSI)



Pick out the most irrelevant word set (HIT Details)

□ Auto-accept next HIT

Requester Luwei

HITs 3

Reward $0.06

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. We undrstand that in many cases it will be hard to tell what the "right" choice is. Use your best judgement, but please be attentive. **We have randomly mixed in a number of gold-standard HITs whose answers are less ambiguous to evaluate the quality of your work.** If you miss a large number of gold-standard HITs, you may be **blocked** from continued participation in this study (and future studies).

After you've done a relatively large number of HITs of a specific task structure, your qualification might be temporarily surpressed. This does not mean you've done anything wrong, but we need to ensure that a variety of workers complete our tasks. This is **NOT a block** and will have **NO negative impact** on you Mturk record.

Instructions

For each HIT, you will see **four** sets of words. Three of them will be related to each other, but **one word set will be out of place**. Your job is to pick up the one word grouping that **does NOT belong with the others**

Background

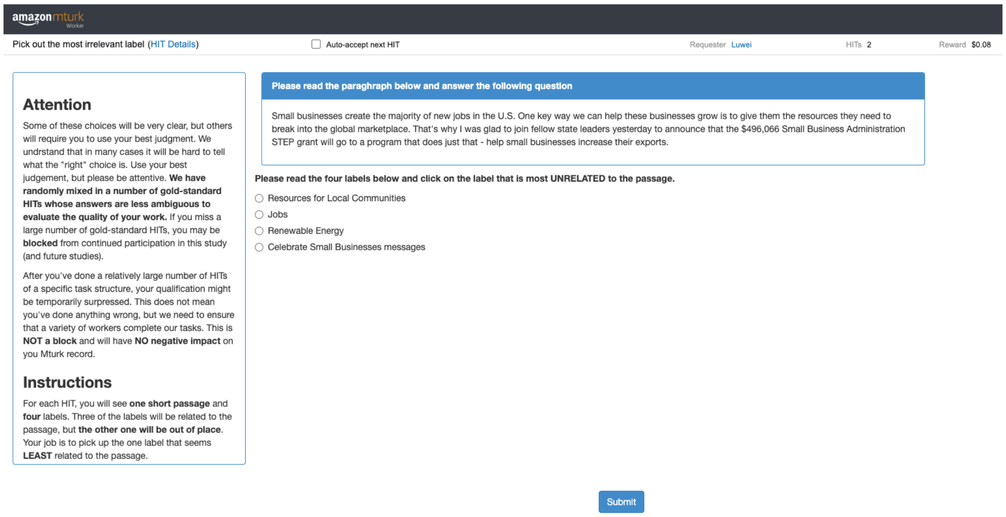
The words you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Please click on the word set below that is most UNRELATED to the other three.

○ family, every, work, country
○ workers, continue, women, everlasting
○ fighting, paid, living, ensure
○ technology, ai, google, machine

2.2 Example HITs for Label Validation

Figure SI4: An Example HIT for Label Intrusion (LI)



Pick out the most irrelevant label (HIT Details)

□ Auto-accept next HIT

Requester Luwei

HITs 2

Reward $0.08

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. We undrstand that in many cases it will be hard to tell what the "right" choice is. Use your best judgement, but please be attentive. **We have randomly mixed in a number of gold-standard HITs whose answers are less ambiguous to evaluate the quality of your work.** If you miss a large number of gold-standard HITs, you may be **blocked** from continued participation in this study (and future studies).

After you've done a relatively large number of HITs of a specific task structure, your qualification might be temporarily surpressed. This does not mean you've done anything wrong, but we need to ensure that a variety of workers complete our tasks. This is **NOT a block** and will have **NO negative impact** on you Mturk record.

Instructions

For each HIT, you will see **one short passage** and **four** labels. Three of the labels will be related to the passage, but **the other one will be out of place**. Your job is to pick up the one label that seems **LEAST** related to the passage.

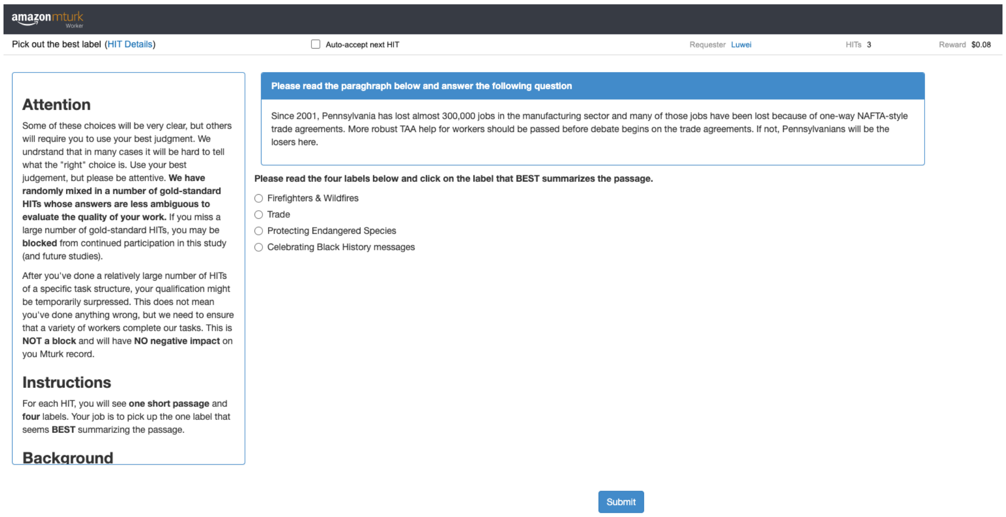
Please read the paragraph below and answer the following question

Small businesses create the majority of new jobs in the U.S. One key way we can help these businesses grow is to give them the resources they need to break into the global marketplace. That's why I was glad to join fellow state leaders yesterday to announce that the $496,066 Small Business Administration STEP grant will go to a program that does just that - help small businesses increase their exports.

Please read the four labels below and click on the label that is most UNRELATED to the passage.

○ Resources for Local Communities
○ Jobs
○ Renewable Energy
○ Celebrate Small Businesses messages

Figure SI5: An Example HIT for Optimal Label (OL)



Pick out the best label (HIT Details) □ Auto-accept next HIT Requester Luwei HITs 3 Reward $0.08

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. We undrstand that in many cases it will be hard to tell what the "right" choice is. Use your best judgement, but please be attentive. **We have randomly mixed in a number of gold-standard HITs whose answers are less ambiguous to evaluate the quality of your work.** If you miss a large number of gold-standard HITs, you may be **blocked** from continued participation in this study (and future studies).

After you've done a relatively large number of HITs of a specific task structure, your qualification might be temporarily surpressed. This does not mean you've done anything wrong, but we need to ensure that a variety of workers complete our tasks. This is **NOT a block** and will have **NO negative impact** on you Mturk record.

Instructions

For each HIT, you will see **one short passage** and **four** labels. Your job is to pick up the one label that seems **BEST** summarizing the passage.

Background

Please read the paraghraph below and answer the following question

Since 2001, Pennsylvania has lost almost 300,000 jobs in the manufacturing sector and many of those jobs have been lost because of one-way NAFTA-style trade agreements. More robust TAA help for workers should be passed before debate begins on the trade agreements. If not, Pennsylvanians will be the losers here.

Please read the four labels below and click on the label that BEST summarizes the passage.

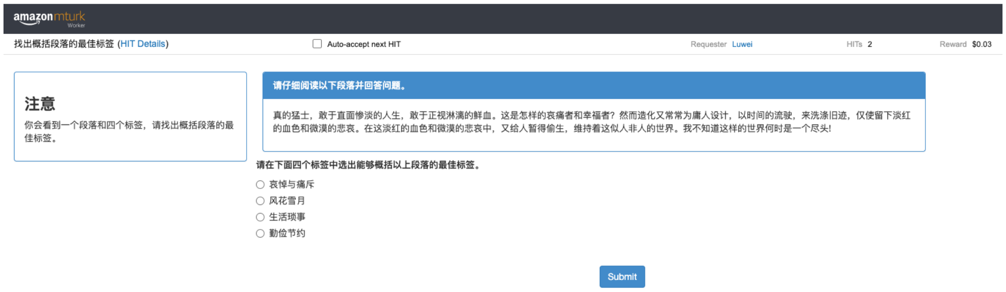
○ Firefighters & Wildfires
○ Trade
○ Protecting Endangered Species
○ Celebrating Black History messages

Submit

2.3 Example HITs in non-English (Chinese)

Our package allows researchers to post tasks in different languages with no need to change any settings. Although we didn't discuss in depth posting validation tasks for non-English corpus, we note this possibility here. Figure SI6 shows an example HIT in Chinese, posted with our R package.

Figure SI6: An Example HIT in non-English (Chinese)



找出概括段落的最佳标签 (HIT Details) □ Auto-accept next HIT Requester Luwei HITs 2 Reward $0.03

注意

你会看到一个段落和四个标签，请找出概括段落的最佳标签。

请仔细阅读以下段落并回答问题。

真的猛士，敢于直面惨淡的人生，敢于正视淋漓的鲜血。这是怎样的哀痛者和幸福者？然而造化又常常为庸人设计，以时间的流驶，来洗涤旧迹，仅使留下淡红的血色和微漠的悲哀。在这淡红的血色和微漠的悲哀中，又给人暂得偷生，维持着这似人非人的世界。我不知道这样的世界何时是一个尽头!

请在下面四个标签中选出能够概括以上段落的最佳标签。

○ 哀悼与痛斥
○ 风花雪月
○ 生活琐事
○ 勤俭节约

Submit

3 Alternative Task Designs and Results

In this section, we present two variants of the Random 4 Word Set Intrusion (R4WSI) model validation task and three other task structures for label validation. In each case, we experimented with these approaches but decided not to recommend them as the primary choices for the reasons discussed below. However, for the sake of transparency we report the procedures and results here as they may be of interest to future researchers.

3.1 R4WSI Tasks with Documents

We experimented with two variations of R4WSI task with the major difference that it shows coders a document in addition to the word sets. R4WSI-Random, chooses a random document from the 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, which are drawn following the same procedure as we did for R4WSI in the main text: the three non-intruder word sets are all chosen from the same topic (the topic most associated with the document) and the intruder set from a different topic.

The second variant is identical to the above R4WSI-Random task with one exception: the documents shown are always one of the ten most representative documents for a specific topic. Thus the only difference between these two variants is how the document is selected. For convenience, we call it R4WSI-Representative. Both variants are summarized in Table SI4.

Table SI4: Additional Task Structures for Coherence Evaluations

	A. Text Presented	B. Question Asked	C. Options Given
R4WSI-Random	A randomly selected document	After reading the above passage, please click on the word set below that is most UNRELATED to the passage.	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
R4WSI-Representative	A document randomly selected from the top ten most representative documents for a given topic	After reading the above passage, please click on the word set below that is most UNRELATED to the passage.	Three word sets (each containing four randomly selected words) from the top twenty high-probability words of one topic and one word set (the intruder) randomly selected from the top twenty high-probability words of another topic

As in the main text, we posted 500 HITs twice, summing to 1000 HITs, for each task-model combination. Note that the R4WSI-Representative task was implemented with four different structural topic models from Models 1-5 in the main text:[1]

1. a 20 topic model (same as Model 2) limited to one EM iteration (which keeps the model from properly converging) with random initialization;

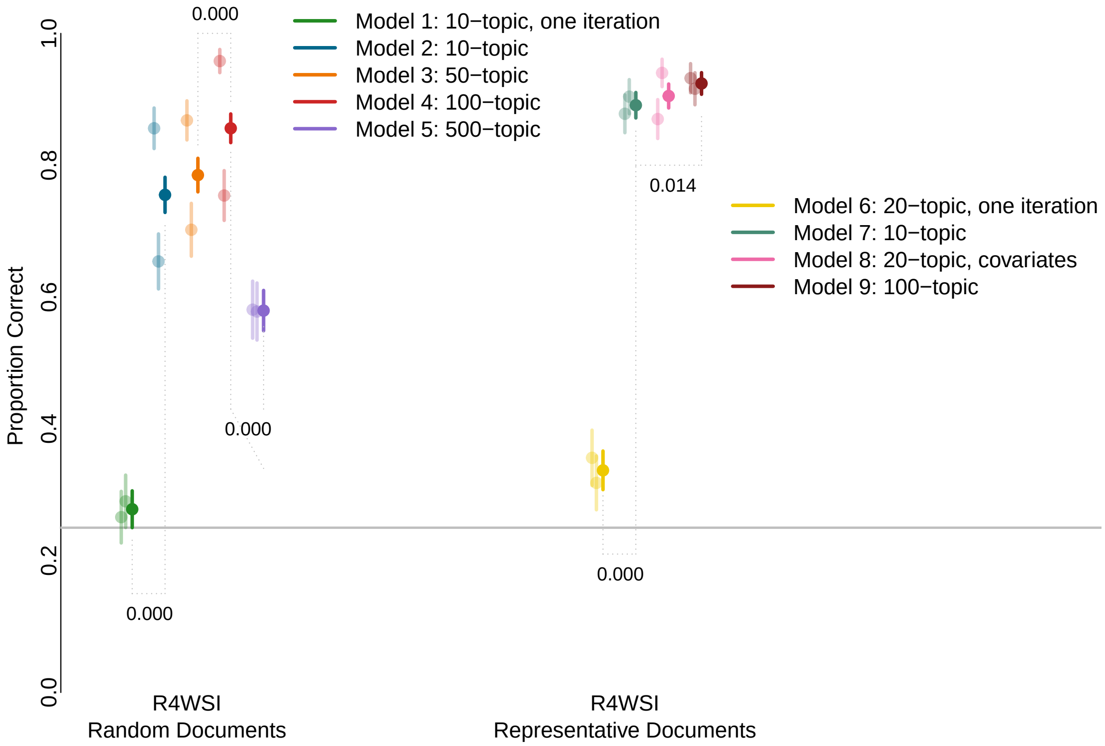
[1] Models 6-9 were from an earlier version of the paper where we fit topic models without holding out any documents.

2. a model with 10 topics and no covariates;
3. a model with 20 topics and adding the senators' names as a covariate, and;
4. a model with 100 topics.

The result can be seen in Figure SI7. The main problem is that the R4WSI-Random scores are not reliable across runs for Model 2-4. The scores differed dramatically even though each run includes identical tasks. We were unable to improve the reliability for this task structure even after multiple attempts to recruit high quality workers.

In R4WSI-Representative, the two identical trials for Model 8 are significantly different indicating lower levels of reliability. Additionally, this task structure provided much lower power to distinguish between models. We think this is because the R4WSI-Representative task only uses the clearest documents and thus results in a ceiling effect.

Figure SI7: Results for Coherence Evaluations



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.

3.2 Two Alternative Label Validation Tasks

We have also designed two alternative label validation tasks that rely only on the labels and high probability words. The *Word Set Intrusion* task copies the R4WSI described in the topic validation section in the main text, simply adding the label at the top of the task. An example is shown in Panel (a) of Figure SI8. The *Optimal Label for Word Sets* (OLW) task presents a topic, represented by the eight highest-probability words, and four labels. One label is for the relevant topic and the others are for other topics in the model but within the specified broad category. An example is shown in Panel (b) of Figure SI8. Both task structures are summarized in words in Table SI5.

Table SI5: Additional Task Structures for Label Validation

	A. Text Presented	B. Question Asked	C. Options Given
WSI	A label[a]	Please click on the word set below that is most UNRELATED to the above label.	Three labels for the top three high-probability topics and one label for other economy-related topics
OLW	The eight highest-probability words for a topic	Please read the four labels below and click on the label that BEST summarizes the word set.	One label for the highest-probability topic and three labels for other economy-related topics

[a] The label for one of ten economy-related topics in this empirical illustration.

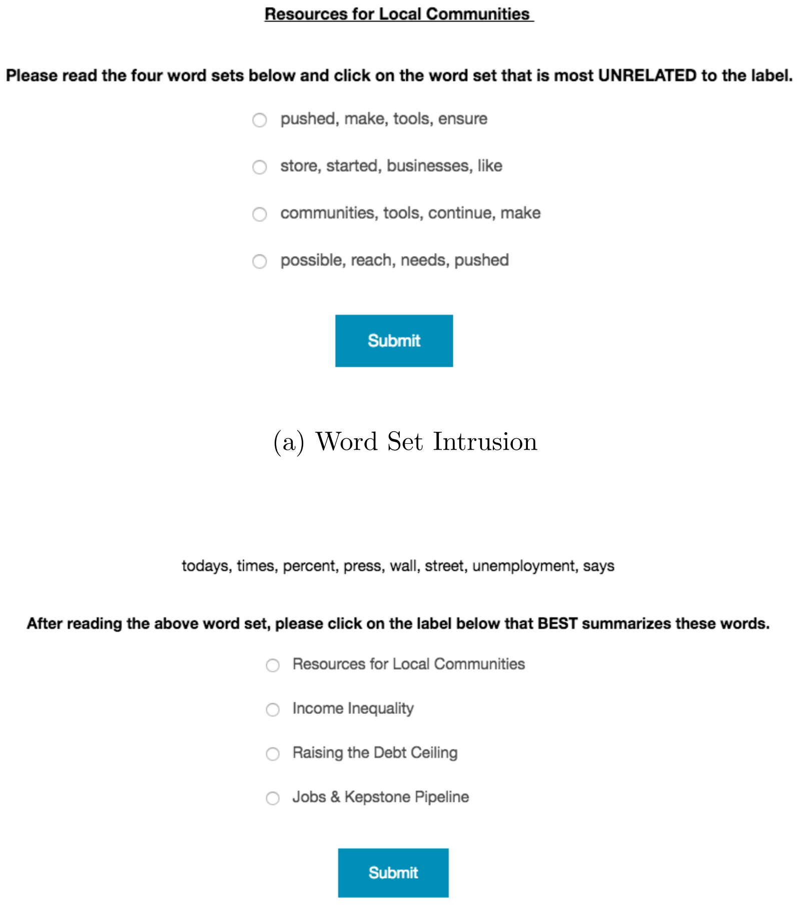
Past experience teaches that word sets are easier to associate with a topic than an actual document since language out of context is easier to re-interpret. This can make associating broad concepts with word sets easier than with documents, but often at the expense of verisimilitude. Thus, it should not be too surprising to find that validations based primarily on word sets will give less clear guidance.

To test these task structures we followed the same basic procedures as discussed in the main text. For each task/coder combination we created 500 tasks that were coded by trained workers on AMT.[2] The results from 4,000 high quality HITs are shown in Figure SI9. However, unlike the main text, these task structures were tested with Model 9 described above. In addition, we created "cursory coder" and "careful coder" labels for only 10 economics-related topics (although we followed the same basic procedures as in the main text to create these competing label sets).

The expected result that the "careful coder" labels were superior, however, was not confirmed by the tasks based around word sets (WSI and OWS). Ex ante this makes sense since the "cursory coder" labels were based only on the word sets to begin with. Given that the "careful coder" labels were intentionally designed to more accurately represent the underlying meaning of the topics as they appear in the actual documents, we infer that label validation should be based on documents rather than word sets alone. In all, based on these results we would recommend that labels be evaluated using the LI and OL.

[2] Workers were paid 0.02/HIT for the OLW task, 0.03/HIT for the WSI task

Figure SI8: Example HITs for Alternative Label Validation Tasks



Resources for Local Communities

Please read the four word sets below and click on the word set that is most UNRELATED to the label.

○ pushed, make, tools, ensure

○ store, started, businesses, like

○ communities, tools, continue, make

○ possible, reach, needs, pushed

(a) Word Set Intrusion

todays, times, percent, press, wall, street, unemployment, says

After reading the above word set, please click on the label below that BEST summarizes these words.

○ Resources for Local Communities

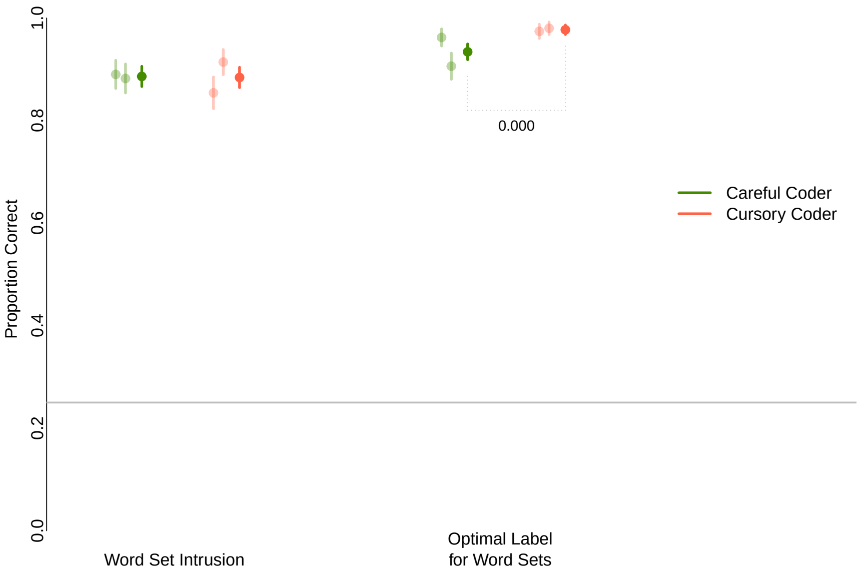
○ Income Inequality

○ Raising the Debt Ceiling

○ Jobs & Kepstone Pipeline

(b) Optimal Label for Word Sets

Figure SI9: 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.

3.3 Selecting among Alternative Labels?

A further variation on our labeling task structure focused on testing alternative labels for individual topics (instead of competing label *sets*). Specifically, we presented our good labels (the "careful coder" labels) to two research assistants and asked each of them to come up with three closely related concepts without looking at the top words or documents.[3] We then selected three out of these six alternative labels, mixing them with the original "careful coder" labels to construct the tasks. In this way, the alternative labels would, by definition, be similar in meanings. However, we wanted to test the expectation that, *on average*, our

[3] The instruction we gave the coders are as follows:

> Some background information: We are researchers who want to know what the US senators are talking about online. We scraped the senators' Facebook pages (all posts) from 2008 to 2017 (10 years) and ran a topic model to discover the topics from this corpus. We selected 10 topics and labeled them below. You job is, by just looking at these labels, to come up with 3 alternative labels for each of the 10 labels.Your alternative labels should be related concepts/phrases that could potentially summarize a similar set of documents. For example, for the label "Terrorism," three alternative labels could be "Counter-Terrorism," "Domestic Security", and "Violent Groups".
>
> The labels: Equal Pay for Women; Healthcare/Reproductive Rights; Agriculture; Student Loan/Debt; Drug Abuse; Higher Education/Job Training; Wall Street/Financial Sector; Government Shutdown/Congressional Budget; Obamacare/Tax Policy; Deficits/Debt/Budget

"careful coder" labels will outperform alternatives.

To be clear, this alternative task structure is identical to the *Optimal Label* task in the main text. The difference is that coders were not choosing between labels of different topics, but instead choosing directly between *four alternative labels for the same topic.* Like other experiment, we post 500 HITs (50 HITs for each topic) twice. The results (including the four alternatives for each topic) are shown in Figure SI10.

Although the sample size allocated to each topic is small (50 HITs) for each topic, the results are fairly consistent from trial to trial. Our preferred labels perform best in eight out of the ten cases for one run and seven out of the ten cases for a second run. In the two topics ("equal pay for women" and "drug abuse") where the alternative labels outperform the original labels, revisiting the document sets indicated that these are indeed good labels for those specific topics.

Although these results seem promising, our concern is that the task structure will overly favor specific labels when available leading to confused results. Conceptually, we might imaging having a topic 'A' that includes sub-components '1' '2' and '3.' For example, assume a topic model had a broad "economic policy" topic that encapsulated the "government shutdown/congressional budget", "Obamacare/tax policy", and "deficits/debt/budget" topics discussed above.[4] For each document, coders would then choose between: economic policy, government shutdown/congressional budget, Obamacare/tax policy, and deficits/debt/budget. While economic policy might better represent the overall topic, each document will fit in better with the more specific label. Since workers do not have access to the whole document set or other model outputs, they will (incorrectly) tend to choose the more specific labels even though they are inappropriate for the broad topic.

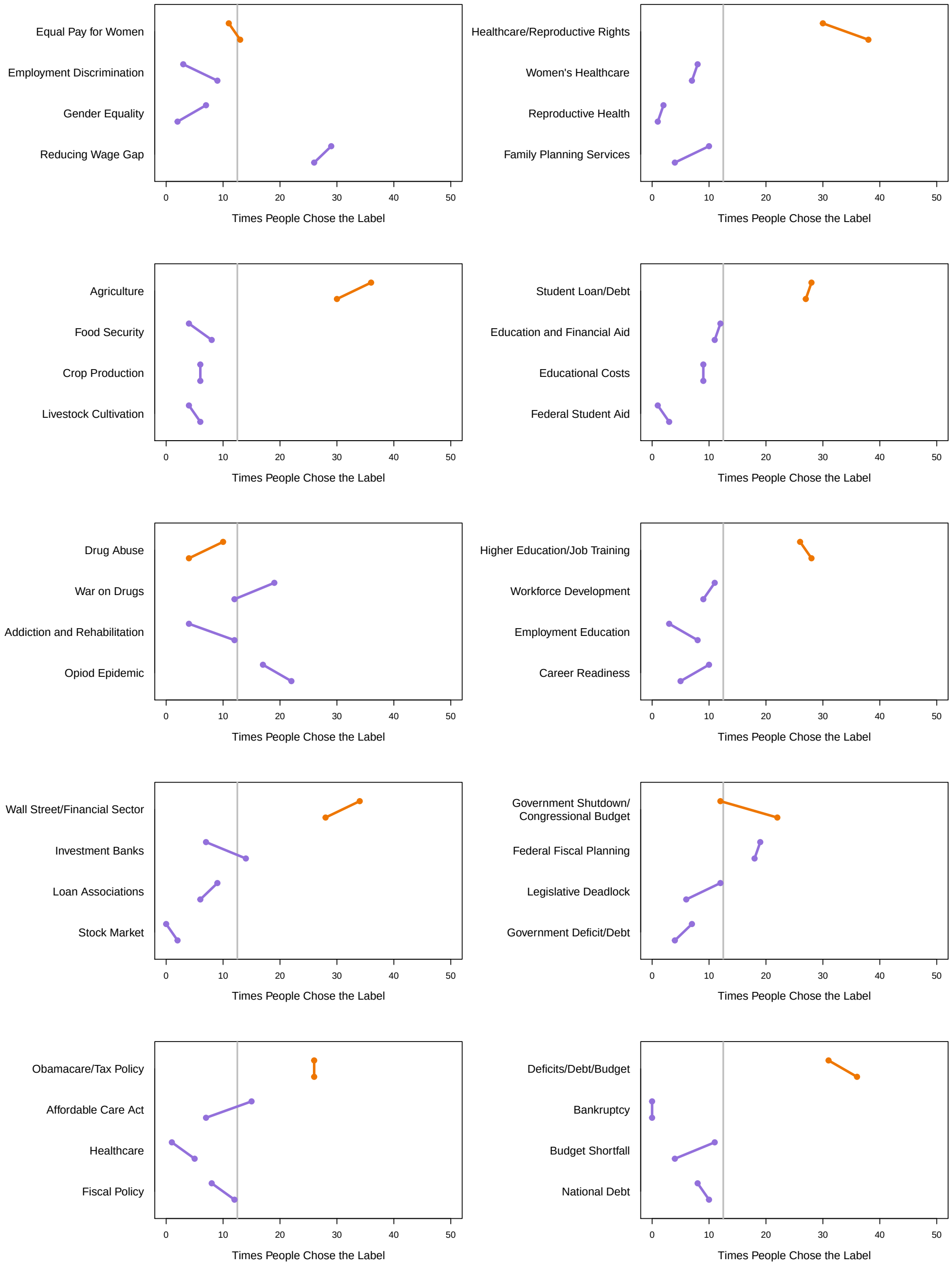
To test this, we re-ran the above experiment but collapsed these three categories and presented respondents with exactly this choice set. The results confirmed our suspicion that coders would strongly prefer the more specific labels. For the 150 relevant tasks, coders chose the broader "economic policy" topic only 20 times using the more specific labels 86% of the time. Meanwhile the other topics were selected nearly an equal proportion of the time (government shutdown/congressional budget was chosen 47 times, deficits/debt/budget chosen 45 times, and Obamacare/tax policy 48 times). A naive analyst looking at these results would conclude that all of the labels are poor, but that the "economic policy" label is worst (even though it is the most appropriate for this setting).

Thus, while this approach may be useful for helping researchers compare competing topic labels, we are hesitant to recommend its widespread adoption since the outputs seem inherently ambiguous. Low scores for a specific label may indicate a poor conceptual mapping but may also simply indicate that the competing topics include sub-components. That is, coders may regularly choose the more specific labels for any given document and never the broader "theme" appropriate for the topic.

Further, in the context of topic models, the goal is not simply to place the "right" label on any given topic but to also ensure that the conceptual category is internally coherent and distinct from neighbors. That is, we are not typically labeling topics in isolation but as a set. The idea is not simply to find the best label for one specific topic, but to find a set of

[4] Collapsing adjacent topics into an overarching topic occurs commonly, for instance, when reducing the number of topics in a model.

Figure SI10: Results for the Alternative Label Validation Task



Note: The "careful coder" labels are in orange and the alternative labels are in purple. The two dots represent two identical trials. The gray vertical lines represent random guessing.

labels that we can clearly distinguish. Thus, this task structure may be useful for testing between competing conceptual labels but only if care is taken to avoid this decomposition fallacy.

4 A Checklist for Crowdsourced Topic Validation

In this section, we provide succinct checklist for future researchers who intend to use topic models to measure concepts. This list is intended to be descriptive of the steps we have found to be useful in practice rather than prescriptive. We doubt there is any finite set of recommendations that will be appropriate in all settings and this should not be read as a narrow set of requirements but rather as a place to start when thinking through these issues. Still it may be useful to provide a more coherent summary of our approach in one location.

When validating topic coherence, here are the steps we follow.

1. Know your corpus and read it in depth. There is no substitute for being familiar with the content of your documents.
2. Fit several models, then use fit statistics and qualitative reading to select one (or several) you wish to consider further. At a minimum this should be done by reading multiple documents for each topic and considering model outputs such as high probability and FREX words. However, in our experience it is easy to get fooled by this alone. Further steps may be to examine model outputs such as fit statistics (e.g., held-out log likelihoods) and surrogate scores.
3. Consider the effects of alternative pre-processing steps, starting values, and tuning parameters. Themes and topics that are highly sensitive to these choices may be harder to uncover reliably with unsupervised methods.
4. Validate the chosen model(s) using the crowdsourced procedure:
   (a) Create an Amazon requester account. You might consider also setting up a worker account to see your own tasks as a worker.
   (b) Generate tasks using our package, be sure to look at them in detail and see if you can do the tasks yourself. If not, you may want to try more models or improve/refine training. Look for common issues or areas of confusion and think about how to set up your training module to explain these issues clearly.
   (c) Design the training module, explain the background to the workers in detail, and make sure the example tasks cover a wide range of your topics. You will want to include several example HITs that illustrate the easier cases but also some that specifically illustrate edge cases or unusual features of your data. The final step is to include a "test" to screen out poor workers or bots and ensue that only qualified workers engage in your tasks. Keep in mind that harder tests may improve worker quality, but may also decrease the pool of workers who can pass it (and may require higher compensation rates).
   (d) Post tasks with the free Mturk sandbox version to make sure everything looks right before spending money on Mturk.
   (e) Remember that the quality of the work is vital. Think carefully about the requirements you want workers to meet (we recommend using master workers). As data comes in, pay attention to the work as it arrives. Be aware to abnormal activities (e.g., submitting a task every second, missing a bunch of gold-standard HITs) and stop these workers early. Workers won't mind so long as you don't "reject" their work (which can have huge effects on their reputation scores and

future income). Simply removing their permission to continue work is enough.

   (f) We recommend that 1/10 HITs should be a "gold standard" HIT where the answer is clear and unambiguous. These can be generated in advance and chosen by the researcher. Our software allows you to include these easily while not contaminating your data. We removed permission from any worker who misses more than two gold HITs.

   (g) Don't let any one worker complete too much of your work. We recommend allowing any one worker to only complete 20% of your HITs. This makes it easier to diagnose if there is a problem with your model or the specific workers who picked up your HITs.

   (h) Decide on a threshold in advance where you will simply discard all of the work from a "bad" worker. We discarded work from two workers who missed four gold HIT.

5. We compared models using a simple two-sample difference in proportions test, although we do not think it is always necessary to evaluate more than one model.
6. Report the results of the validation including mean, standard error, and sample size. Authors might also report their training module (or a summary of the module) and basic information about any issues with worker quality. While there is no absolute minimum standard for what score is "high enough," the results should clearly be better than random guessing. This should be only one validation reported along with other forms of model validation.

When validating topic labels, here is our advice.

1. Think about the use-case: Are you looking for a broader concept (e.g., politics) that is captured a set of topics? Or by just one topic? Remember that fine-grained distinctions will by their nature be harder to detect, less reliable, and noisier. If high levels of precision/accuracy are crucial to your design, you may want to go with broader themes (clusters of topics) rather than single topics. Try fitting models with a larger number of topics. It may be easier to combine several smaller, coherent topics rather than try and force meaning onto a single large topic.
2. Label the topics by carefully reading the high probability words, frequent and exclusive words, many representative documents, and other relevant information. Preferably, have multiple people do so independently and reach a consensus. Our method is about validating labels, but you need to take time and give careful thought to developing them in the first place.
3. Label the relevant topics both on and off the concept (e.g., about politics and not about politics). However, in our experience some topics will not be interpretable, mixing themes or focusing on unimportant features of language such as style. Do not try and force a meaning onto your topics where they do not exist. You can simply exclude incoherent topics from your further validation (although they should also be excluded from your eventual analysis).
4. Follow the above tips, now making sure the training module captures the concept you intend to measure. Keep your attention on the worker quality.

5. The LI and OL task structures are both appropriate. The OL task is better suited for single-membership models or where you tend to think of documents as mostly being in one conceptual bin. Where the mixture of topics is more important, the LI task will be more suitable.
6. Report the results with other forms of validation discussed in the main text. You can show the overall score, but may want to focus on the "within" and "between" results when using the broader categorization approach discussed in the main text. The reported results should include the mean score, standard error, and sample size. Authors might also report their training module (or a summary of the module) and basic information about any issues with worker quality.

5 Software and Working Example

Along with the paper, we provide free-to-use software to facilitate the topic and label validation procedures for future researchers. This section serves as a user manual.[5]

5.1 R Package: `validateIt`

The R package, `validateIt`, to implement the methods we proposed in the paper is currently being prepared for submission to the *Comprehensive R Archive Network* (CRAN). Users can install the most recent development version from `Github` using the `devtools` package. First, users would have to install `devtools` using the following code. Note that this step only has to be done once:

```
if(!require(devtools)) install.packages("devtools")
```

Then you can load the package and use the function `install_github`

```
library(devtools)
install_github("Luwei-Ying/validateIt", dependencies = TRUE)
library(validateIt)
```

Note that this will install all the packages suggested and required to run our package. It may take a few minutes the first time, but this only needs to be done on the first use. In the future users can update to the most recent development version using the same code.

5.2 Fitting Candidate Topic Models and Labeling

From here, we provide example code to post validation tasks. It shows the entire procedure of validating topic model outcomes using our proposed methods. This procedure consists five major steps: 1) Fitting Candidate Topic Models and Labeling; 2) Training and Certification; 3) Preparing Tasks Locally; 4) Posting Tasks to Mturk and Maintaining Them Online; and 5) Retrieving and Evaluating Results.

[5] This section features simplified instructions for using the proposed method. It is for illustration purposes. To view the original files and code for our study, please refer to the replication materials.

Researchers need to fit several candidate topic models to be validated later. To illustrate the methods, we fit 5 topic models using the `R` packages `stm` with 500 topics, 100 topics, 50 topics, 10 topics, 10 topics but restricted to only 1 EM iteration, as described in the main manuscript. Topic models generated by other software will also work as long as the users keep track of a) the documents in the topic model; b) the vocabulary in the topic model; c) the matrix of topic conditional word probabilities, usually known as the beta matrix; and d) the matrix of the topic distribution for each document, usually known as the theta matrix.

While this section does not intend to provide detailed instructions on fitting topic models, we want to point out that we did not stem the words. We also randomly select 10% of the documents (16364 out of 163642 documents) and **held out** 50% of the tokens in these documents so we will be able to compare the results from our methods with held-out likelihood. Code for this procedure is included below.

```
library(stm)
# step 1: pre-process, obtain the stm object
# define the customized stop words outside of the function
customstopwords <- c("senator", "senate", "sen", "facebook", "please", "today",
                     "will", tolower(state.name), tolower(state.abb), "rhode",
                     "hampshire", "jersey", "york", "carolina", "dakota")
docs = textProcessor(documents = as.character(Senate$text),
                     metadata = Senate,
                     lowercase = TRUE,
                     removestopwords = TRUE,
                     removenumbers = TRUE,
                     removepunctuation = TRUE,
                     stem = FALSE,
                     wordLengths = c(3, 20),
                     language = "en",
                     striphtml = TRUE,
                     customstopwords = customstopwords)
stmPrep <- prepDocuments(docs[[1]], docs[[2]], meta = docs$meta)
save(stmPrep, file = "Corpus/stmPrep.Rdata")

# step 2: heldout some words from some documents
heldout <- make.heldout(stmPrep$documents, stmPrep$vocab, seed = 123)
save(heldout, file = "Corpus/heldout.Rdata")
documents <- heldout$documents
vocab <- heldout$vocab

# step 3: fit stm with different topic numbers, i.e., different k
stm <- stm(documents, vocab, k = 10)
save(stm, file = "Models/unstemmed/stm.Rdata")
```

To label the topics, we strongly recommend users to follow the practice of the “Careful Coder”: carefully read the high-probability words and frequent & exclusive words (FREX), as well as fifty representative documents per topic. Ideally, future researchers will have more than one coder to independently label the topics so they can assess inter-coder reliability.

5.3 Training and Certification

First of all, researchers need to register a Mturk requester account here: `https://www.mturk.com/`.

For each of the task structures, Word Intrusion (WI), Top 8 Word Set Intrusion (T8WSI), Random 4 Word Set Intrusion (R4WSI), Label Intrudion (LI), and Optimal Label (OL), researchers need to make a training module to introduce the background and get online workers familiar with the tasks. While doing so, we provide five practice questions, where people's answers are NOT scored, then eight test questions. The workers need to get 7 or more out of the 8 questions correct to receive the qualification.

The training modules are written in .xml files, each with a separate .xml file coding the answers. Section 6.1 - 6.5 present the training modules used for this paper. For .xml formatting, see the example files "Question.xml" and "Answer.xml" in our replication package.

We then create the qualifications on Mturk:

1) Specify the Mturk options.

```
library(pyMTurkR)
Sys.setenv(AWS_ACCESS_KEY_ID = AWS_ACCESS_KEY_ID)
Sys.setenv(AWS_SECRET_ACCESS_KEY = AWS_SECRET_ACCESS_KEY)

# change sandbox = F when ready to run on MTurk
options(pyMTurkR.sandbox = T)

# use this to test that the pyMturkR settings are correct
AccountBalance()
```

2) Read in the .xml file for questions and answers.

```
TestQuestions <- paste0(readLines("T8WSIQuestion.xml", warn = FALSE),
                        collapse = "")
TestKey <- paste0(readLines("T8WSIAnswer.xml", warn = FALSE),
                  collapse = "")
```

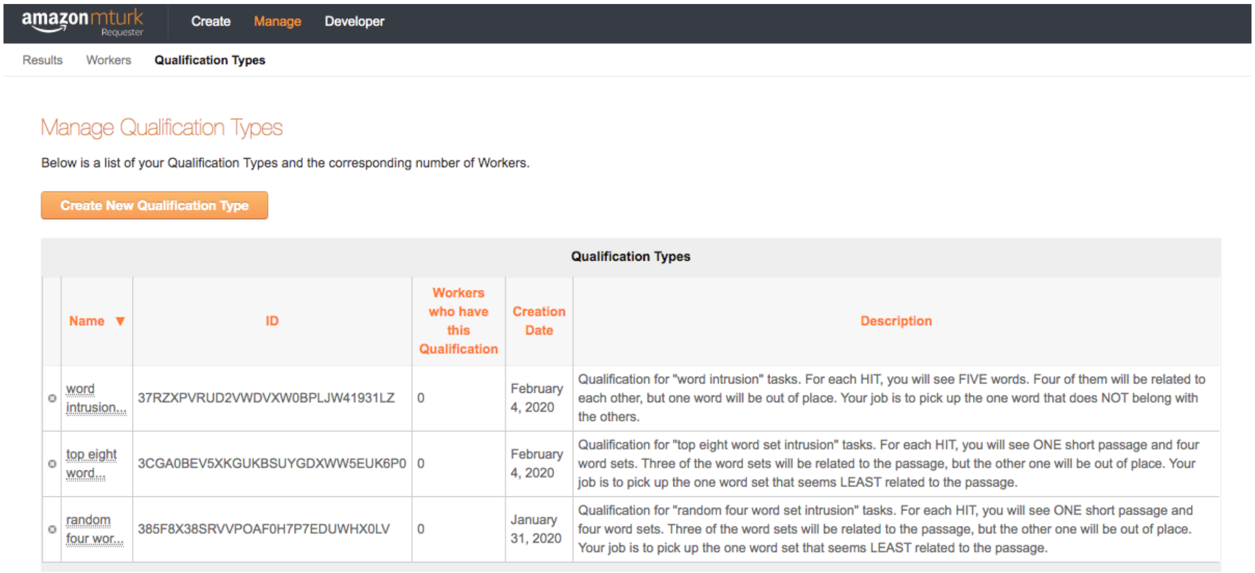
3) Create the qualification.

```
T8WSIQual <- CreateQualificationType(
             name = "top eight word set intrusion qualification",
             description = "Qualification for "top eight word
                           set intrusion" tasks.",
             status = "Active", # allows qual to remain active for users
             test = TestQuestions, # pass questions for test
             test.duration = 60 * 60, # test duration, in seconds
             retry.delay = NULL, # how long until worker can retry test;
                                 NULL means never
```

```
            answerkey = TestKey)
```

The created qualification can be seen on the requester's dashboard under the "Manage" tab after a few minutes.

Figure SI11: Check the qualification, which should appear on the dashboard



Manage Qualification Types

Below is a list of your Qualification Types and the corresponding number of Workers.

Qualification Types

	Name ▼	ID	Workers who have this Qualification	Creation Date	Description
⊗	word intrusion...	37RZXPVRUD2VWDVXW0BPLJW41931LZ	0	February 4, 2020	Qualification for "word intrusion" tasks. For each HIT, you will see FIVE words. Four of them will be related to each other, but one word will be out of place. Your job is to pick up the one word that does NOT belong with the others.
⊗	top eight word...	3CGA0BEV5XKGUKBSUYGDXWW5EUK6P0	0	February 4, 2020	Qualification for "top eight word set intrusion" tasks. For each HIT, you will see ONE short passage and four word sets. Three of the word sets will be related to the passage, but the other one will be out of place. Your job is to pick up the one word set that seems LEAST related to the passage.
⊗	random four wor...	385F8X38SRVVPOAF0H7P7EDUWHX0LV	0	January 31, 2020	Qualification for "random four word set intrusion" tasks. For each HIT, you will see ONE short passage and four word sets. Three of the word sets will be related to the passage, but the other one will be out of place. Your job is to pick up the one word set that seems LEAST related to the passage.

4) Save the qualification in case changes need to be made in the future.

```
save.image("Qualifications.RData") # you may need to set file path
```

5) Update a qualification if needed, e.g., fix a typo.

```
load("Qualifications.RData")
T8WSIQual <- UpdateQualificationType(
             qual = T8WSIQual$QualificationTypeId, # keep same qualification id
             description = T8WSIQual$Description,
             status = "Active", # manually set status
             test = TestQuestions, # update test
             test.duration = 60 * 60, # manually set test duration
             retry.delay = NULL, # manually set retry delay
             answerkey = TestKey) # update answer key
```

5.4 Preparing Tasks

Researcher will prepare tasks locally before sending them to Mturk.

5.4.1 Preparing Tasks for Topic Validation

For topic validation, our package, `validateIt`, requires a) the documents in the topic model (`docs`); b) the vocabulary in the topic model (`vocab`); c) the beta matrix in the topic model (`beta`); and d) the theta matrix in the topic model (`theta`).

In the main manucript, we fit *unstemmed* models. The `combMass()` function takes in a `stm` output and combine mass for words with the same root. The output, `newMass`, is a list of two: the combined vocabulary matrix where words are completed to the most frequent form in that specific topic (`newvocab`) and the combined beta matrix in correspondance with the combined vocabulary (`newbeta`). This step is not required in using our crowd-sourcing methods. We see this procedure as analogous to fitting a stemmed topic model but retain the complete form of words.

```
newMass <- combMass(stm)
```

Now, researchers start formatting the tasks:

1) The `validateTopic()` function creates tasks of the desired type and number. Below is the example code for T8WSI, the structure of which task is the most complicated. We have left out the documents where a portion of words have been held out. The default threshold is 20, meaning that our algorithm draws from the pool of top 20 highest probability words.

```
T8WSItasks <- validateTopic(type = "T8WSI",
                            n = 500,
                            text = stmPrep$meta$post_text[-heldout$missing$index],
                            vocab = newMass[[1]],
                            beta = newMass[[2]],
                            theta = stm50k$theta[-heldout$missing$index,],
                            thres = 20)
```

For WI and R4WSI tasks, leave out the `docs` and `theta` arguments. The default threshold is still 20.

```
# WI
WItasks <- validateTopic(type = "WI",
                         n = 500,
                         vocab = newMass[[1]],
                         beta = newMass[[2]],
                         thres = 20)


# R4WSI
R4WSItasks <- validateTopic(type = "R4WSI",
                            n = 500,
                            vocab = newMass[[1]],
                            beta = newMass[[2]],
                            thres = 20)
```

All word-drawing processes are probability-based (mass-based). Specifically, in preparing WI tasks, the function a) orders the word probability (beta[k,]) for each topic k, then b) randomly draws 4 words from top 1 to `thres` words **based on their corresponding probabilities in k**, and then c) randomly draws 1 intruder word from top 1 to `thres` words from another topic, $\neg k$ **based on it corresponding probability in $\neg k$**.

In preparing R4WSI, the function a) for each topic k, randomly draws 12 words from top 1 to `thred` words **based on their corresponding probabilities in k** and randomly assign them to 3 different word sets, then b) randomly draws 4 intruder words from top 1 to `thred` words from another topic, $\neg k$ **based on their corresponding probabilities in $\neg k$**.

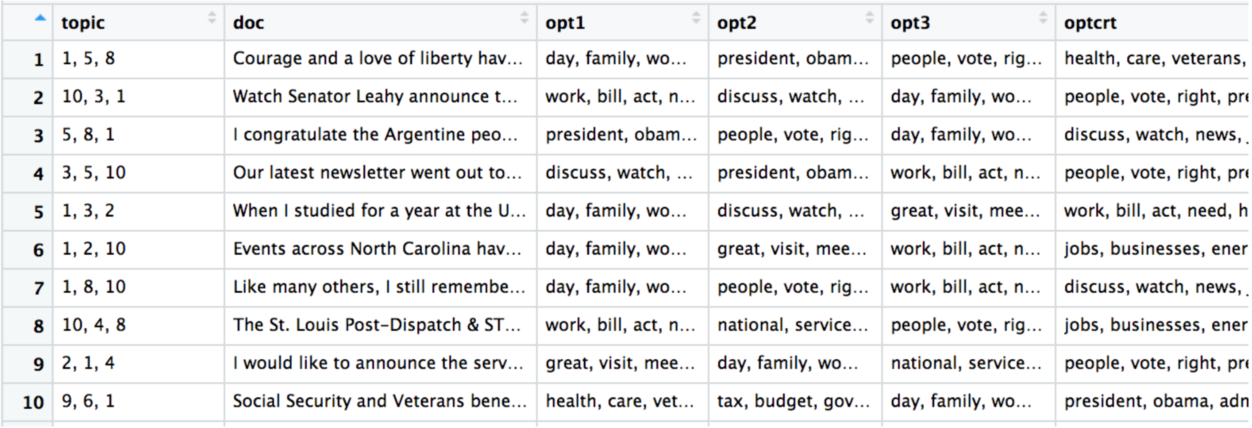
In preparing T8WSI, the function a) uses `vocab` and `beta` to find the top 8 words for each topic k, then b) randomly sample a document from the document pool (`docs`) and calculate the top 3 high probability topics associated with that document using `theta`, and then c) sample an intruder topic from other lower probability topics. Notice that the word drawing procedure here is not probability-based as the function is always looking for the top 8 words regardless of their mass.

The output while specifying "T8WSI" looks as below, where each row represents a task. Each task contains an indicator of the topic (`topic`), a randomly drawn document associated with that topic (`doc`), three non-intruder word sets (`opt2 - opt3`), and an intruder word set (`optcrt`).

2) In addition to generating tasks for validation, we suggest preparing a series of gold-standard HITs to monitor the quality of the works in the future. These gold-standard HITs are the "easy" tasks whose answers are unambiguous.

While preparing gold-standard HITs for the current paper, we select 50 HITs for each task structure from pilot runs on Mturk where two different workers have agreed on the answers in two identical rounds. These selection processes were **not random** and we also modified the documents and words to make the answers even clearer. Future researchers could simply generate more tasks than they need and hand pick the gold-standard HITs

Figure SI12: Output of `validateTopic()`



	topic	doc	opt1	opt2	opt3	optcrt
1	1, 5, 8	Courage and a love of liberty hav...	day, family, wo...	president, obam...	people, vote, rig...	health, care, veterans,
2	10, 3, 1	Watch Senator Leahy announce t...	work, bill, act, n...	discuss, watch, ...	day, family, wo...	people, vote, right, pr
3	5, 8, 1	I congratulate the Argentine peo...	president, obam...	people, vote, rig...	day, family, wo...	discuss, watch, news,
4	3, 5, 10	Our latest newsletter went out to...	discuss, watch, ...	president, obam...	work, bill, act, n...	people, vote, right, pr
5	1, 3, 2	When I studied for a year at the U...	day, family, wo...	discuss, watch, ...	great, visit, mee...	work, bill, act, need, h
6	1, 2, 10	Events across North Carolina hav...	day, family, wo...	great, visit, mee...	work, bill, act, n...	jobs, businesses, ener
7	1, 8, 10	Like many others, I still remembe...	day, family, wo...	people, vote, rig...	work, bill, act, n...	discuss, watch, news,
8	10, 4, 8	The St. Louis Post-Dispatch & ST...	work, bill, act, n...	national, service...	people, vote, rig...	jobs, businesses, ener
9	2, 1, 4	I would like to announce the serv...	great, visit, mee...	day, family, wo...	national, service...	people, vote, right, pr
10	9, 6, 1	Social Security and Veterans bene...	health, care, vet...	tax, budget, gov...	day, family, wo...	president, obama, adn

they deem as clear from the excessive tasks. An example of T8WSI gold-standard HITs reads like this:

> *The Keystone XL pipeline represents not only thousands of jobs and growth for the nation's economy, but also a big step toward American energy independence. We can become energy independent in America within five to seven years, but we must commit to moving forward with important projects like the Keystone XL pipeline.*

After reading the above passage, please click on the set of words below that is most unrelated to the passage.

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

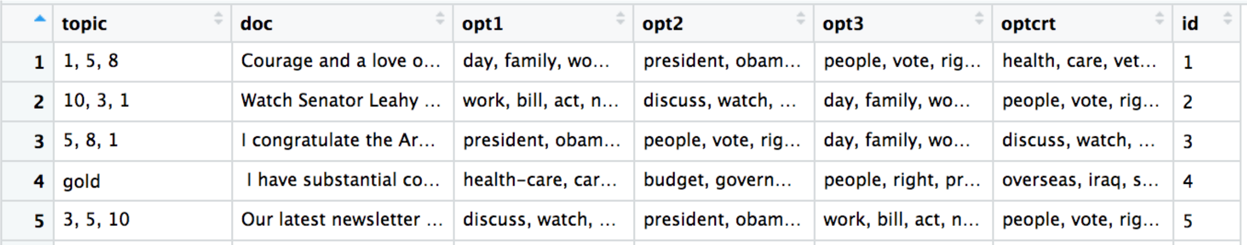
3) Randomly mix in the gold-standard HITs. The `mixGold()` function ensures that one gold-standard HIT would show up every $\frac{\text{\# total tasks}}{\text{\# gold standard tasks}}$ number of tasks. It also assigns a unique id for each of the task.

```
goldT8WSI <- read.csv("goldT8WSI.csv", stringsAsFactors = FALSE)
allT8WSItasks <- mixGold(tasks = T8WSItasks, golds = goldT8WSI)
```

The output looks as below, where the topic column for gold-standard HITs indicates "gold."

4) Next, record the prepared tasks to a specified path. While doing so, the function will create a list of two, where the first element is the above data frame of prepared tasks and the second element only keeps the documents and word sets (in the case of WI only keeps the words and in the case of R4WSI only keeps the word sets) with randomized order, leaving out other meta data, e.g., the ids. Users can assign the record to an object for immediate use. They can always load it later as well.

Figure SI13: After mixing in gold-standard HITs



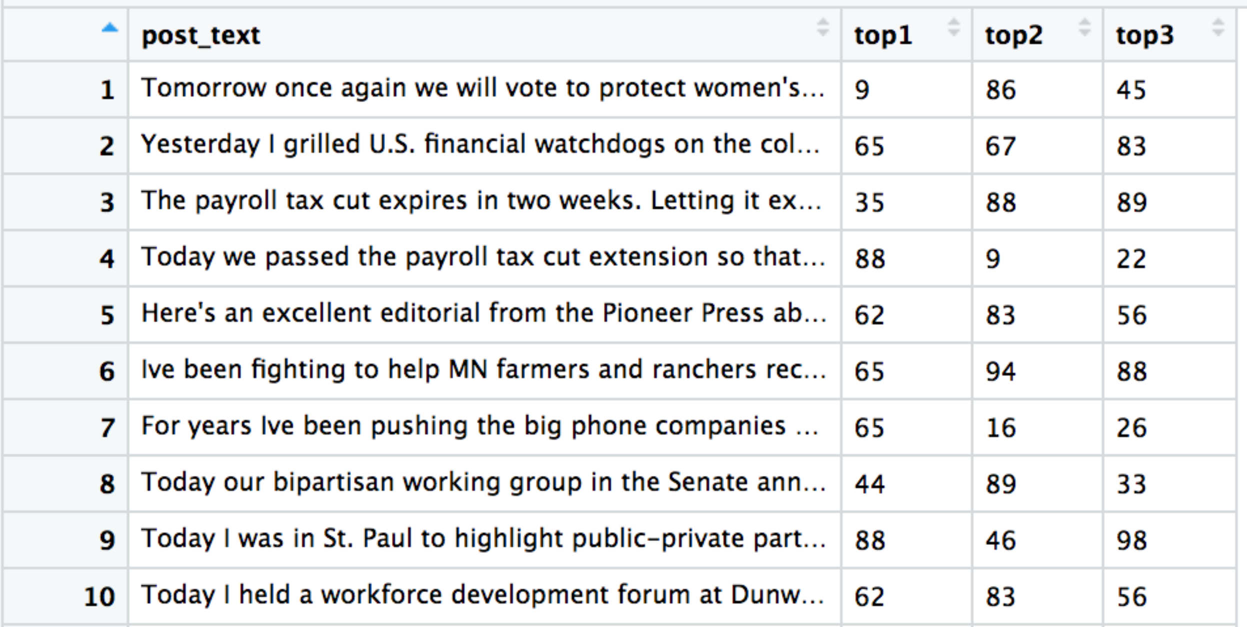
	topic	doc	opt1	opt2	opt3	optcrt	id
1	1, 5, 8	Courage and a love o...	day, family, wo...	president, obam...	people, vote, rig...	health, care, vet...	1
2	10, 3, 1	Watch Senator Leahy ...	work, bill, act, n...	discuss, watch, ...	day, family, wo...	people, vote, rig...	2
3	5, 8, 1	I congratulate the Ar...	president, obam...	people, vote, rig...	day, family, wo...	discuss, watch, ...	3
4	gold	I have substantial co...	health–care, car...	budget, govern...	people, right, pr...	overseas, iraq, s...	4
5	3, 5, 10	Our latest newsletter ...	discuss, watch, ...	president, obam...	work, bill, act, n...	people, vote, rig...	5

```
record <- recordTasks(type = "T8WSI", tasks = allT8WSItasks,
                      path = "T8WSI/record.Rdata")
```

5.4.2 Preparing Tasks for Label Validation

The procedure for preparing label validation tasks locally is almost identical to that for preparing topic validation tasks. One extra step is that users need to predict the high probability topics for their documents and define a document pool, from which they would like to draw documents. Typically, the document pool for Label Intrusion (LI) tasks contains all documents whose top three high probability topics are among the pre-defined category (in our example, domestic topics). The document pool for Optimal Label (OL) tasks, on the other hand, contains all documents whose top one high probability topic is among the pre-defined category (in our example, domestic topics). The document pool file – like the one depicted in Figure SI14 – must contain a document column and one or three top column(s) named as "top1" (and "top2"/"top3").

Figure SI14: Structure of the Document Pool Document



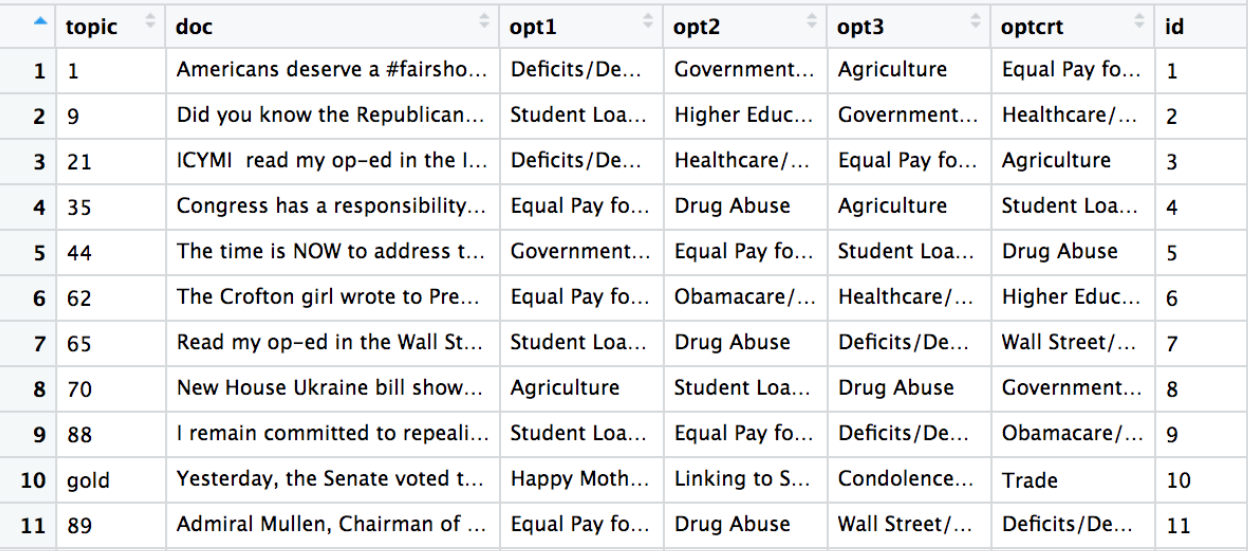
	post_text	top1	top2	top3
1	Tomorrow once again we will vote to protect women's...	9	86	45
2	Yesterday I grilled U.S. financial watchdogs on the col...	65	67	83
3	The payroll tax cut expires in two weeks. Letting it ex...	35	88	89
4	Today we passed the payroll tax cut extension so that...	88	9	22
5	Here's an excellent editorial from the Pioneer Press ab...	62	83	56
6	Ive been fighting to help MN farmers and ranchers rec...	65	94	88
7	For years Ive been pushing the big phone companies ...	65	16	26
8	Today our bipartisan working group in the Senate ann...	44	89	33
9	Today I was in St. Paul to highlight public–private part...	88	46	98
10	Today I held a workforce development forum at Dunw...	62	83	56

Now the users will use the function `validateLabel()`, specifying either `type = "LI"` or type = "OL". text.predict specifies the document pool. text.name specifies the name of the document column of the document pool data frame. labels are the labels given by human coders, in the same order with labels.index. Users could choose to add additional intruder labels through the labels.add argument. In our example for illustration, we validate 10 domestic labels while adding 10 international labels as cross-category intruders.

```
documentpool <- read.csv("documentpool.csv", stringsAsFactors = FALSE)
# "documentpool.csv" is equivalent to "top1domText.csv" in our replication materials.
OLtasks <- validateLabel(type = "OL",
                         n = 500,
                         text.predict = documentpool,
                         text.name = "post_text",
                         labels = c("Equal Pay for Women",
                                    "Healthcare/Reproductive Rights",
                                    "Agriculture",
                                    "Student Loan/Debt",
                                    "Drug Abuse",
                                    "Higher Education/Job Training",
                                    "Wall Street/Financial Sector",
                                    "Government Sutdown/Congressional Budget",
                                    "Obamacare/Tax Policy",
                                    "Deficits/Debt/Budget"),
                         labels.index = c(1, 9, 21, 35, 44, 62, 65, 70, 88, 89),
                         labels.add = c("International Trade",
                                        "Praising Active Military/Military Units",
                                        "Terrorism",
                                        "Military Sexual Assault",
                                        "Nuclear Deterrence/International Security",
                                        "Air Force",
                                        "Honoring Specific Veterans",
                                        "Honoring Veterans/Heroes",
                                        "Military Operations/Armed Conflicts",
                                        "Veterans Affairs/Veterans Healthcare"))
```

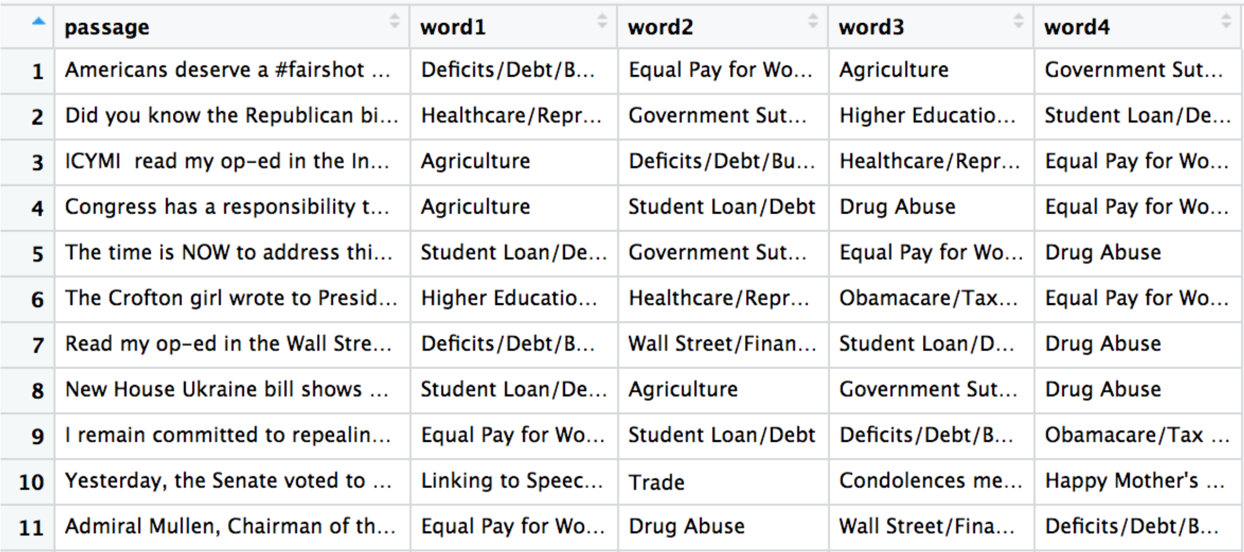
Like the topic intrusion tasks, we have randomly mixed 50 gold-standard HITs into 500 tasks for both task structure. We recommend future researchers to adopt this “gold-standard HITs” approach as well. The recordTasks() function records the tasks at a specified local directory. The two elements in the output list of recordTasks() is shown in Figure SI15 and SI16.

Figure SI15: Local Record with Identifiers (`recordTasks()` Output 1)



	topic	doc	opt1	opt2	opt3	optcrt	id
1	1	Americans deserve a #fairsho...	Deficits/De...	Government...	Agriculture	Equal Pay fo...	1
2	9	Did you know the Republican...	Student Loa...	Higher Educ...	Government...	Healthcare/...	2
3	21	ICYMI read my op-ed in the I...	Deficits/De...	Healthcare/...	Equal Pay fo...	Agriculture	3
4	35	Congress has a responsibility...	Equal Pay fo...	Drug Abuse	Agriculture	Student Loa...	4
5	44	The time is NOW to address t...	Government...	Equal Pay fo...	Student Loa...	Drug Abuse	5
6	62	The Crofton girl wrote to Pre...	Equal Pay fo...	Obamacare/...	Healthcare/...	Higher Educ...	6
7	65	Read my op-ed in the Wall St...	Student Loa...	Drug Abuse	Deficits/De...	Wall Street/...	7
8	70	New House Ukraine bill show...	Agriculture	Student Loa...	Drug Abuse	Government...	8
9	88	I remain committed to repeali...	Student Loa...	Equal Pay fo...	Deficits/De...	Obamacare/...	9
10	gold	Yesterday, the Senate voted t...	Happy Moth...	Linking to S...	Condolence...	Trade	10
11	89	Admiral Mullen, Chairman of ...	Equal Pay fo...	Drug Abuse	Wall Street/...	Deficits/De...	11

Figure SI16: Local Record, Randomized Order, i.e., (`recordTasks()` Output 2)



	passage	word1	word2	word3	word4
1	Americans deserve a #fairshot ...	Deficits/Debt/B...	Equal Pay for Wo...	Agriculture	Government Sut...
2	Did you know the Republican bi...	Healthcare/Repr...	Government Sut...	Higher Educatio...	Student Loan/De...
3	ICYMI read my op-ed in the In...	Agriculture	Deficits/Debt/Bu...	Healthcare/Repr...	Equal Pay for Wo...
4	Congress has a responsibility t...	Agriculture	Student Loan/Debt	Drug Abuse	Equal Pay for Wo...
5	The time is NOW to address thi...	Student Loan/De...	Government Sut...	Equal Pay for Wo...	Drug Abuse
6	The Crofton girl wrote to Presid...	Higher Educatio...	Healthcare/Repr...	Obamacare/Tax...	Equal Pay for Wo...
7	Read my op-ed in the Wall Stre...	Deficits/Debt/B...	Wall Street/Finan...	Student Loan/D...	Drug Abuse
8	New House Ukraine bill shows ...	Student Loan/De...	Agriculture	Government Sut...	Drug Abuse
9	I remain committed to repealin...	Equal Pay for Wo...	Student Loan/Debt	Deficits/Debt/B...	Obamacare/Tax ...
10	Yesterday, the Senate voted to ...	Linking to Speec...	Trade	Condolences me...	Happy Mother's ...
11	Admiral Mullen, Chairman of th...	Equal Pay for Wo...	Drug Abuse	Wall Street/Fina...	Deficits/Debt/B...

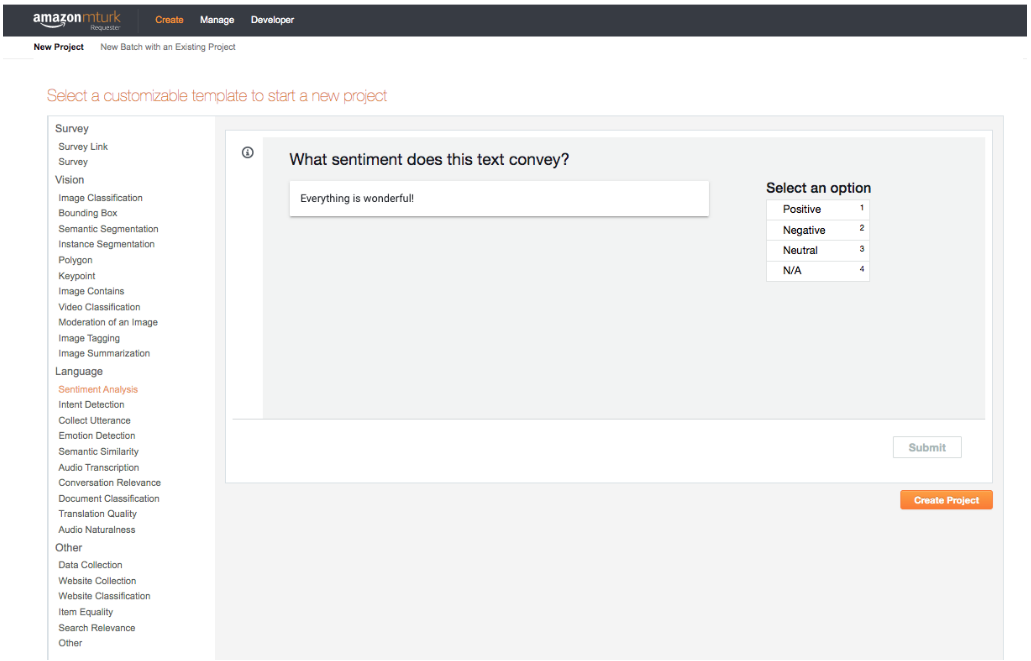
5.5 Posting Tasks and Maintaining

Now, researchers should be able to interact with Mturk and post tasks. They will do so by first manually specifying the basic tasks settings and then sending tasks through the API.

1) Login to the Mturk requester page, click "New Project" under the "Create" tab. Click the "Sentiment Analysis" tab (which allows us to specify our customized layout later) on the left. Then, click "Create Project."

2) Following the prompts, specify the properties. In particular, be sure to add an additional qualification created in section "Training and Certification" under the "Worker requirements" tab. Our tasks in the paper require a "master" qualification as well. Then, click "Design Layout."

Figure SI17: Create a Project: Step One



Select a customizable template to start a new project

Survey
Survey Link
Survey
Vision
Image Classification
Bounding Box
Semantic Segmentation
Instance Segmentation
Polygon
Keypoint
Image Contains
Video Classification
Moderation of an Image
Image Tagging
Image Summarization
Language
Sentiment Analysis
Intent Detection
Collect Utterance
Emotion Detection
Semantic Similarity
Audio Transcription
Conversation Relevance
Document Classification
Translation Quality
Audio Naturalness
Other
Data Collection
Website Collection
Website Classification
Item Equality
Search Relevance
Other

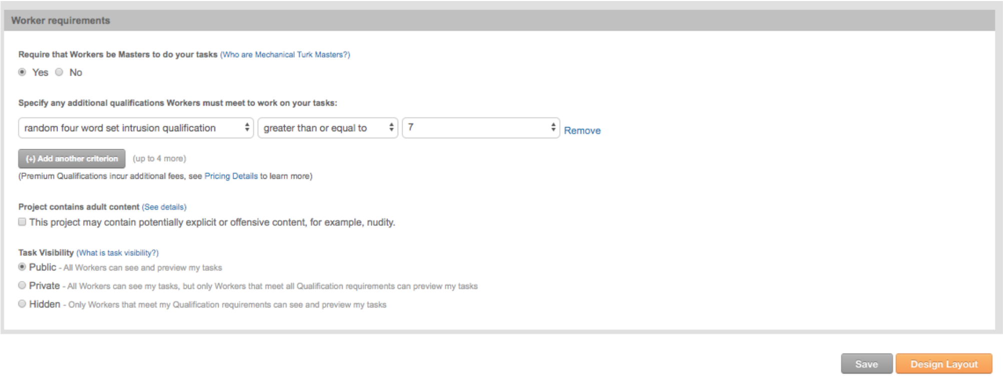
What sentiment does this text convey?

Everything is wonderful!

Select an option

Positive	1
Negative	2
Neutral	3
N/A	4

Figure SI18: Create a Project: Step Two


Worker requirements

Require that Workers be Masters to do your tasks (Who are Mechanical Turk Masters?)

● Yes ○ No

Specify any additional qualifications Workers must meet to work on your tasks:

random four word set intrusion qualification | greater than or equal to | 7 | Remove

(+) Add another criterion (up to 4 more)

(Premium Qualifications incur additional fees, see Pricing Details to learn more)

Project contains adult content (See details)

□ This project may contain potentially explicit or offensive content, for example, nudity.

Task Visibility (What is task visibility?)

● Public - All Workers can see and preview my tasks

○ Private - All Workers can see my tasks, but only Workers that meet all Qualification requirements can preview my tasks

○ Hidden - Only Workers that meet my Qualification requirements can see and preview my tasks

3) Replace the default layout file with our customized layout. The example .html file can be found as a separate file in our replication package (Layout.html). Notice that researchers can change the instructions for their own tasks, which will later appear on the left of the workers' screen. For the tasks in our paper, we specifically tell them:

> Some of these choices will be very clear, but others will require you to use your best judgment. We understand that in many cases it will be hard to tell what the "right" choice is. Use your best judgement, but please be attentive. **We have randomly mixed in a number of gold-standard HITs whose answers are less ambiguous to evaluate the quality of your work.** If you miss a large number of gold-standard HITs, you may be **blocked** from continued participation in this study (and future studies).

> After you've done a relatively large number of HITs of a specific task structure, your qualification might be temporarily suppressed. This does not mean you've done anything wrong, but we need to ensure that a variety of workers complete our tasks. This is **NOT a block** and will have **NO negative impact** on you Mturk record.

4) Preview the Layout and finish creating tasks.

From here, we can do everything through the API, sending instructions from `R` to interact with the Mturk platform.

5) Load the prepared tasks from section 5.4.

```
load("T8WSI/record.Rdata")
```

6) Send tasks to Mturk using the `sendTasks()` function. Users must specify a path to **record the HITids immediately** as they are vital for future procedures. The function will record the HITids in a list format where the first element is a vector of HIT ids returned by Mturk and the second element is **the mapping of the local HIT ids to the Mturk HIT ids**. The `tasksids` argument allows users to post specific tasks by specifying the task ids (numeric form) in a vector. If `tasksids` left unspecified, all tasks in the record will be posted.

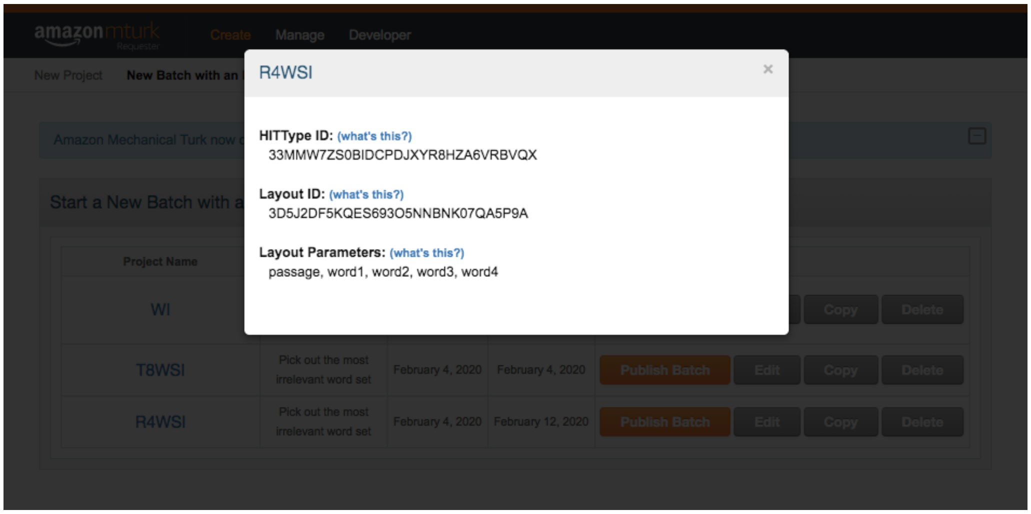
```
HITids <- sendTasks(hit_type = "FIND_IT_FROM_DASHBOARD",
                    hit_layout = "FIND_IT_FROM_DASHBOARD",
                    type = "T8WSI",
                    tasksrecord = record,
                    tasksids = c(1:10, 15),
                    HITidspath = "T8WSI/HITids/testIDs.Rdata")

# Console responds
> Sending task to MTurk
> HITids saved to T8WSI/HITids/testIDs.Rdata
```

Notice that `hit_type` and `hit_layout` can be found from the MTurk requester's dashboard by clicking the project name, e.g., "T8WSI":

Once posted, worker should be able to see the tasks on the Mturk platform. They can get the certification by going through the training module and passing the test. Figure SI20 depicts an example task from the workers' perspective, we see that a "master" qualification is also required.

Figure SI19: Finding HITType ID and Layout ID



R4WSI

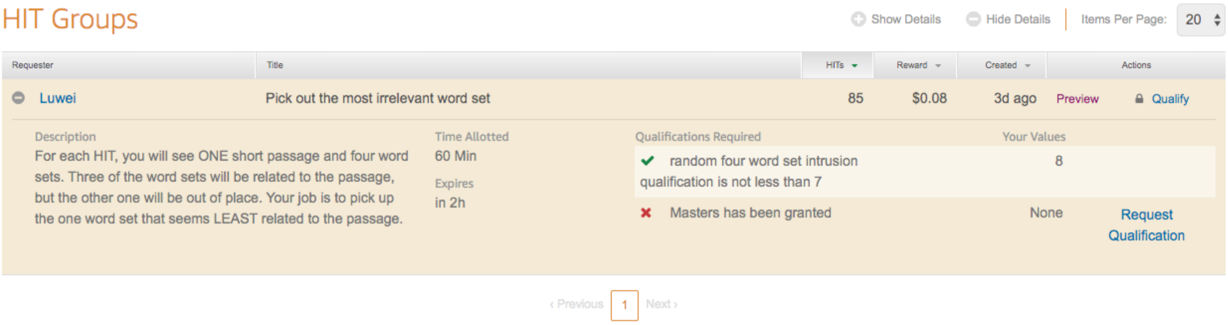
HITType ID: (what's this?)
33MMW7ZS0BIDCPDJXYR8HZA6VRBVQX

Layout ID: (what's this?)
3D5J2DF5KQES693O5NNBNK07QA5P9A

Layout Parameters: (what's this?)
passage, word1, word2, word3, word4

Project Name			
WI			
T8WSI	Pick out the most irrelevant word set	February 4, 2020	February 4, 2020
R4WSI	Pick out the most irrelevant word set	February 4, 2020	February 12, 2020

Figure SI20: How tasks look on Mturk



Requester	Title	HITs	Reward	Created	Actions	
Luwei	Pick out the most irrelevant word set	85	$0.08	3d ago	Preview	Qualify

Description
For each HIT, you will see ONE short passage and four word sets. Three of the word sets will be related to the passage, but the other one will be out of place. Your job is to pick up the one word set that seems LEAST related to the passage.

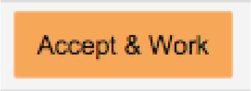
Time Allotted
60 Min

Expires
in 2h

Qualifications Required	Your Values	
✓ random four word set intrusion qualification is not less than 7	8	
✗ Masters has been granted	None	Request Qualification

Once the workers get the qualification, the "lock" icon will become "Accept & work".

Figure SI21: Workers can start working after seeing this "Accept & work" icon



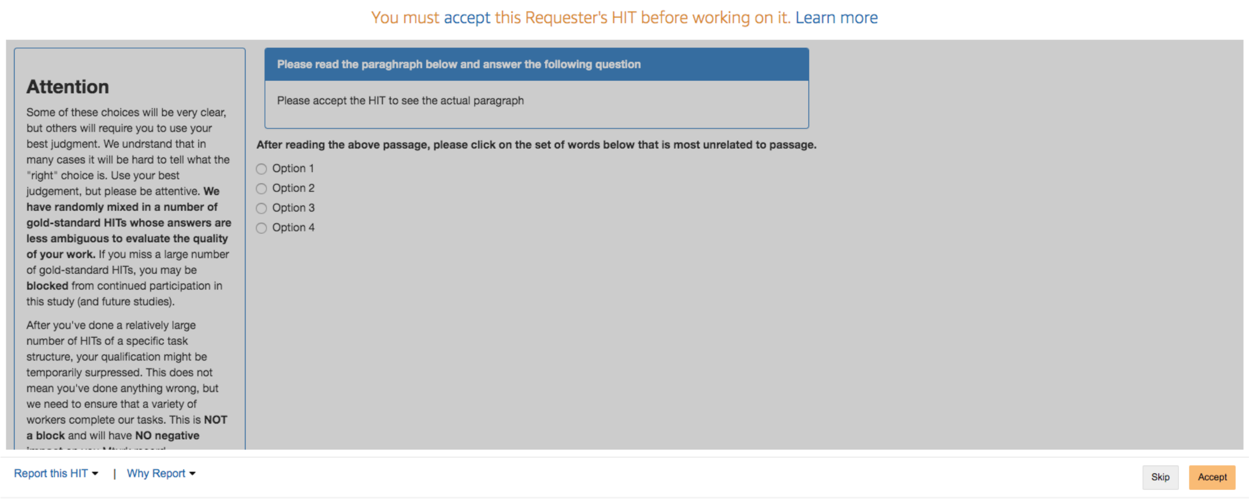
Before they accept the HIT, they would **NOT** be able to see the actual options or document, which feature prevents them from selecting the tasks.

They'll only be able to see the actual tasks after accepting. Notice that they can ALWAYS see the instructions on the left.

Researchers are able to extend the time of their HITs before or after the HITs expire.

```
for(i in HITids[[1]]){
  ExtendHIT(hit = i, add.seconds = 3600)
}
```

Figure SI22: Before Accepting the HIT



You must accept this Requester's HIT before working on it. Learn more

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. We undrstand that in many cases it will be hard to tell what the "right" choice is. Use your best judgement, but please be attentive. **We have randomly mixed in a number of gold-standard HITs whose answers are less ambiguous to evaluate the quality of your work.** If you miss a large number of gold-standard HITs, you may be **blocked** from continued participation in this study (and future studies).

After you've done a relatively large number of HITs of a specific task structure, your qualification might be temporarily surpressed. This does not mean you've done anything wrong, but we need to ensure that a variety of workers complete our tasks. This is **NOT a block** and will have **NO negative**

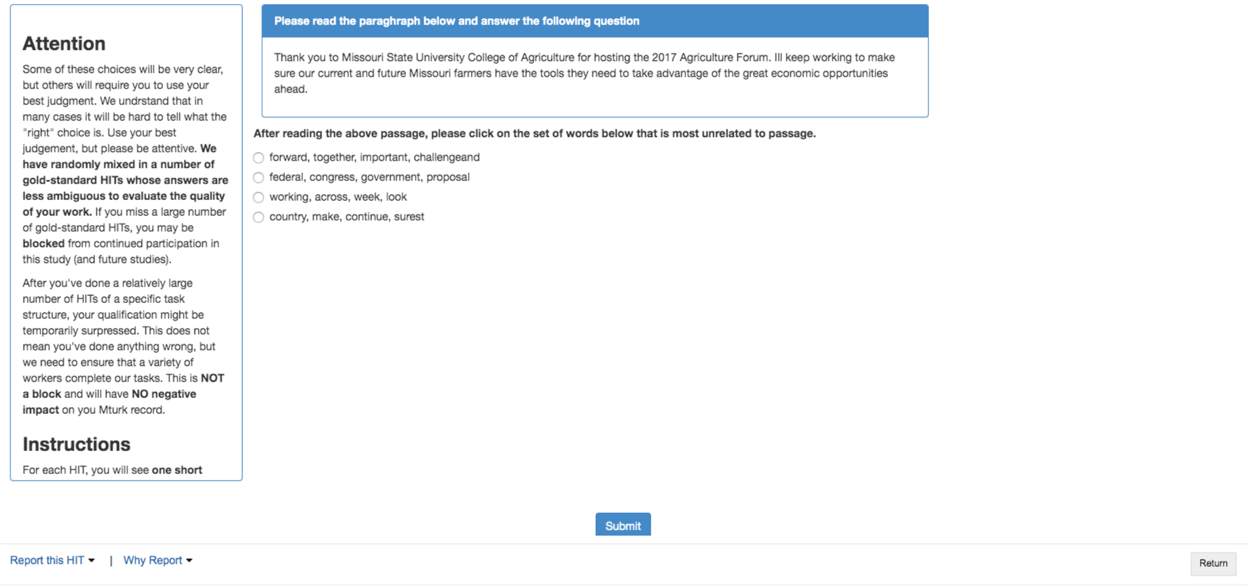
Please read the paraghraph below and answer the following question

Please accept the HIT to see the actual paragraph

After reading the above passage, please click on the set of words below that is most unrelated to passage.

○ Option 1
○ Option 2
○ Option 3
○ Option 4

Figure SI23: After Accepting the HIT



Attention

Some of these choices will be very clear, but others will require you to use your best judgment. We undrstand that in many cases it will be hard to tell what the "right" choice is. Use your best judgement, but please be attentive. **We have randomly mixed in a number of gold-standard HITs whose answers are less ambiguous to evaluate the quality of your work.** If you miss a large number of gold-standard HITs, you may be **blocked** from continued participation in this study (and future studies).

After you've done a relatively large number of HITs of a specific task structure, your qualification might be temporarily surpressed. This does not mean you've done anything wrong, but we need to ensure that a variety of workers complete our tasks. This is **NOT a block** and will have **NO negative impact** on you Mturk record.

Instructions

For each HIT, you will see **one short**

Please read the paraghraph below and answer the following question

Thank you to Missouri State University College of Agriculture for hosting the 2017 Agriculture Forum. Ill keep working to make sure our current and future Missouri farmers have the tools they need to take advantage of the great economic opportunities ahead.

After reading the above passage, please click on the set of words below that is most unrelated to passage.

○ forward, together, important, challengeand
○ federal, congress, government, proposal
○ working, across, week, look
○ country, make, continue, surest

5.6 Retrieving and Evaluating Results

The `getResults()` function allows the researchers to retrieve results from a batch. It would work **no matter the batch has been completed or not**. In that sense, this function can also be used to check batch status. Notice that `batch_id` here can be any string that helps the researchers to refer to the batch in the future. It is first specified here. The `retry`, if `TRUE`, retries retrieving results from Mturk API at most five times, given that it fails retrieving all results the first time. The default is `retry = TRUE`.

```
testresults <- getResults(batch_id = "testbatch",
                          hit_ids = HITids,
                          retry = FALSE)
```

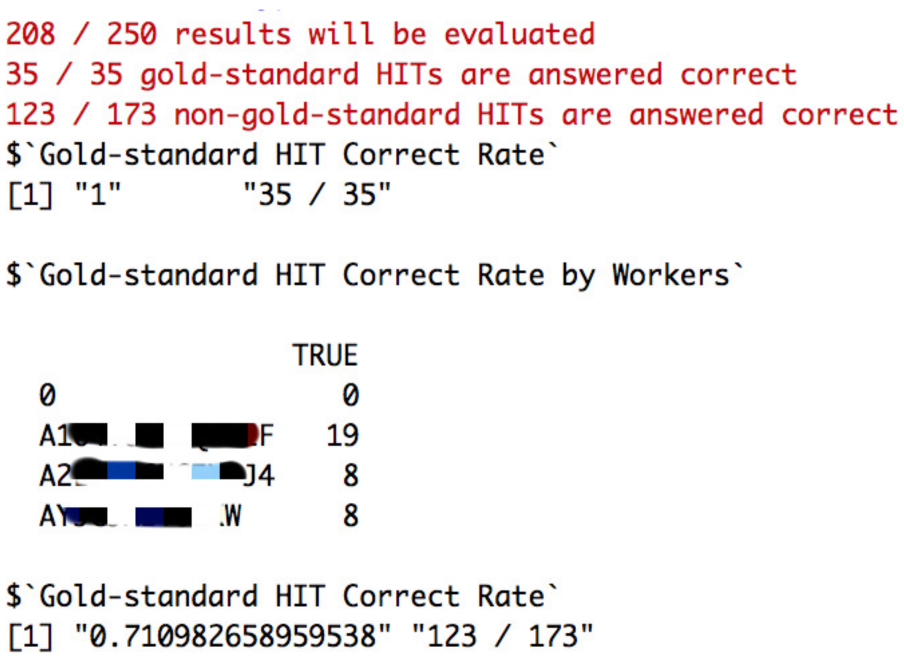
```
# Console returns
> Start getting HITs...
> 208 / 250 results retrieved
```

We highly recommend users to record the results immediately, especially when the batch is finished as the results might be deleted by the Mturk platform at some point.

To evaluate results, use `evalResults()`. This function identifies workers who consistently give poor-quality work and also returns the rate that human workers agree with the machine prediction.

```
evalResults(results = testresults,
            key = record,
            type = "T8WSI")
```

Figure SI24: Evaluating Results: Output from `evalResults()`



```
208 / 250 results will be evaluated
35 / 35 gold-standard HITs are answered correct
123 / 173 non-gold-standard HITs are answered correct
$`Gold-standard HIT Correct Rate`
[1] "1"       "35 / 35"

$`Gold-standard HIT Correct Rate by Workers`

                 TRUE
  0                 0
  [redacted]       19
  [redacted]        8
  [redacted]        8

$`Gold-standard HIT Correct Rate`
[1] "0.710982658959538" "123 / 173"
```

In rare cases, researchers will encounter workers who keep giving work of poor quality. Therefore, researcher can reject HITs from these workers and ban them from participating in future tasks in two steps.

1) Identify these workers and revoke their qualifications.

```
workers_to_ban <- c("WORKER1", "WORKER2")
for(worker in workers_to_ban){
  AssignQualifications(qual = "QUALIFICATION_ID",
                       workers = worker,
                       notify = TRUE,
                       value = 0)
}
```

2) Reject their qualifications.

```
AssignmentsToReject <-
    testresult$assignment_id[testresults$worker_id %in% workers_to_ban]
RejectAssignments(assignments = unique(AssignmentsToReject),
                  feedback = "The quality of your work in the most recent
                  batch of HITs did not pass an audit. Your HITs from this
                  batch have been rejected and your qualification has
                  been deactivated.")
```

6 Workers' Training and Performance

This section first presents the training modules for the five tasks used in the paper. The second part provides some statistics and discussion about the workers' performance.

6.1 Training Module for Word Intrusion

Completing this training module qualifies you to complete Word Intrusion HITs.

Basic instructions
1. For each HIT, you will see FIVE words.
2. Four of them will be related to each other, but one word will be out of place.
3. Your job is to pick up the one word that does NOT belong with the others.

Background

The words you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o job
o business
o day
o work
o economy

Answer: The correct answer is "day". "job", "business", "work" and "economy" are all related to economic activities. "day" is not.

Practice HIT #2

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o energy

o water
o climate
o small
o clean

Answer: The correct answer is “small”. “energy”, “water”, “climate” and “clean” are all connected to one another under the theme of environmental policy concerns. However, “small” is irrelevant to this theme.

Practice HIT #3

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o country
o america
o oil
o nation
o freedom

Answer: The correct answer is “oil”. “country”, “america”, “nation” and “freedom” are related to America and American values. However, “oil” is irrelevant to this theme.

Practice HIT #4

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o farm
o school
o students
o university
o high

Answer: The correct answer is “farm”. “school”, “students”, “university” and “high” are all related to education. However, “farm” is not.

Practice HIT #5

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o veterans
o service
o proud
o meeting
o force

Answer: The correct answer is “meeting”. “veterans”, “service”, “proud” and “force” are all related to veterans affairs and the armed services. However, “meeting” is irrelevant.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o tax
o budget
o debt
o bipartisan
o spending

Answer: [**Not Available in the Real Test**] “bipartisan”

Test HIT #2

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o homework
o security
o nuclear
o iran
o hearing

Answer: [**Not Available in the Real Test**] “homework”

Test HIT #3

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o watch
o news
o committee
o live
o show

Answer: [**Not Available in the Real Test**] “committee”

Test HIT #4

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o vegetable
o congress
o president
o republicans
o house

Answer: [**Not Available in the Real Test**] “vegetable”

Test HIT #5

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o health
o bill
o act
o airplane
o legislation

Answer: [**Not Available in the Real Test**] “airplane”

Test HIT #6

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o agriculture
o happy
o farmers
o industry
o trade

Answer: [**Not Available in the Real Test**] “happy”

Test HIT #7

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o sanders
o social
o night
o billion
o spending

Answer: [**Not Available in the Real Test**] “night”

Test HIT #8

Please read the five words below, and choose one that is most IRRELEVANT to the other four.

o nuclear-power
o discuss
o watch
o spoke
o read

Answer: [**Not Available in the Real Test**] “nuclear-power”

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

6.2 Training Module for Top 8 Word Set Intrusion

Completing this training module qualifies you to complete Top Eight Word Set Intrusion HITs.

Basic instructions
1. For each HIT, you will see ONE SHORT PASSAGE and FOUR word sets.
2. Three of the word sets will be related to the passage, but the other one will be out of place.
3. Your job is to pick up the one word set that seems LEAST related to the passage.

Background

The passages and words you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will highlight the important text and provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

> *Heartening story in the Lawrence Journal-World about 89-year-old WWII Veteran and The University of Kansas Football Team Alumnus Bryan Sperry. He stole the show in Saturday's alumni flag football game.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o veterans, service, national, honor, military, thank, proud, women
o great, students, school, office, visit, thanks, state, county
o watch, senate, morning, last, discuss, live, read, news
o sanders, tax, budget, new, bernie, said, debt, pay

Answer: The correct answer is "sanders, tax, budget, new, bernie, said, debt, pay". The passage is about Bryan Sperry. It is a local interest story that focuses on the achievements of a veteran a football alumni involved in a flag football game.

The word set "veterans, service, national, honor, military, thank, proud, women" clearly relates to his identity as "an 89-year-old WWII Veteran.

The word set "great, students, school, office, visit, thanks, state, county" relates to his identity as an alumnus of the University of Kansas Football Team and the fact that this post is celebrating a local event/achievement.

The passage also references local news coverage (the post is actually about a news story), which is somewhat related to the set "watch, senate, morning, last, discuss, live, read, news."

The word set "sanders, tax, budget, new, bernie, said, debt, pay", however, seems to relate public policy and are not related to a local non-politcial event.

Practice HIT #2

> *Earlier this week I spoke at a town hall in support of Referendum 74 at First Baptist Church in Seattle. Stand with me – and thousands of Washington families – and click below to volunteer in support of marriage equality!*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o watch, senate, morning, last, discuss, live, read, news
o day, family, people, one, life, every, years, world
o energy, water, climate, change, oil, federal, clean, regulations
o great, students, school, office, visit, thanks, state, county

Answer: The correct answer is "energy, water, climate, change, oil, federal, clean, regulations". The passage is in support of a state referendum on gay marriage. It also references a town hall (a local political meeting) held at a church.

Two word sets "great, students, school, office, visit, thanks, state, county" and "watch, senate, morning, last, discuss, live, read, news" relates to the fact that the post references the Senator appearing at a local political gathering.

Another word set "day, family, people, one, life, every, years, world" seems to be related to the fact that the issue concerns "marriage" and "families."

The final word set "energy, water, climate, change, oil, federal, clean, regulations", however, seems to primarily be about environmental policy.

Practice HIT #3

> *The London Metropolitan Police Department has great taste in cars! They're patrolling the streets of London in BMW X5s made in – you guessed it – South Carolina.*
>
> *I'm in London this weekend attending the annual Farnborough International Airshow in support of the Boeing 787 Dreamliners also made in South Carolina. #PalmettoPride #PalmettoPrideintheUK*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o new, jobs, state, help, businesses, work, business, economy
o veterans, service, national, honor, military, thank, proud, women
o president, congress, people, senate, american, americans, government, obama
o great, students, school, office, visit, thanks, state, county

Answer: The correct answer is "president, congress, people, senate, american, americans, government, obama". The passage expresses pride in the cars and airplanes made in South

Carolina. The passage mentions cars, the police, and Boeing airplanes manufactured in South Carolina, and an airshow in London.

One word set “new, jobs, state, help, businesses, work, business, economy” clearly relates to the economy, and is related to the cars and planes manufactured in South Carolina.

A second word set “great, students, school, office, visit, thanks, state, county” again relates to the local nature and focus of the post.

A third word set “veterans, service, national, honor, military, thank, proud, women” is a bit more difficult. Probably it is related to both the reference to the police and the international airshow.

The fourth set of words “president, congress, people, senate, american, americans, government, obama”, has no real connection with the passage.

Practice HIT #4

> *Today’s executive order from President Trump is more about extreme xenophobia than extreme vetting. This executive order is the equivalent of a ”Keep Out” sign posted at America’s borders.*
>
> *Turning away immigrants based on their nationality and religion is un-American and in direct opposition to everything for which our Founding Fathers fought. President Trump may not call it a Muslim ban, but it is, and runs afoul of our morals and values.*
>
> *We must ensure that we have the strongest safeguards in place to keep terrorists from ever reaching our shores. And we must fully and thoroughly vet all refugees to screen out any potential terrorist threats.*
>
> *But as conflict and war force millions around the world from their homeland, the United States should welcome more refugees, not less. Suspending the U.S. refugee resettlement program will endanger refugees’ lives and tear families apart.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o day, family, people, one, life, every, years, world
o great, students, school, office, visit, thanks, state, county
o president, congress, people, senate, american, americans, government, obama
o committee, security, hearing, secretary, united, states, department, senate

Answer: The correct answer is “great, students, school, office, visit, thanks, state, county”. The passage is about President Trump’s executive order on immigrants, which is a national-level policy discussion.

The word set “president, congress, people, senate, american, americans, government, obama” is about national politics and, therefore, relevant.

Another word set “committee, security, hearing, secretary, united, states, department, senate” seems to relate to politics and perhaps national security.

The word set “day, family, people, one, life, every, years, world” is a bit tricky. However, it seems to relate to social issues and is triggered by the mention of families.

The final word set “great, students, school, office, visit, thanks, state, county” seems to refer to local politics and local events and is not clearly related to the discussion of national immigration policy.

Practice HIT #5

> *Women in New Mexico and across the country deserve to make their own decision about family planning. Laws to protect a women's right to choose what's best for her body and well-being should not be restricted to what state she lives in. On the 42nd anniversary of the landmark Roe v. Wade decision, I remain committed to strengthening women's reproductive rights and freedom to make their own personal health care choices. #Roe42*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o president, congress, people, senate, american, americans, government, obama
o energy, water, climate, change, oil, federal, clean, regulations
o health, bill, act, care, legislation, help, need, bipartisan
o day, family, people, one, life, every, years, world

Answer: The correct answer is "energy, water, climate, change, oil, federal, clean, regulations". The passage advocates for a pro-choice position and frames abortion as personal health care choices (family planning).

The word set "health, bill, act, care, legislation, help, need, bipartisan" is about health care and, therefore, relevant.

Another word set "day, family, people, one, life, every, years, world" seems to relate to social issues and is triggered by the mention of families.

The word set "President, congress, people, senate, american, americans, government, obama" is a bit more ambiguous. It relates to legislative politics and the passage is talking about an important issue in legislative politics. Thus, it is also relevant.

The final word set "energy, water, climate, change, oil, federal, clean, regulations" is about environmental issues and not relevant to the passage.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

> *Being conservative means controlling spending and costs. Those without insurance are forced to seek care in expensive ER's after their conditions have worsened. Hospitals are legally required to treat patients in ERs regardless of whether they can pay and they pass the cost of treatment onto the privately insured, increasing total costs. Increasing coverage and allowing more people to manage their health care is cheaper for society. The Cassidy-Collins would accomplish this.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o conservative, congress, people, senate, american, society, government, obama
o spending, tax, budget, new, cost, price, debt, pay
o military, veterans, service, national, defence, threat, attack, iran
o health, bill, act, care, legislation, help, need, insurance

Answer: [**Not Available in the Real Test**] "military, veterans, service, national, defence, threat, attack, iran"

Test HIT #2

> *Today Sen. Judd Gregg and I introduced The Bipartisan Tax Fairness and Simplification Act of 2010, which will help middle-class taxpayers by streamlining and modernizing the outdated tax code. The proposal includes fiscally responsible tax cuts to help working families struggling to make ends meet and also eliminates the corporate tax break that encourages companies to invest overseas rather than creating jobs in the U.S.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o water, river, environment, change, new, future, proud, great
o senate, bill, act, introduce, legislation, help, need, bipartisan
o invest, jobs, state, help, campany, work, business, economy
o sanders, tax, budget, cut, bernie, said, debt, pay

Answer: [**Not Available in the Real Test**] "water, river, environment, change, new, future, proud, great"

Test HIT #3

> *Obama simultaneously could ruin Putin's day and brighten the lives of millions of Americans. All Obama needs is the courage to tell the environmental Left to let him do the right thing.*
>
> *Today is the last day for the public to comment on whether the State Department should approve a presidential permit for the #KeystoneXL pipeline. This article should give you some ideas of things to include in a comment. Here is where you can leave your comment: [link omitted]*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o day, family, people, public, life, every, millon, world
o energy, water, climate, change, oil, federal, clean, regulations
o president, congress, people, senate, american, state, government, obama
o veterans, service, national, honor, military, thank, proud, women

Answer: [**Not Available in the Real Test**] "veterans, service, national, honor, military, thank, proud, women"

Test HIT #4

> *Look, here is the bottom line. We remain the only nation in the industrialized world that doesn't guarantee health care to all people. We have 29 million people who are uninsured, yet we are spending far, far more per capita on health care than do the people of any other country.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o health, bill, act, care, legislation, help, need, bipartisan
o students, school, office, class, young, educare, college, county
o president, congress, people, senate, american, americans, government, obama
o spending, tax, budget, new, bernie, said, debt, pay

Answer: [**Not Available in the Real Test**] "students, school, office, class, young, educare, college, county"

Test HIT #5

> *Like so many Rhode Islanders, I am deeply disturbed by President-elect Trump's appointment of Steve Bannon to be his chief strategist in the White House. As Executive Chairman of Breitbart, Bannon served as a conduit for some of the worst sentiments in our society – hatred and violence on the basis of race, religion, gender, and way of life. As CEO of Trump's campaign, he employed the hateful code of the white supremacist movement to leverage prejudice and fear, for political gain. Now, the President-elect wants Bannon to help guide his administration. Many Rhode Islanders have called and written to tell me how unsettling Bannon's presence in the White House is, and I share their concern. If President-elect Trump will not denounce the dangerous ideas Steve Bannon has represented, it is up to the rest of us to stand firm against them.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o students, school, office, class, young, educare, college, county
o president, congress, people, senate, american, americans, government, obama
o day, family, people, one, life, every, equal, world
o committee, appointment, hearing, represent, united, states, department, senate

Answer: [**Not Available in the Real Test**] "students, school, office, class, young, educare, college, county"

Test HIT #6

> *I wish Democrats would show some interest in how and why President Obama and Susan Rice got it so wrong about the true nature of the Benghazi attack. I wish Democrats would show a little interest about why Secretary Clinton was clueless about the multiple security requests coming from Benghazi and how she allowed our mission to become a death trap. I wish Democrats would show a little interest in finding out whether Mike Morell, the former #2 at the CIA, lied to Congress and the American people about a protest that never happened. Democrats seem*

> *to be more interested in protecting the Obama Administration than they are in getting the truth.*
>
> *If Republicans win a Senate majority in 2014, one of the first things I will insist on is hearings that actually get to the bottom of what happened before, during, and after the Benghazi attack.*
>
> *The families of those who lost loved ones, and the American people, deserve nothing less than a full accounting.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o day, family, people, one, life, every, years, world
o committee, security, hearing, secretary, united, states, department, senate
o president, congress, people, senate, american, americans, government, obama
o company, jobs, invest, help, business, work, unemployment, economy

Answer: [**Not Available in the Real Test**] "company, jobs, invest, help, business, work, unemployment, economy"

Test HIT #7

> *Tomorrow, I'll be attending the White House bipartisan summit on health care reform hosted by President Obama. Share your ideas on health care reform with me now – so I can share your feedback with President Obama and Congressional leaders on Thursday!*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o military, veterans, service, national, defence, threat, attack, iran
o health, bill, act, care, legislation, help, need, bipartisan
o president, congress, people, senate, american, americans, government, obama
o watch, senate, morning, share, discuss, live, feedback, news

Answer: [**Not Available in the Real Test**] "military, veterans, service, national, defence, threat, attack, iran"

Test HIT #8

> *Today Chairman Johnson held a hearing on duplication, waste, and fraud in federal programs.*
>
> *Seven years later, with less than half of GAO's recommendations even implemented, GAO estimates that this report has resulted in actual savings of $75 billion. One simple idea has saved American taxpayers tens of billions of dollars. However, there are still hundreds of recommendations that have gone unimplemented, and very little actual duplication in the federal government has been addressed. I am pleased that the Trump administration is taking this problem seriously. The Executive Order signed by President Trump and the memorandum by OMB Director Mulvaney that followed will result in a plan to reorganize and*

> *streamline the federal government and help it better serve the American people. This is long overdue.*

After reading the above passage, please click on the set of words below that is most UNRELATED to passage.

o sanders, tax, budget, new, bernie, said, debt, pay
o committee, order, hearing, federal, united, states, department, senate
o president, administration, congress, people, senate, americans, government, obama
o students, school, office, class, young, educare, college, county

Answer: **[Not Available in the Real Test]** "students, school, office, class, young, educare, college, county"

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

6.3 Training Module for Random 4 Word Set Intrusion

Completing this training module qualifies you to complete Random Four Word Set Intrusion HITs

Basic instructions

1. For each HIT, you will see FOUR word sets.
2. Three of the word sets will be related to one another, but the other one will be out of place.
3. Your job is to pick up the one word set that seems LEAST related to others.

Background

The words you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

Please click on the word set below that is most UNRELATED to the other three.

o bill, govern, vote, cut

o senate, congress, pass, plan

o tax, budget, house, reform

o report, today, press, emergency

Answer: The correct answer is "respond, today, first-aid, emergency". While three word sets are related to policy-making institutions and procedures ("bill, govern, vote, cut", "senate, congress, pass, plan", "tax, budget, house, reform"), the set of words "respond, today, first-aid, emergency" is not. Rather it seems to relate to time sensitive response and emergencies.

Practice HIT #2

Please click on the word set below that is most UNRELATED to the other three.

o great, thank, visit, city

o state, enjoy, tour, center

o reform, obamacare, federal, year

o happy, team, celebrate, park

Answer: The correct answer is "reform, obamacare, federal, year". While the other three word sets relate to celebration of local events ("great, thank, visit, city", "state, enjoy, tour, center", and "happy, team, celebrate, park"), the set of words "reform, obamacare, federal, year" is not but rather relates to health care reform policy.

Practice HIT #3

Please click on the word set below that is most UNRELATED to the other three.

o nation, student, service, school

o honor, education, veteran, state

o social, call, secure, post

o learn, high, college, proud

Answer: The correct answer is "social, call, secure, post". Three of the word sets ("nation, student, service, school", "honor, education, professor, state", and "learn, high, college, proud") relate to education. The other word set ("social, call, secure, post") is not.

Practice HIT #4

Please click on the word set below that is most UNRELATED to the other three.

o day, family, country, people

o proud, learn, university, service

o women, american, live, make

o work, life, nation, love

Answer: The correct answer is "proud, learn, university, college". While the three other word sets seems to be generally about American ideals and values ("day, family, country, people", "women, american, live, make", and "work, life, nation, love"), the set of words “proud, learn, university, college” is instead more closely related to education.

Practice HIT #5

Please click on the word set below that is most UNRELATED to the other three.
o job, business, help, work
o confirm, judge, nominate, rule
o energy, economy, state, new
o develop, continue, import, protect

Answer: The correct answer is “confirm, judge, nominate, rule”. While the other three word sets are related to the economy and businesses (“job, business, help, work”, “energy, economy, state, new” and “develop, continue, import, protect”), the set of words “confirm, judge, nominate, rule” is more clearly related to judicial appointment and legal issues.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

Please click on the word set below that is most UNRELATED to the other three.
o discuss, join, week, senate
o hear, talk, share, live
o many, world, nation, america
o office, morning, watch, meet

Answer: [**Not Available in the Real Test**] “many, world, nation, america”

Test HIT #2

Please click on the word set below that is most UNRELATED to the other three.
o thank, visit, good, time
o health, care, veteran, help
o need, provide, support, work
o drug, program, prevent, ensure

Answer: [**Not Available in the Real Test**] “thank, visit, good, time”

Test HIT #3

Please click on the word set below that is most UNRELATED to the other three.
o right, law, court, senate
o vote, justice, supreme, constitution
o judge, nominate, rule, investigate
o news, report, read, said

Answer: [**Not Available in the Real Test**] “news, report, read, said”

Test HIT #4

Please click on the word set below that is most UNRELATED to the other three.

o president, action, security, administration
o elect, decision, political, confirm
o nation, threat, trump, deal
o secretary, congress, immigrant, defense

Answer: [**Not Available in the Real Test**] "elect, decision, political, confirm"

Test HIT #5

Please click on the word set below that is most UNRELATED to the other three.
o family, legislation, children, bill
o news, today, said, report
o governor, state, press, call
o federal, emergency, post, continue

Answer: [**Not Available in the Real Test**] "family, legislation, children, bill"

Test HIT #6

Please click on the word set below that is most UNRELATED to the other three.
o west, proud, community, learn
o university, force, military, nation
o student, service, member, thank
o veteran, help, need, support

Answer: [**Not Available in the Real Test**] "veteran, help, need, support"

Test HIT #7

Please click on the word set below that is most UNRELATED to the other three.
o administration, deal, threat, state
o small, support, community, economy
o create, continue, import, protect
o job, business, help, work

Answer: [**Not Available in the Real Test**] "administration, deal, threat, state"

Test HIT #8

Please click on the word set below that is most UNRELATED to the other three.
o justice, elect, supreme, political
o decision, protect, general, constitution
o right, law, court, senate
o bill, american, tax, people

Answer: [**Not Available in the Real Test**] "bill, american, tax, people"

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

6.4 Training Module for Label Intrusion

Completing this training module qualifies you to complete Label Intrusion HITs

Basic instructions
1. For each HIT, you will see ONE SHORT PASSAGE and FOUR labels.
2. Three of the labels will be related to the passage, but the other one will be out of place.
3. Your job is to pick up the one label that seems LEAST related to the passage.

Background

The passages you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Using a computer algorithm, we divided these posts into different categories based on the words they contain. The labels you see are our tentative summaries of those issues or topics.

We want to see if you can identify the label that does not belong with the post.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will highlight the important text and provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

> *Alexander Cosponsor's Cut, Cap, and Balance Act: At a time when we're borrowing 40 cents of every dollar we spend, I will support serious proposals like the "Cut, Cap and Balance Act" to reduce out-of-control Washington spending. The final version of any such legislation should have an appropriate balance between reductions in both entitlement and discretionary spending, but this bill is a good start.*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Government Spending
o Legislation
o American Politics
o Emergency

Answer: The correct answer is "Emergency". The passage talks about a legislation on the federal budget cut. Thus, the labels "Legislation" and "Government Spending" are clearly relevant. This is an important issue in “American Politics,” which is also relevant as a label. The remaining label “Emergency,” however, is out of place because the budget cut is usually not an emergent event.

Practice HIT #2

> *Earlier this week I spoke at a town hall in support of Referendum 74 at First Baptist Church in Seattle. Stand with me – and thousands of Washington families – and click below to volunteer in support of marriage equality!*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Local Events
o Public Meeting
o Environment
o Human Wellbeing

Answer: The correct answer is “Environment”. The passage is in support of a state referendum on gay marriage. It also references a town hall (a local political meeting) held at a church. Two word sets “Local Events” and “Public meeting” relate to the fact that the post references the Senator appearing at a local political gathering. Another word set “Human Wellbeing” seems to be related to the fact that the issue concerns “marriage” and “families.” The final label ”Environment,” however, is not mentioned.

Practice HIT #3

> *The London Metropolitan Police Department has great taste in cars! They’re patrolling the streets of London in BMW X5s made in – you guessed it – South Carolina.*
>
> *I’m in London this weekend attending the annual Farnborough International Airshow in support of the Boeing 787 Dreamliners also made in South Carolina. #PalmettoPride #PalmettoPrideintheUK*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Economy
o Local Events
o Presidential Politics
o Police

Answer: The correct answer is “Presidential Politics”. The passage expresses pride in the cars and airplanes made in South Carolina. Those products are about “Economy.” South Carolina is local and the airshow belongs to “Local Events.” Since the police department is mentioned, the label ”Police” should also be relevant. “Presidential Politics” has no real connection with the passage.

Practice HIT #4

> *Today's executive order from President Trump is more about extreme xenophobia than extreme vetting. This executive order is the equivalent of a "Keep Out" sign posted at America's borders.*
>
> *Turning away immigrants based on their nationality and religion is un-American and in direct opposition to everything for which our Founding Fathers fought. President Trump may not call it a Muslim ban, but it is, and runs afoul of our morals and values.*
>
> *We must ensure that we have the strongest safeguards in place to keep terrorists from ever reaching our shores. And we must fully and thoroughly vet all refugees to screen out any potential terrorist threats.*
>
> *But as conflict and war force millions around the world from their homeland, the United States should welcome more refugees, not less. Suspending the U.S. refugee resettlement program will endanger refugees' lives and tear families apart.*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Immigration
o Local Events
o Presidential Politics
o Security and Defence

Answer: The correct answer is "Local Events". The passage is about President Trump's executive order on immigrants. The labels "Presidential Politics" and "Immigration" are thus clearly relevant. The sentence "We must ensure that we have the strongest safeguards in place to keep terrorists from ever reaching our shores." implies that "Security and Defence" are also concerned. "Local Events" is not clearly related to the discussion of national immigration policy.

Practice HIT #5

> *Women in New Mexico and across the country deserve to make their own decision about family planning. Laws to protect a women's right to choose what's best for her body and well-being should not be restricted to what state she lives in. On the 42nd anniversary of the landmark Roe v. Wade decision, I remain committed to strengthening women's reproductive rights and freedom to make their own personal health care choices. #Roe42*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Legislation
o Environment
o Healthcare
o Daily Life

Answer: The correct answer is "Environment". The passage advocates for a pro-choice position in legislative politics and thus "Legislation" is clearly relevant. It frames abortion as personal "Healthcare" choices (family planning). The label "Daily Life" is triggered by the mention of families. “Environment” is not relevant to the passage.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

> *The unexpected and tragic attacks on Pearl Harbor that occurred 75 years ago today–a date that even now still lives in infamy–triggered America to go to war. As a result of the unprovoked attacks in 1941, many of our brave servicemen and women lost their lives. Today and every year on December 7th, I hope we each take a moment to reflect on and pay tribute to the sacrifices made by those who were injured or lost at Pearl Harbor. And let us each re-dedicate ourselves to ensuring their lives were not lost in vain by honoring and upholding the quintessential American values and freedoms they were fighting for.*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o American History
o Security and Defence
o Education
o Honoring Veterans

Answer: [Not Available in the Real Test] “Education”

Test HIT #2

> *Today Sen. Judd Gregg and I introduced The Bipartisan Tax Fairness and Simplification Act of 2010, which will help middle-class taxpayers by streamlining and modernizing the outdated tax code. The proposal includes fiscally responsible tax cuts to help working families struggling to make ends meet and also eliminates the corporate tax break that encourages companies to invest overseas rather than creating jobs in the U.S.*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Immigration
o Legislation
o Economy
o Tax and Budget

Answer: [Not Available in the Real Test] “Immigration”

Test HIT #3

> *Obama simultaneously could ruin Putin's day and brighten the lives of millions of Americans. All Obama needs is the courage to tell the environmental Left to let him do the right thing.*
>
> *Today is the last day for the public to comment on whether the State Department should approve a presidential permit for the #KeystoneXL pipeline. This article should give you some ideas of things to include in a comment. Here is where you can leave your comment: [link omitted]*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Human Wellbeing
o Environment
o Presidential Politics
o Honoring Veterans

Answer: [Not Available in the Real Test] "Honoring Veterans"

Test HIT #4

> *Look, here is the bottom line. We remain the only nation in the industrialized world that doesn't guarantee health care to all people. We have 29 million people who are uninsured, yet we are spending far, far more per capita on health care than do the people of any other country.*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Healthcare
o Education
o Government Spending
o Human Wellbeing

Answer: [Not Available in the Real Test] "Education"

Test HIT #5

> *In Warroad, MN, Hockeytown USA, with 3 proud U.S. Hockey silver medalists: Rubin Bjorkman (1952), Gordon Christian (1956) and Henry Boucha (1972). Warroad boasts 7 Olympic hockey medalists with Gigi Marvin and TJ Oshie set to win more. Oshie's grandpa played on 1948 Warroad team! Go USA hockey!*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Healthcare
o Announcement
o Game
o Celebration

Answer: [Not Available in the Real Test] "Healthcare"

Test HIT #6

> *I wish Democrats would show some interest in how and why President Obama and Susan Rice got it so wrong about the true nature of the Benghazi attack. I wish Democrats would show a little interest about why Secretary Clinton was clueless about the multiple security requests coming from Benghazi and how she allowed our mission to become a death trap. I wish Democrats would show a little interest in finding out whether Mike Morell, the former #2 at the CIA, lied to Congress and the American people about a protest that never happened. Democrats seem to be more interested in protecting the Obama Administration than they are in getting the truth.*
>
> *If Republicans win a Senate majority in 2014, one of the first things I will insist on is hearings that actually get to the bottom of what happened before, during, and after the Benghazi attack.*
>
> *The families of those who lost loved ones, and the American people, deserve nothing less than a full accounting.*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o The Public
o Political Parties
o Presidential Politics
o Economy

Answer: [Not Available in the Real Test] "Economy"

Test HIT #7

> *Tomorrow, I'll be attending the White House bipartisan summit on health care reform hosted by President Obama. Share your ideas on health care reform with me now – so I can share your feedback with President Obama and Congressional leaders on Thursday!*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Honoring Veterans
o Healthcare
o American Politics
o Public Participation

Answer: [Not Available in the Real Test] "Honoring Veterans"

Test HIT #8

> *Yesterday I met with Anjali Lall and Maneesh Apte, U.S. Presidential Scholar recipients who recently graduated from Davies High School in Fargo, ND. The United States Presidential Scholar Program annually awards up to 141 high school graduates nationwide, honoring students who demonstrate exceptional academic achievement. Congratulations Anjali and Maneesh!*

Please read the four labels below and click on the label that is most UNRELATED to the passage.

o Honorary Statement
o Education
o President
o Economy

Answer: [**Not Available in the Real Test**] "Economy"

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

6.5 Training Module for Optimal Label

Completing this training module qualifies you to complete Optimal Label HITs.

Basic instructions

1. For each HIT, you will see ONE SHORT PASSAGE and FOUR labels.
2. Your job is to pick up the one label that seems BEST summarizing the passage.

Background

The passages you see are taken from U.S. senators' official Facebook postings. These postings can be on any issue or topic. Some postings are about certain legislation, some are about federal or statewide events, and some are about the administration, etc.

Using a computer algorithm, we divided these posts into different categories based on the words they contain. The labels you see are our tentative summaries of those issues or topics.

We want to see if you can identify the label that best summarizes the post.

Attention

Some of these choices will be very clear, but others will require you to use your best judgment. It is critical that you read each word grouping carefully. Scanning or reading quickly will result in low-quality evaluations, and you may be blocked from continued participation in this study (and future studies).

Part 1

The next 5 questions are example HITs.

We will highlight the important text and provide you with the correct answer. These will not be scored and will not count for or against your qualification.

Practice HIT #1

> *Alexander Cosponsor's Cut, Cap, and Balance Act: At a time when we're borrowing 40 cents of every dollar we spend, I will support serious proposals like the "Cut, Cap and Balance Act" to reduce out-of-control Washington spending. The final version of any such legislation should have an appropriate balance between reductions in both entitlement and discretionary spending, but this bill is a good start.*

Please read the four labels below and click on the label that BEST summarizes the passage.

o Insurance
o Legal/Law
o Immigration
o Budget Cut

Answer: The correct answer is "Budget Cut". The passage talks about a legislation on the federal budget cut. The label "Budget Cut" best summarizes it. "Legal/Law", "Immigration", and "Immigration" are not relevant.

Practice HIT #2

> *In Warroad, MN, Hockeytown USA, with 3 proud U.S. Hockey silver medalists: Rubin Bjorkman (1952), Gordon Christian (1956) and Henry Boucha (1972). Warroad boasts 7 Olympic hockey medalists with Gigi Marvin and TJ Oshie set to win more. Oshie's grandpa played on 1948 Warroad team! Go USA hockey!*

Please read the four labels below and click on the label that BEST summarizes the passage.

o Economy
o Holiday Greetings
o Celebration Messages
o Information Sources

Answer: The correct answer is "Celebration Messages". The passage celebrates a town called "Warroad" for having many hockey medalists. "Celebration Messages" is thus the optimal label. The purpose of this message is not greeting people, so "Holiday Greetings" is not the best answer. While providing people with some information, the passage is not mainly about "Information Sources." "Economy" is not relevant.

Practice HIT #3

> *Yesterday I met with Anjali Lall and Maneesh Apte, U.S. Presidential Scholar recipients who recently graduated from Davies High School in Fargo, ND. The United States Presidential Scholar Program annually awards up to 141 high school graduates nationwide, honoring students who demonstrate exceptional academic achievement. Congratulations Anjali and Maneesh!*

Please read the four labels below and click on the label that BEST summarizes the passage.

o Law Enforcement
o Honoring Veterans
o Education
o Presidential Politics

Answer: The correct answer is “Education”. The passage congratulates a student for receiving the U.S. Presidential Scholar Award. It is clearly an educational issue and, thus, the label “Education” is optimal. Although the award’s name has the word, presidential, in it, the passage is not about presidential politics. “Law Enforcement” and “Honoring Veterans” are not implied.

Practice HIT #4

> *The unexpected and tragic attacks on Pearl Harbor that occurred 75 years ago today–a date that even now still lives in infamy–triggered America to go to war. As a result of the unprovoked attacks in 1941, many of our brave servicemen and women lost their lives. Today and every year on December 7th, I hope we each take a moment to reflect on and pay tribute to the sacrifices made by those who were injured or lost at Pearl Harbor. And let us each re-dedicate ourselves to ensuring their lives were not lost in vain by honoring and upholding the quintessential American values and freedoms they were fighting for.*

Please read the four labels below and click on the label that BEST summarizes the passage.
o Terrorist Attack
o Honoring Veterans
o Water Pollution
o American History
Answer: The correct answer is “Honoring Veterans”. The passage honors victims of the attacks on Pearl Harbor and specifically mentions “brave servicemen and women.” “Honoring Veterans” is the best summary. Although the attack took place 75 years ago, this message is not mainly about “American History.” The Pearl Harbor attack is not a “Terrorist Attack.” “Water Pollution” is not relevant.

Practice HIT #5

> *The Red River Corridor Fund received more than $1.1 million in federal funding to incentivize private companies to help small businesses and local economies grow. When our small businesses have an environment they can thrive in, it creates more jobs and helps grow our communities and state. These funds will help encourage more private-public partnerships across the state so small businesses, private lenders, and the federal government can work collaboratively to build North Dakota’s already growing economy.*

Please read the four labels below and click on the label that BEST summarizes the passage.
o Unemployment Rate
o Economy
o Local Issues
o Healthcare
Answer: The correct answer is “Economy”. The passage advertises that North Dakota received federal funds supporting small businesses and local economies. “Economy” is relevant. “Unemployment Rate” is attempting given that the passage mentioned “creates more

jobs". However, it is not optimal because it cannot cover other aspects of the economy such as "funding" and "private-public partnerships." "Local Issues" is too vague. "Healthcare" is not relevant.

Part 2

The next 8 questions are your test HITs.

You must answer at least 7 of the test HITs correctly to receive the qualification.

Test HIT #1

> *I'm hosting a live telephone town hall call this Monday, April 24 at 5:45pm MT. If you're available to talk on the phone, sign up to get a call. You can also participate and ask questions by streaming the event online. Click on the link below to do both: [link omitted]*

Please read the four labels below and click on the label that BEST summarizes the passage.

o Drug Abuse
o Legislative Politics
o Education
o Town Hall Meeting

Answer: [Not Available in the Real Test] "Town Hall Meeting"

Test HIT #2

> *Today, Congress has unanimously passed the Clay Hunt SAV Act to improve veteran suicide prevention programs. We must do everything we can to ensure our veterans and military families have access to high-quality mental health services. By improving treatment, intervention, and outreach, the Clay Hunt SAV Act will help more veterans get the help that they need and have earned. No veteran should ever have to wait for critical mental health care. Learn more: [link omitted]*

Please read the four labels below and click on the label that BEST summarizes the passage.

o Veteran Healthcare
o Economy
o Drug Abuse
o Nation Building

Answer: [Not Available in the Real Test] "Veteran Healthcare"

Test HIT #3

> *Today I met with Judge Neil Gorsuch, nominee to serve on the U.S. Supreme Court. Judge Gorsuch is a smart, diligent and thoughtful jurist. I am impressed by his remarkable commitment to the Constitution and the separation of power our founders envisioned. Oklahoma is within the jurisdiction of the 10th Circuit*

> *Court of Appeals, on which Judge Gorsuch sits, and I have seen his judicial philosophy first hand. In Burwell v. Hobby Lobby, he wrote a concurring opinion in favor of the Oklahoma company's position, and upholding our First Amendment right to religious freedom. Judge Gorsuch is a qualified mainstream jurist who is well respected by conservatives and liberals alike and I urge Senate Democrats to allow and up or down vote on Gorsuch. I look forward to confirming him to the bench.*

Please read the four labels below and click on the label that BEST summarizes the passage.

o Supreme Court
o Congratulation Messages
o Military
o Religious Freedom

Answer: [Not Available in the Real Test] "Supreme Court"

Test HIT #4

> *Recent actions by Iran have once again demonstrated their reluctance to curb their confrontational nuclear missile program. These dangerous, provocative actions by a country considered a state sponsor of terror should carry consequences. The decision to increase sanctions should come at no surprise. A line was drawn, and Iran deliberately defied the agreement. I stand by the President on his decision to implement a quick and deliberate strategy. It should send a clear message that Iran's reckless behavior will not be tolerated.*

Please read the four labels below and click on the label that BEST summarizes the passage.

o Justice
o Security
o Honoring Veterans
o Government Spending

Answer: [Not Available in the Real Test] "Security"

Test HIT #5

> *Deadline extended: January 23 is the new deadline to apply for federal disaster assistance from Hurricane Matthew. Register with FEMA Federal Emergency Management Agency online: [link omitted]*

Please read the four labels below and click on the label that BEST summarizes the passage.

o Emergency Management
o Holiday Greetings
o Foreign Policy
o Public Hearing

Answer: [Not Available in the Real Test] "Emergency Management"

Test HIT #6

> *Interested in attending one of our nation's military academies? My office is currently accepting applications from North Dakota students interested in attending the U.S. Military Academy at West Point, the U.S. Naval Academy, U.S. Air Force Academy or the U.S Merchant Marine Academy. The deadline is Oct. 5.*

Please read the four labels below and click on the label that BEST summarizes the passage.
o Healthcare
o Doctor
o American History
o Application
Answer: [**Not Available in the Real Test**] "Application"

Test HIT #7

> *Today I brought together local business development and community leaders to discuss the essential role of rural communities in North Dakota's economic success and ways to support new investment and growth in small towns across the state. Energy development in the Bakken has created unprecedented growth but we have to make sure this development's success flows across the entire state. Rural communities like Ellendale are an important part of that effort, and by supporting hardworking North Dakotans and small businesses in these areas, we can continue to sustainably grow our state for future generations.*

Please read the four labels below and click on the label that BEST summarizes the passage.
o Local Businesses
o Nation Building
o Infrastructure Construction
o Public Meeting
Answer: [**Not Available in the Real Test**] "Local Businesses"

Test HIT #8

> *While under oath during his confirmation hearing, Attorney General Jeff Sessions misled his colleagues in U.S. Senate about his communications with Russia Ambassador Sergey Kislyak. He should resign immediately. This is exactly why we need an independent special counsel to investigate the Russian hacking of our presidential election and any ties the Trump campaign may have had to Russia.*

Please read the four labels below and click on the label that BEST summarizes the passage.
o Family Memory
o Information Sources
o Games
o Legal Institutions and Offices
Answer: [**Not Available in the Real Test**] "Legal Institutions and Offices"

Before you submit your answers...

You will only have 1 chance to take this test. Make sure that you are satisfied with all of your answers above before submitting.

If you become qualified to participate in the HITs, please continue to fully read each future HIT and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, you may be blocked from continued participation in this study (and future studies).

6.6 Workers' Performance

To assess workers' performance, we randomly mixed in a gold-standard HIT into every ten HITs. Once posted a batch, we checked its progress sporadically. 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. Among all workers who have completed HITs in the batches demonstrated in the main manuscript, we have rejected and replaced work from only three of them who missed more than 4 gold-standard HITs each. These thresholds (2 and 4) are not iron rules. Future researchers can decide on their own standards based on the difficulty of their designed gold-standard HITs.

The gold-standard HIT approach yields consistent results. In the main manuscript, none of the identical pairs we presented are significantly different from each other.

Another way to assess the consistency of workers' performance across identical trails is to assess the "agreement rate," which we present in Tables SI6 and SI7. The numbers reflect the frequency that workers from the two trials get the task either right or wrong simultaneously, as opposed to one get it right but the other get it wrong.

Table SI6: Workers' Agreement Rate in the Topic Validation Tasks

	Model 1	Model 2	Model 3	Model 4	Model 5
WI	0.816	0.696	0.724	0.696	0.684
T8WSI	0.62	0.606	0.68	0.72	0.704
R4WSI	0.714	0.754	0.85	0.884	0.728

We realize that these numbers are hard to interpret. In general, a high agreement rate is desirable. However, the agreement rate would, theoretically, be a function of the correct rate. This introduces a floor effect. Consider the case, for instance, where workers got 95% of the HITs correct, agreement rates can only go as low as 90%. For this reason, we did not include these tables in the main text, but only as supplementary information.

Table SI7: Workers' Agreement Rate in the Label Validation Tasks

	Careful Coder	Cursory Coder
LI-Within	0.764	0.746
LI-Across	0.864	0.896
OL-Within	0.836	0.824
OL-Across	0.904	0.856

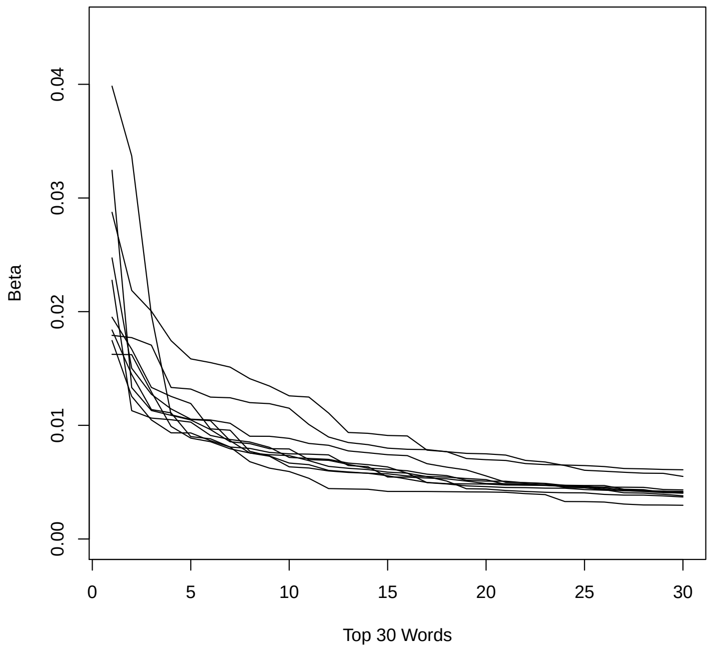
7 More on the Corpus, Topic Model Fit, and Labeling

This section presents more background information and discussions on the structural topic models in the paper and their labeling process. Specifically, we first discuss the distributions of word probability in the topic models, which is the reason we chose to draw words based on their probability in the topic validation tasks. We then present some word clouds to give people a different perspective of the topics in different topic models. Finally, we present the representative documents based on which we came up with the topic labels.

7.1 Word Mass Distribution

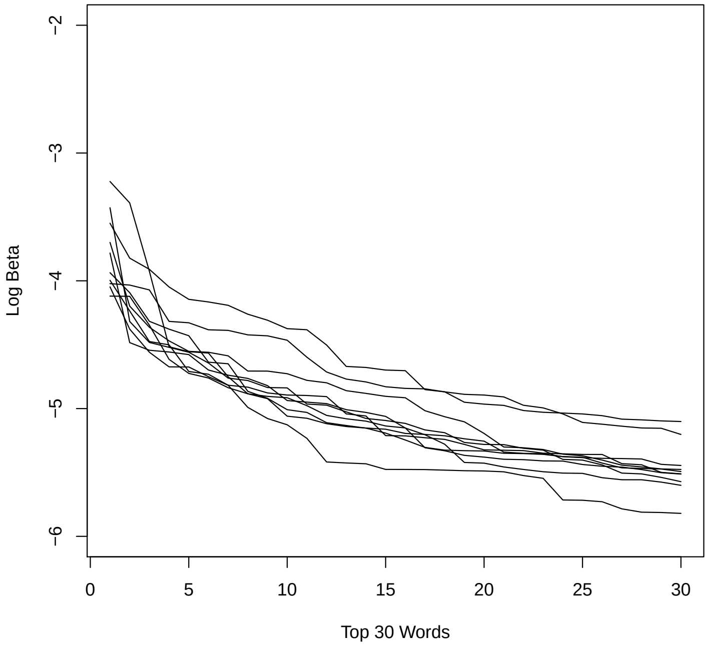
One major difference between our topic validation tasks and those from Chang et al. (2009) is that we randomly draw words based on their probabilities in a given topic from a given model. We made this choice because the word probabilities could vary a lot among the top 20 high probability words. Figures SI25 and SI26 demonstrate this point with Model 2 (the 10-topic model) from our paper, where `beta` refers to the word probability.

Figure SI25: Word Mass: Beta



Furthermore, Figure SI27 depicts the total word probability covered by the top n words for the four converged topic models (Model2, Model3, Model4, and Model5) from our paper. The shapes of the curves are similar, although the specific numbers vary across different models. On average, except Model2, the top 20 words cover about 50% of the total word probability.

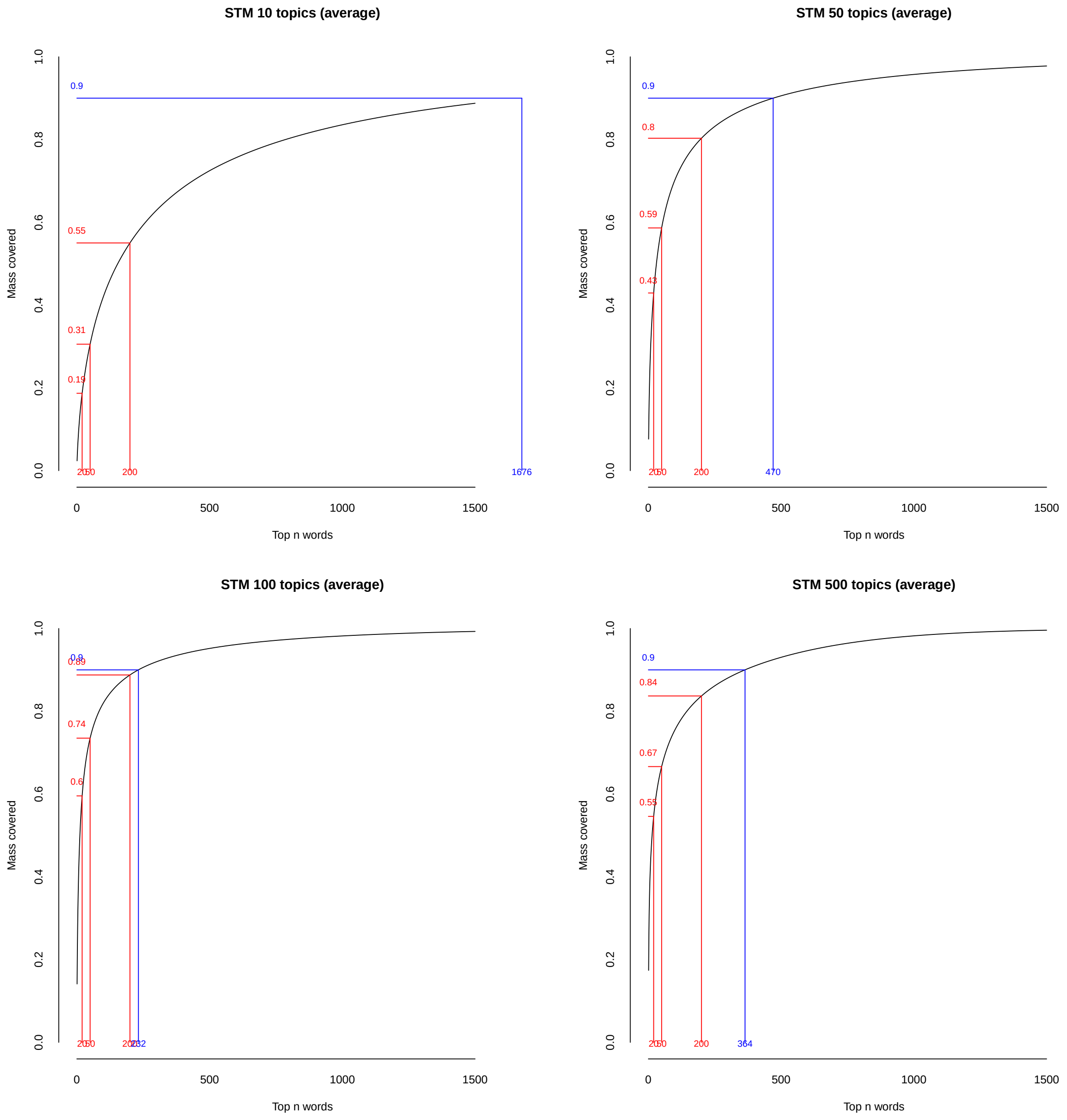
Figure SI26: Word Mass: Log-Beta



7.2 Word Clouds

For the four structural topic models validated in the paper, we randomly select 6 topics from each of them and present the word clouds below. Just from the word cloud, it does not seem obvious that the four converged models (Model2, Model3, Model4, and Model5) are of different quality.

Figure SI27: Word versus Mass



STM 10 topics (average)

STM 50 topics (average)

STM 100 topics (average)

STM 500 topics (average)

Figure SI28: Word Cloud: Six topics from **model 1** (the 10-topic model, 1 iteration)

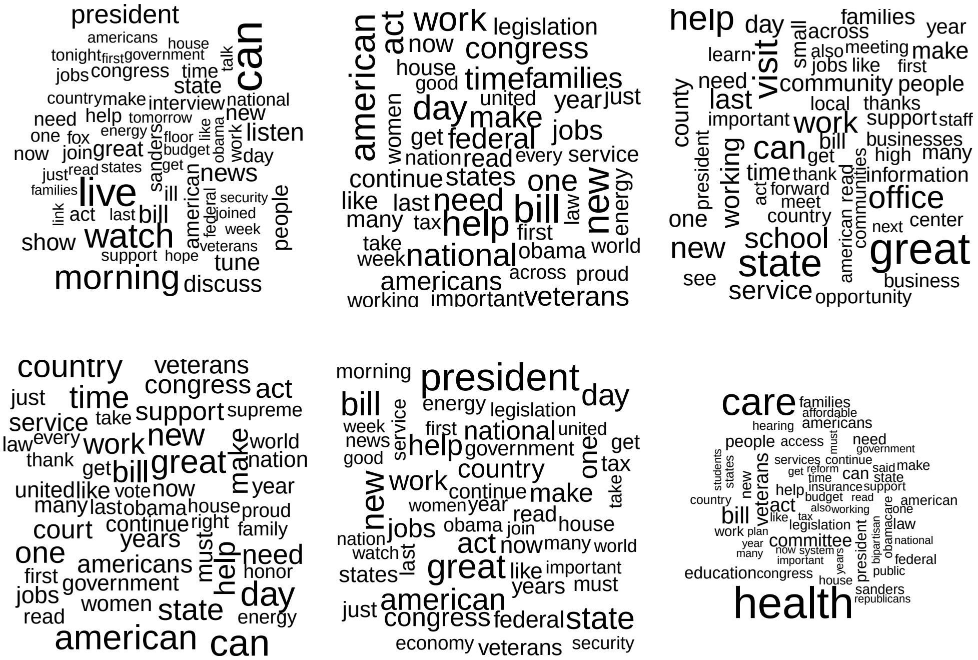


Figure SI29: Word Cloud: Six topics from **model 2** (the 10-topic model)

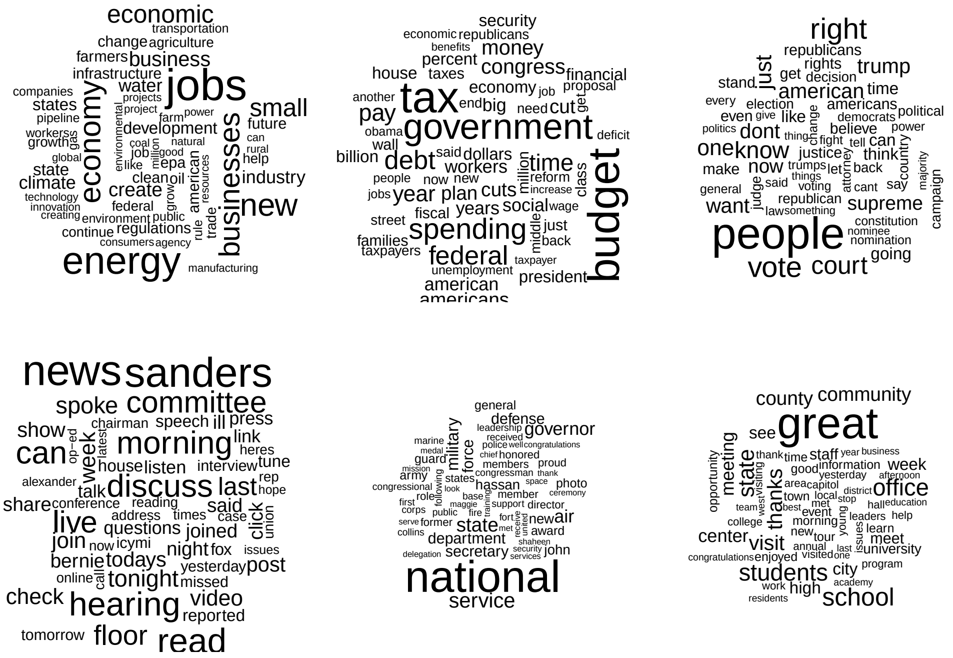


Figure SI30: Word Cloud: Six topics from **model 3** (the 50-topic model)

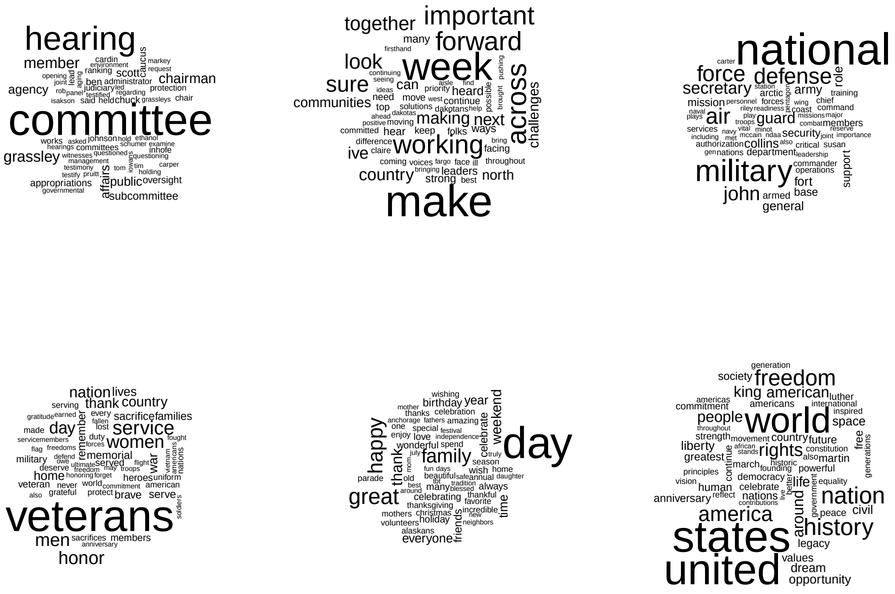
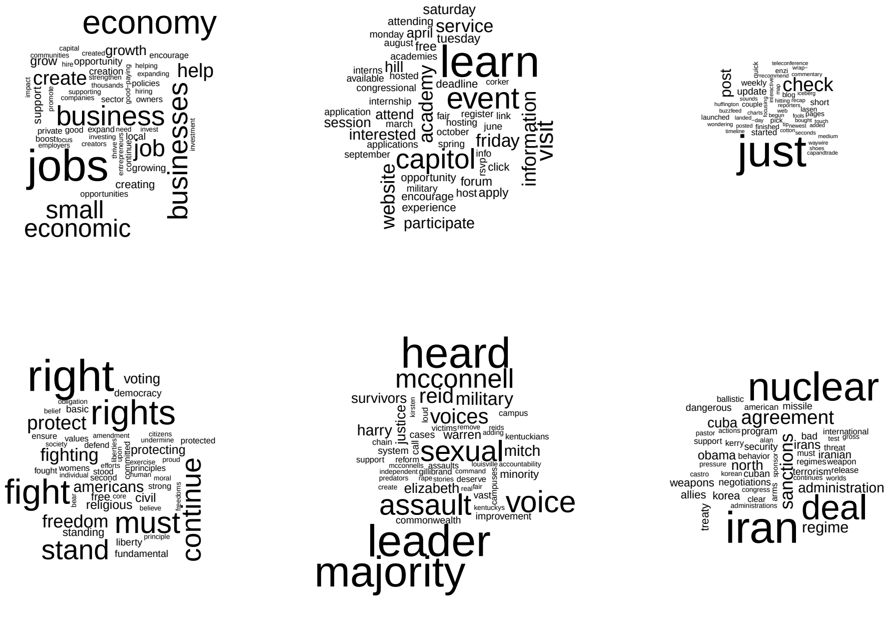
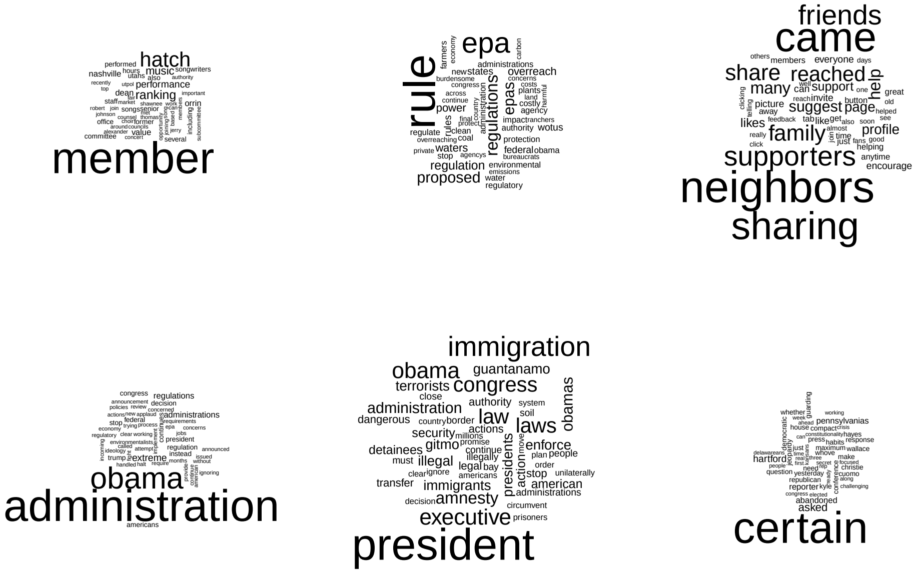


Figure SI31: Word Cloud: Six topics from **model 4** (the 100-topic model)



SI: 67

Figure SI32: Word Cloud: Six topics from **model 5** (the 500-topic model)



7.3 Representative Documents for Labeling

For the label validation tasks, we asked the two careful coders to read carefully the top 50 representative documents. Table SI8 shows the top 5 documents for each domestic topic with which they come up with the labels.

Table SI8: Representative Documents for Domestic Topics

Label (Careful Coder)	Document
Equal Pay for Women	(1) In Indiana, women working full-time make an average of 75 cents for every dollar earned by their male colleagues. Women in Indiana and around the country deserve to earn equal pay for equal work. When women are paid less, it not only impacts their ability to save for retirement but also means families have less to spend on groceries, rent, and other basic expenses. #EqualPayDay (2) Sweat, sweat, sweat! Work and sweat, cry and sweat, pray and sweat! Zora Neale Hurston #Hardwork

	(3) Women earn less than men regardless of any other factor. I support #EqualPay for equal work & that's why I voted for the Paycheck Fairness Act that builds on the Lilly Ledbetter Fair Pay Act by closing loopholes that enable pay discrimination against women. Share this graphic if you agree that women should earn equal pay for equal work! (4) Today is Equal Pay Day. North Dakota women earn 73 cents for every $1 a man earns. I support the Paycheck Fairness Act because there should be equal pay for equal work. Like/share if you agree. (5) In Michigan, women earn 74 cents for every dollar men earn. The Gender Pay Gap doesnt just hurt women - it hurts entire families, affecting retirement, health care coverage, pensions and much more. Today is Equal Pay Day, a date that symbolizes how far into the year women must work to earn what men earned in the previous year. It is time to end the Gender Pay Gap and pay women fair wages for their work.
Healthcare/ Reproductive Rights	(1) Planned Parenthood provides critical access to health care for thousands of Montanans. #StandWithPP (2) Courtney from Wisconsin stands with Planned Parenthood because without them, she wouldn't have access to quality and affordable womens health care. TrumpCare rolls back the clock on womens health by cutting funding for maternity care and defunding Planned Parenthood. (3) I #StandwithPP for the millions who depend on their critical health care services. #PinkOut (4) California health centers received funds to expand primary care services by hiring qualified health care providers. (5) Planned Parenthood is providing women across Michigan with critical preventative care, including cancer screenings and wellness exams. Today, I met with Planned Parenthood Advocates of Michigan to let them know I #StandWithPP and the millions of women who rely on them for quality health care. #PinkOut
Agriculture	(1) Antibiotic overuse unchecked: "About four-fifths of all antibiotics sold in the U.S. go to livestock and poultry." (2) Montana ranchers produce the best beef in the world and we cant have beef imports from Brazil and Argentina jeopardizing the livelihood of Montana producers. (3) Today we recognize and thank all who contribute to Iowas agriculture industry, among them Iowas farmers, producers, ranchers: [Link Omitted] #AgDay (4) SenatorLeahy: #VT farmers, sugar makers and rural farms benefited from REAP in 2008 Farm Bill. #FarmBill 2012 continues this good work.

	(5) Senator Shaheen discussed improvements made to the dairy program in this years farm bill during a tour of McNamara Dairy farm in Plainfield. Employees of McNamara Dairy milk 175 Holstein cows, bottle their own milk and sell it to customers up and down the Connecticut River Valley. (Plainfield, NH. July 20, 2012)
Student Loan/ Debt	(1) #ObamaEconomy: Inflation-adjusted median household income has fallen from $54K in '08, to only $52K in 2013 while inflation-adjusted per capita income has fallen from $29,173 in 2008 to $28,829 (2) Recent college graduates are struggling in this slow economic recovery. We must keep student loan interest rates at the current rate and prevent them from doubling, but we must pay for it. (3) In our latest installment of Correspondence from the Commute, Chris replies to Bailey from Hockessin about Joseph Kony and the LRA. (4) Americans have over $1.3 trillion of student debt. My In #TheRedAct will allow struggling borrowers to refinance their student loans and take advantage of lower interest rates the same way people refinance a mortgage, a car loan or business debt. (5) FACT: Total student loan debt now surpasses both credit card & auto loan debt. #HigherEdNotDebt
Drug Abuse	(1) DEA Implements Cornyn-Klobuchar Law to Help Curb Prescription Drug Abuse: (2) Today, I teamed up with Senator Sheldon Whitehouse in hosting a roundtable with the Alliance to Prevent the Abuse of Medicines on how we can work together to combat prescription drug abuse and help struggling families and communities in Ohio. To prevent drug abuse and better help the tens of thousands of Ohioans struggling with addiction, we need a comprehensive strategy that starts from the bottom up. Our bipartisan Comprehensive Addiction and Recovery Act builds on proven methods to enable law enforcement to respond to the heroin epidemic and supports long-term recovery by connecting prevention and education efforts with treatment programs. Like my video to join me in the fight against drug abuse. [Link Omitted] (3) Today I re-introduced the Budgeting for Opioid Addiction Treatment (LifeBOAT) Act to establish reliable funding to expand access to substance abuse treatment in West Virginia. (4) Help prevent the misuse of addictive prescription drugs on Prescription Drug Take-Back Day at a location near you [Link Omitted] (5) KSDK NewsChannel 5 previews Claire's roundtable discussion today at NCADA to tackle the opioid and heroin epidemic with advocates, providers, and law enforcement.

Higher Education/ Job Training	(1) Very productive college affordability roundtable with students from University of Wisconsin-Parkside, Gateway Technical College, Carthage College & University of Wisconsin - Whitewater. Higher education should be a path to shared prosperity, not a path into suffocating debt. (2) Honored to be at Northwestern Michigan College yesterday for the expansion of their Community College Skilled Trades Equipment Program. With this new program, NMC will be able to offer wider array of educational opportunities - from marine and aviation tech to welding and nursing. (3) Happy to join Community College of Rhode Island (CCRI) today to announce a $2.5 million federal grant to create new pathways and opportunities in advanced manufacturing. CCRI's Accelerated Pathways in Advanced Manufacturing program emphasizes a learn and earn model that provides opportunities for adult learners to acquire new knowledge and skills that are linked with jobs in high-growth industries. (4) As part of her efforts to help New Hampshire workers develop the skills and innovative thinking needed for jobs in the 21st century economy, Governor Maggie Hassan announced today that five New Hampshire companies have been awarded job training grants to help them train 171 workers in new skills. The job training grants total $144,973.50, and the companies contributed matching funds to bring the total amount for training workers to $289,947. (5) Continuing her efforts to help New Hampshire workers develop the skills, knowledge and innovative thinking necessary for jobs in the 21st century economy, Governor Hassan announced today that five companies will receive job training grants to help them train 125 workers in new skills. The job training grants total $72,536.48. The five companies contributed matching funds, bringing the total funds for training workers to $145,072.96.
Wall Street/ Financial Sector	(1) SenJohnMcCain: Wall Street bonuses: Goldman Sachs $16.7 billion; JP Morgan $8.7 billion; main street $0 your tax doll... (2) This recession was caused by the greed, the recklessness and the illegal behavior on Wall Street. [Link Omitted] (3) Will question CEO Jamie Dimon about J.P. Morgan's $2 billion plus losses at a Senate Banking hearing around 11am. Tune in here: [Link Omitted] (4) I'm working with Consumer Financial Protection Bureau head Rich Cordray to protect consumers from predatory for-profit schools & big bank tricks. (5) No, AIG. Criticism of Wall Street bonuses is not equivalent to lynchings. Offensive.
Government Shutdown/ Congressional Budget	(1) Republicans in Congress are passing bills to reopen and fund essential functions of our government. It's the Democrats in Congress who refuse to negotiate. #HarryReidsShutdown

	(2) How does federal government claim preemption in Arizona when federal government fails to act? (3) This week, Puerto Rico argued before the Supreme Court that it should be able to restructure its debt. If Puerto Rico were a country, it could go to the IMF. If it were a city, it could declare bankruptcy. But because Puerto Rico is a U.S. territory, it can do neither. Without help, Puerto Rico is at the mercy of Wall Street vulture funds, creditors that specialize in preying on borrowers who are in trouble. The Republican leadership needs to step in immediately and help millions of American families who are caught in an economic catastrophe. (4) Skipping the budgeting, authorizing and appropriating committee process in lieu of last-minute, omnibus, must-pass, stop-gap bills diminishes the role of Congress. It diminishes the role of individual members of Congress and reduces the input of the people from that members state or district. By skipping steps, Congress allows the Administration to regulate and spend with little fear of being checked by the legislative branch. Our countrys founders gave Congress the power of the purse. The Senate and House should right the ship no matter what party controls the White House. Using the committee system and considering legislation as it should be considered would go a long way. It would put a stop to a lot of partisanship, some of which is simply frustration from leaders preventing members from legislating. (5) An essential premise of good government is that Congress should authorize programs and activities before it funds them. If we relinquish our responsibility to regularly review and reform these programs, all of our government funding will essentially operate on auto-pay.
Obamacare/ Tax Policy	(1) #99countymeetings 50ppl at Mapleton Q&amp;A EPAregs VoterID MedicareFraud Obamacare/SupremeCt Farmbill F&amp;F ElectionFinanceRecorm SocSec (2) Courtesy of ObamaCare: Broken promises, plagued by failure & cost skyrocket for millions. (3) #ObamaCare creates fewer options at "much higher premiums, higher out-of-pocket costs, higher taxes on the costs that they[everyday Americans] do incur, and fewer jobs and fewer hours for those who are employed." (4) Millions who lost their insurance and reenrolled through OCare are facing higher premiums: [Link Omitted] (5) Premiums are soaring, patients choices are dwindling. Obamacare must be repealed and replaced. #RepealObamacare
Deficits/ Debt/ Budget	(1) The President's budget proposes deficits of $5.3 trillion and $8.1 trillion in new debt over the next decade. Where's the Balanced Budget? (2) Why we must cut spending and why simply raising taxes won't solve our fiscal problems

	(3) Today the Congressional Budget Office released a report that shows positive impact of the budget caps. If our nation is serious about balancing our budget and reducing Americas debt, real, substantive budget reforms and savings will have to be on the table during any spending negotiations. (4) Voted NO on raising debt ceiling without addressing spending problem. Need to both prevent a debt limit crisis today & debt crisis tomorrow. (5) The combination of new spending with the President's proposed budget freeze will only maintain an unsustainable status quo. We need serious cuts, not a continuation of the trillion dollar deficits that are jeopardizing our nations fiscal stability.

Table SI9 shows the top 5 documents for each domestic topic with which they come up with the labels.

Table SI9: Representative Documents for International Topics

Label (Careful Coder)	Document
International Trade	(1) Sherrod reintroduces China trade plan to discourage currency manipulation that harms American exports. As our trade deficit continues to widen, our need to level the playing field for American manufacturers and workers becomes more urgent. (2) RT @GrassleyOffice: Heres #TheScoop by @ChuckGrassley [Link Omitted] #4jobs @RedCross @Mail4Heroes #fastandfurious #NationalGuard ... (3) Chinas economy is simply too large for it to be artificially propped up by a blatantly manipulated Yuan. Chinas actions are deliberate and are designed to give China a competitive advantage in the marketplace. I believe free and fair trade is beneficial, but I also know that Chinese currency manipulation and intellectual property theft is doing harm to our economy. (4) I am pleased that the U.S. Trade Representative took action against Chinas unfair trade practices. Its policies give Chinese companies an unfair advantage and hurt American workers and products. (5) Today, I urged the International Trade Commission to protect Ohio Steelworkers. Ohio has the best steelworkers in the country and they deserve to compete on a level playing field with our competitors. That is why I am working to promote exports of Ohio steel and protect our steelworkers from unfair competition: [Link Omitted]
Praising Active Military/ Military Units	(1) 150 soldiers from the 1st Stryker Brigade Combat Team, 25th Infantry Division arrived home to Fort Wainwright yesterday from a deployment in Afghanistan. Welcome home! Check out these great homecoming pictures:

	(2) The South Dakota Army National Guard 196th Maneuver Enhancement Brigade returned after a 10-month deployment to Kuwait. I was honored to welcome them home today. (3) Welcome support from Hollywood to #SaveSaeed Abedini and Kenneth Bae. (4) Welcome home from Afghanistan SPC Jessica McKim, The U.S. Army! (5) Yesterday, the 82nd Airborne Division welcomed its 49th commander, Maj. Gen. Erik Kurilla. Join me in thanking Maj. Gen. Kurilla and the 48th commander, Maj. Gen. Richard Clarke for their service to the 82nd Airborne Division.
Terrorism	(1) Whether its in Jerusalem, or Paris, or New York, the civilized people of the world are under siege by violent Islamic extremists. We must redouble our efforts to win this war on terrorism. In 2016, we must remain vigilant against Islamic extremism. We need to do everything in our power to protect Americans from terrorism. (2) Washington must wake up to this fact: the crowd-sourcing of domestic terror is a reality. While slow-moving, federal bureaucracies look for card-carrying terrorists, the Islamic State and al Qaeda are crowd-sourcing their jihad. We must adapt and develop a long-term strategy to name and defeat our enemy. That enemy is not the empty label "extremism" but the ideology of militant Islam.

	(3) This weekend as Americans celebrate the birth of our nation and our freedoms, the horrific events yesterday are a sobering reminder of the radical Islamic terrorists who seek to destroy us. Twenty more innocent civilians are dead in the most recent attack by radical Islamic terrorism this time the target was an upscale neighborhood in Dhaka, #Bangladesh. Terrorists throwing grenades and shouting Allahu Akbar attacked a popular restaurant frequented by diplomats and students, reportedly slitting the throats of some of their victims and singling out those who could not recite verses from the Koran. Over the past few years, the authorities in Bangladesh have insisted that terrorist incidents in their country were somehow isolated from the global scourge of radical Islamic terrorism, the product of homegrown militants responding to domestic grievances. This attack, however, appears designed by ISIS to prove otherwise, and that the Islamic State is functioning not only in Syria and Iraq but also in Bangladesh, just as it is in Turkey, France, Belgium and the United States. This isnt just a wake-up call anymore it is a screaming siren of an alarm that demands our urgent attention. Just as the Bangladesh attack was a demonstration of the Islamic States determination to take their war on civilization to another Muslim-majority nation, the recent attacks in the United States from San Bernardino to Orlando demonstrate the United States is, likewise, not immune. We are fortunate to have friends and allies in Europe, the Middle East, and Asia who see this enemy for what it is, and we should do everything we can to partner with them not to struggle against generic violent extremism but to actually take the fight to the radical Islamic terrorists. (4) The attack at the Istanbul Ataturk Airport in Turkey is yet another tragic reminder of the threats we face from terrorists who indiscriminately employ senseless violence with no regard for innocent lives or their own. It's a reminder that we must remain vigilantnot only here at home, but also around the globein our efforts to prevent terrorist threats at our transportation hubs and other vulnerable locations. (5) While this is a milestone that we have all awaited, we must remember that al Qaeda and its affiliates are not dependent on one man and we must remain vigilant in our efforts to disrupt and destroy terrorist networks that threaten our Nation and allies.
Military Sexual Assault	(1) Momentum continues to build in support of the Military Justice Improvement Act, which would create an unbiased military justice system by transferring the decision-making authority over which cases go to trial from the chain of command to independent trained military prosecutors where it belongs. I hope you'll join me and my colleagues in supporting this commonsense measure to reform the way the military handles sexual assault cases so that victims can get the fair shot at justice they deserve. (2) Military sexual assault victims were silenced in Congress. Lend your voice in support of the Military Justice Improvement Act to give them a fair shot at justice. #MJIA

	(3) Yesterday, I stood alongside sexual assault survivors, advocates, and a bipartisan coalition of U.S. Senators to renew our push for real reform of the military justice system. Our bipartisan Military Justice Improvement Act would remove prosecutorial authority in sexual assault cases from the chain of command and place it in the hands of independent trained military prosecutors where it belongs. Only then will we truly create an unbiased system of military justice where assailants are held accountable and survivors, like Samantha Jackson, get the justice they deserve. (4) Of the estimated 26,000 incidents of unwanted sexual contact and sexual assaults that occured in 2012, only 3,374 were reported. Time and again, victims tell us the reason there is such a drop off in reporting is because the decision-making in cases of sexual assault lies within the chain of command. Until this authority is moved outside the chain of command, into the hands of an independent military prosecutor, victims will not have the confidence they will receive the justice they deserve. This is why I'm fighting to pass the Military Justice Improvement Act, to make sure the voices of victims–like Navy veteran and MST survivor Brian Lewis–are heard and that we create an objective and accountable military justice system. Please share this graphic to let Brian's voice, and that of so many others, be heard. #MJIA (5) Victims of sexual assault in the military tell us time and time again that the reason they don't report sexual assaults is because the decisionmaking lies within their own chain of command. If we want to truly reform the military justice system and give victims of sexual assault a real chance at justice, we must remove this decisionmaking power from the chain of command and place it in the hands of independent trained military prosecutors. Right now, victims' voices are not being heard in Congress, such as U.S. Marine Vet & #MST survivor Stacey Thompson. Please lend your voice in support of the Military Justice Improvement Act and make sure the voices of victims like Stacey are heard: [Link Omitted] #MJIA
Nuclear Deterrance/ International Security	(1) Today, Iran test-fired ballistic missiles in violation of international sanctions. This launch demonstrates the regimes continued disregard for international sanctions and the grave consequences of this Administration blindly trusting the Iranian regime to uphold its commitments. Instead of making concessions to the Iranian regime and freeing up more than $100 billion dollars in sanctions relief, the Obama Administration should change course and increase pressure on Iran by imposing additional unilateral sanctions to stop their belligerent behavior once and for all. (2) The more we learn about the Iran deal, the more concerned I am that it not only fails to prevent Iran from obtaining a nuclear weapon, it emboldens them through tens of billions of dollars in sanctions relief, a phased out lifting of a UN arms embargo and the ability to test more advanced centrifuges.

	(3) The long list of concessions in the Presidents Iran deal includes the end of a long-standing arms embargo. This misguided concession, coupled with the immediate and imprudent sanctions relief, makes this not only a weak deal, but a dangerous one, too. (4) Even President Obama is admitting that Iran would be able to develop a nuclear weapon in "a matter of months" after the Iran deal expires. I strongly oppose the deal, which paves the path for Iran to become a nuclear power and threatens the security of Israel and our partners in the Middle East. (5) Freeing American prisoners of conscience in Iranian prisons should be a pre-condition to any negotiations with Iran. #FreeAmir
Air Force	(1) Today, Brig. Udo K. "Karl" McGregor, commander of the 452nd Air Mobility Wing, March ARB, Calif., passed the 730th Air Mobility Training Squadron guidon to Lt. Col. Jonathan M. Philebaum, 730th AMS commander, Altus AFB, during the squadron reactivation and assumption of command at hangar 517. (Photo courtesy of U.S. Air Force) (2) It was an honor to spend Friday at Pease Air National Guard Base, onboard the KC-135A Stratotanker with the New Hampshire National Guard. During the refueling mission, I watched as the crew honed in on a EC-130J Commando Solo aircraft not far from Pittsburgh and, with laser like precision, guided a refueling hose into the aircraft. (3) Advocated for Warriors of the North-Grand Forks Air Force Base, Minot Air Force Base, and North Dakota National Guard during meeting with U.S.Airforce. (4) U.S. Sen. Cory Booker and the Adjutant General of the New Jersey National Guard, Brig. Gen. Michael L. Cunniff, discuss the flight parameters of an F-16 fighter jet during a visit to the 177th Fighter Wing in Egg Harbor Township, N.J. on Aug. 5, 2016. It was the first visit to the New Jersey Air National Guard base by Sen. Booker, who received a briefing by the Wing Commander, U.S. Air Force Col. John R. DiDonna, flew an F-16 simulator and got a chance to get an up close look at an F-16C Fighting Falcon fighter jet on the flight line at the Atlantic City Air National Guard base. (U.S. Air National Guard photo by Master Sgt. Andrew J. Moseley/Released) (5) Enjoyed meeting with Col. Gentry Boswell, Commander of the 28th Bomb Wing at Ellsworth Air Force Base and Command Chief, CMSgt Sonia Lee today to discuss the status of the B-1 fleet at Ellsworth.
Honoring Specific Veterans	(1) What an honor to present the Distinguished Flying Cross to WWII Veteran and Hudson-area resident Alfred LeFeber. A long overdue recognition for a true patriot. (2) Yesterday, I was honored to present Roger Piasecki of Kearney with a WWII Medal & Lapel Pin for his father Walter's military service. #GreatestGeneration

	(3) Today I participated in a ceremony honoring Mildred Pretzer of Charlottesville, a World War II veteran who served as part of the Womens Army Corps (WAC). I proudly presented Ms. Pretzer with the Army Good Conduct Medal, the WAC Service Medal, the American Campaign Medal, the World War II Victory Medal and the Honorable Service Lapel Pin. (4) Today, I had the privilege of escorting James Gray, a WWII Veteran from Malvern as he visited the WWII Memorial for the first time. Mr. Gray is a former POW and Purple Heart Recipient who fought in the Battle of the Bulge where he was captured and spent six months as a German POW. Thank you for your service to our great country, Mr. Gray! (5) It was an honor to present lost medals to World War II Veteran and hero Julious Elmore in Magnolia today.
Honoring Veterans/ Heroes	(1) Honor the heroes who ran into burning buildings. Honor the brave who sacrificed their lives for strangers. And never forget those lost on 9/11/2001 and 9/11/2012. (2) Today we remember and honor the brave men and women that sacrificed their lives for our freedom. We are forever grateful and you will never be forgotten. (3) Thank you to the men and women who have worn, and continue to wear, the uniform of our nation's military. As America pauses today to remember the service and sacrifice that our veterans and active duty members have endured, we are forever in debt to you and your families for your selfless service. (4) Today, let us pause to remember the heroes we have lost in the service of our great nation. Thank you to all who serve to protect our freedom. (5) This #MemorialDay, we honor the brave men and women who paid the ultimate sacrifice to preserve our nation's freedom
Military Operations/ Armed Conflicts	(1) Sen. Corker visited the Kilis refugee camp on the Turkish border with Syria today for the second time since the conflict began. Members of the camps refugee leadership expressed dismay and disappointment at the lack of U.S. support for the Syrian opposition. Specifically, refugee leaders noted the importance of helping arm and equip the vetted, moderate opposition under the leadership of Gen. Salim Idriss. (2) Congressional authorization for the use of military force against ISIS, Al-Qaeda, and the Taliban will make clear to our warfighters, our allies, and our adversaries that we are united #AUMF (3) U.S. Senator Tim Kaine leads a Congressional Delegation (CODEL) to the Middle East region focused on the U.S. mission against ISIL and the ongoing humanitarian crisis in Iraq and Syria. (4) The drawdown of US troops in Iraq and use of force in Libya without a plan have enabled ISIS to amass power and territory with little pushback.

	(5) ISIS has possession of and is now using illegal chemical weapons against the Kurds in Iraq. Earlier this year, I visited Iraq and met with the leaders of Iraqi Kurdistan, and I firmly believe that the U.S. must increase its support for the Iraqi Kurdish Peshmerga forces who are valiantly fighting against ISIS.
Veterans Affairs/ Veterans Healthcare	(1) I have been and will continue to be committed to reforming the VA to ensure our veterans receive the care they deserve. For our veterans in need of help with the VA, my office is here to help! (2) Senator Grassley has made it clear that the comments from the Secretary of Veterans Affairs comparing wait times for VA treatment with wait times for rides at Disneyland were unacceptable and that the Secretary should make amends. Senator Grassley has worked to improve veterans experience at the VA, including pushing for and receiving an acknowledgement from the VA that veterans have experienced problems with the Veterans Choice program and a pledge to improve services to Iowa veterans and veterans across the country. (3) "Our veterans deserve the best medical care available, but far too often, veterans suffering from PTSD, Traumatic Brain Injury, and other psychological impacts slip through the cracks." (4) The Department of Veterans Affairs needs to fix problems in helping veterans get appointments outside the VA when needed. Senator Grassley is working with the VA on veterans' frustrations. (5) Today, I toured the Portland Vet Center with Steven Reeves from the U.S. Department of Veterans Affairs. Our veterans deserve the highest quality care possible, and our 5 Vet Centers in Maine need to be adequately staffed in order to provide vital services to those who served our country.

References

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.

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.

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.

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.

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.

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.

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, 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.