CDC Study Text Analysis:

Initial Findings and Figures

Context on the topic modeling framework used:

Structural Topic Model (‘STM’) is topic modeling algorithm that is used to identify latent topics within a large corpus of text data (Roberts et al., 2019). STM generates topic model solutions based on document-topic and topic-word distributions. Specifically, as Roberts et al. (2019) explain, “A topic is defined as mixture over words where each word has a probability of belonging to a topic. And a document is a mixture over topics, meaning that a single document can be composed of multiple topics. As such, the sum of topic proportions across all topics for a document is one, and the sum of the word probabilities for given topic is one”.

In this study, we follow steps for estimating and evaluating a structural topic outlined by Roberts et al. (2019). In our data, each participants’ reason stated is treated as a document in which the STM algorithm generates a topic model for all reasoning data based on words and their co-occurrences used by participants, and then by clustering word groupings (e.g. topics) that best describe the reasons participants stated. That is, by accounting for word frequency and the distance between words used in the participants’ stated reasons, the topic model infers the topics that are expressed in the reasoning data. In addition, one advantage of structural topic models, unlike other topic models models (Blei et al., 2003), is that STM is able to account for covariates in the broader data set in which the text was generated that may be of interest to a researcher. Specifically, covariates that a researcher believes may account for prevalence of word use or content can be controlled for with STM.

**Reasons For Data**

The following figures and tables represent outcomes derived from a 3-topic solution. The following graphics are the pieces of evidence that we would use as a team to decide which topic model solution to use. Deciding between topic model solutions is both a subjective and objective (e.g. fit statistics) process. My goal in this write up was to present a sample of the most commonly used pieces of information that researchers use to evaluate topic model fit. What we will do is compare the following 3-topic solution with alternative models; say 2, 4, and 5-topic models. Through subjective and data-driven evaluation, we ultimately build a case for the best possible representation of the latent constructs that we believe are in the sample data.

**Figure 1**

*Highest Word Probabilities for Each Topic in Reasons For Decision*

Chart, bar chart

Description automatically generated

Figure 1 shows the words that are most common in each of the 3 topics identified by this model. This graphic is a good way to subjectively assess whether the model output is semantically meaningful, which is to say, do these word groupings meaningfully represent a latent construct?

**Figure 2**

*Diagnostic Values by Number of Topics*

Diagram, engineering drawing

Description automatically generated

Figure 2 presents common metrics for selecting and evaluating the best topic models from a data-driven perspective. This step is similar to a scree plot in EFA. To generate these plots, the diagnostic metrics were calculated for topic models that were fit ranging from k = 2 to k = 10 topics. One important metric is semantic coherence, which is a measure of quality within topics. Specifically, semantic coherence is maximized when a topic model converges on a solution in which high probability works for a topic tend to naturally co-occur in the documents that the model is fed (Mimno et al., 2011). We can see in Figure 2 that semantic coherence is best at k = 2 and worse at k = 10. Held-out likelihood represents how well each model represents words in the documents, where lower values for held-out likelihood are preferred. The same number of recommended topics is given with regards to held-out likelihood.

**Figure 3**

*Expected Frequency of Topic Occurrence Across Participants’ Reasons*

Chart

Description automatically generated

Moving on with the three topic solution, there are a host of follow up analyses that are run to iteratively evaluate the goodness-of-fit of the 3-topic model. Figure 3 shows the expected frequency of an identified latent topic in the data set. For our study, this means that topic 2, which may have something to do with keeping safe and healthy (more on this later), is expected to occur around 35% of the time in our sample data.

**Figure 4**

*Words Associated with Topics Related to Reasons For Decision by Fit Metric*

Graphical user interface, text

Description automatically generated

Figure 4 is similar to the histogram in Figure 1 in that the words most associated with a given topic are shown. However, this figure also shows the most common words for a given topic when conditioning on a specific fit criteria. For example, in the figure, ‘marginal FREX’ weights words by their overall frequency andd how exclusive they are to the topic, therefore giving higher weights to words that appear less frequently in other topics (Roberts et al., 2019)

**Table 1**

*Correlation Between Topics*

A picture containing text, scoreboard

Description automatically generated

Correlations between topics are presented in Table 1. Positive correlations indicate that both topics are likely to be discussed within a document (Roberts et al., 2019). Notice, the strong negative correlations. This is a good sign for the 3-topic solution as we would expect that topics would not co-occur too frequently in one participant reason. The reasons that participants state are short, so it is less likely that two or more constructs were expressed in participants’ reasons.

**Figure 5**

*Estimated Effects of Topics Related to Reasons For Decision on Attitudes*

Text

Description automatically generated

Figure 5 shows the estimated relationship between participants’ attitudes and the latent topics identified by the 3-topic model solution.

**Figures 6 – 8**

*Estimated Reasons For Topic Proportion for Level of Participant Attitude*

Chart, line chart

Description automatically generated

Note: Participant attitude\_ave is plotted on the x-axis; y-axis represents expected proportion of Topic 1; These graphs show the estimated proportion of a topic occurring across all reasons stated for a given level of attitude. In the first graph, we see that the likelihood of Topic 1 occurring goes down as attitude towards decision goes up.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

**Reasons Against Data**

In parallel to the evaluation and initial analyses above, the figures below were generated to assess a 3-topic model solution for the *reasons against* data.

**Figure 9**

*Highest Word Probabilities for Each Topic in Reasons Against Decision*

Chart, bar chart

Description automatically generated

**Figure 10**

*Words Associated with Topics Related to Reasons Against Decision by Fit Metric*

Text

Description automatically generated

**Figure 11**

*Estimated Effects of Topics Related to Reasons Against Decision on Attitudes*

Text

Description automatically generated

**Figure 12**

*Estimated Reasons Against Topic Proportion for Level of Participant Attitude*

Chart, line chart

Description automatically generated

**References**

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v091.i02