**Text-mining open-ended survey responses using structural topic modeling:
A practical demonstration to understand parents' coping methods during COVID-19
pandemic in Singapore**

Gerard Chung¹, MSW

Maria Rodriguez², MSW, PhD

Paul Lanier¹, MSW, PhD

Daniel Gibbs¹, MSW, JD

¹The University of North Carolina at Chapel Hill, School of Social Work

²University of Buffalo, School of Social Work

Correspondence regarding this article should be address to Gerard Chung Siew Keong, MSW, School of Social Work, University of North Carolina at Chapel Hill, 325 Pittsboro Street CB# 3550, Chapel Hill, NC 27599. Email: gcsk1982@live.unc.edu

**Text-mining open-ended survey responses using structural topic modeling:
A practical demonstration to understand parents' coping methods during COVID-19
pandemic in Singapore**

Abstract

Objective: Open-ended survey questions crucially contribute to researchers' understandings of respondents' experiences. However, analyzing open-ended responses using human coders is labor-intensive and prone to inconsistencies. Structural topic modeling (STM) is a text mining method that discover topics from textual data. We demonstrate the use of STM to analyze open-ended survey responses to understand how parents cope during COVID-19 lock-down in Singapore. **Method:** We administered online surveys to 199 parents in Singapore during the COVID-19 lock-down. To show a STM analysis, we demonstrated a workflow that includes steps in data preprocessing, model estimation, model selection, and model interpretation.

Results: An 18-topic model best fitted the data based on model diagnostics and researchers' expertise. Prevalent coping methods described by respondents include "Spousal Support", "Routines/Schedules" and "Managing Expectations". Topic prevalence for some topics varies with respondents' levels of parenting stress and whether parents were fathers or mothers.

Conclusion: STM offers an efficient, valid, and replicable way to analyze textual data such as open-ended survey responses and case notes that can complement researchers' knowledge and skills. STM can be used as part of a multistage research process or to support other analyses such as clarifying quantitative findings and identifying preliminary themes from qualitative data.

Keywords: Parenting Stress; COVID-19 Coronavirus; Structural Topic Modeling STM; Text-mining; Machine learning; Singapore

TEXT MINING OPEN-ENDED SURVEY RESPONSES

Survey research is one of the most common data collection methods in social work (Rubin, 2013). Unlike observational methods, surveys rely on what respondents report, rather than on observations of what respondents do (Rubin, 2013). To collect respondents' responses, most surveys in social work research use close-ended questions that offer a limited set of response options (e.g., Likert response items) for respondents to choose among (Krosnick, 1999; Rubin & Babbie, 2013). These response options are typically categories generated by the researcher or a concept they employ. Far fewer surveys use open-ended questions to understand respondents' attitudes and beliefs as expressed in their own words. One reason for this comparative lack of open-ended survey questions is that the work of coding and analyzing unstructured data from open-ended survey questions is labor-intensive. Moreover, human coding can generate inconsistent data (especially when there are multiple coders) and can make the data susceptible to researcher biases (Jackson & Trochim, 2002). In general, researchers consider close-ended survey questions "easier to ask, code, and analyze" (Geer, 1991, p. 360).

Despite these issues, we argue that social work research stands to benefit from more frequent use of open-ended survey questions. Compared to close-ended survey questions, qualitative open-ended survey questions can better capture respondents' experiences in their own words; they are better at identifying concepts and perspectives that researchers do not know about or know little about; and they provide opportunities for respondents to explain or clarify quantitative findings from close-ended survey questions (Jackson & Trochim, 2002; Pietsch & Lessmann, 2018; Roberts et al., 2014). Although one aim of research is to produce knowledge that is generalizable across people and settings, understanding the experiences of individuals – especially social work clients – remains a hallmark goal of social work research (Thyer, 2012).

Recent developments in computational social science and textual digitalization have given social work researchers new methodological tools (e.g., see work by (Patton et al., 2020; Rodriguez & Storer, 2020) to help them analyze unstructured textual data. The aim of the present study is to demonstrate the use of a text-mining tool, Structural Topic Modeling (STM), in analyzing open-ended survey responses collected from a sample of 199 parents in Singapore. We asked these parents to describe in their own words how they were coping with parenting stress in a lockdown setting during the COVID-19 pandemic. Our following discussion demonstrates that STM is an efficient, replicable, and effective method for identifying different types of coping used by parents based on an analysis of their self-reported open responses to an open-ended survey question.

Challenges to Analyzing Open-Ended Survey Questions

Geer (1991) argued that close-ended questions dominated survey questions, not because open-ended questions were inherently weaker indicators. Rather, responses from open-ended questions proved more difficult and demanded more resources to code and analyze than responses from closed-ended questions. The process of coding qualitative data by human coders typically involves the following steps: first, the researcher defines the themes and provides examples to the coders demonstrating how to code the textual data. The researcher often relies on existing theoretical knowledge or selected responses from the samples to generate a codebook. Second, human coders are trained to use the codebook and they began coding the textual responses using the codebook. Third, if discrepancies in coding occur, the coders and the researcher discuss and resolve them. Finally, the researcher will check the final coded results for accuracy in coding and agreement between the human coders. Although these four steps improve consistency in coding among multiple coders, the analysis is time-consuming, laborious, and

costly (Jackson & Trochim, 2002; Roberts et al., 2014). Further, coding decisions made by researchers and human coders can still yield errors or inconsistencies, undermining the reliability and validity of the results. Consequently, because of these challenges in analyzing open responses, researchers have tended to avoid using open-ended questions, and closed-ended questions have come to dominate most surveys (Geer, 1991; Roberts et al., 2014). However, excluding open-ended questions from surveys deprives researchers of the valuable contributions that open-ended response data can make to social work research.

Benefits of Open-Ended Survey Questions

The key function of open-ended survey questions is to give respondents opportunities to describe their experiences in more detail without the restrictions or biases inherent to fixed-choice format questions (Burg et al., 2015). Open-ended responses are useful for gathering new information about a topic, supplementing quantitative findings, and exploring different dimensions of respondents' experiences (Jackson & Trochim, 2002). When a topic is new and research knowledge is limited, exploratory open-ended questions can help to elicit the information needed to define a closed question or inform the development of future studies (Pietsch & Lessmann, 2018). Open-ended questions also better reflect the types of questions that respondents typically hear in everyday life (e.g., "How good was the movie?"), as compared to close-ended questions that only offer respondents a limited set of responses (e.g., "On a scale of 1 to 4, how would you rate the movie?"). Because open-ended questions have fewer cues and are less likely to be biased by researchers' analytical or methodological framework, respondents' may also exhibit lower research reactivity to the questions (Roberts et al., 2014).

Despite these benefits, there are several reasonable critiques of open-ended survey questions. For instance, open-ended survey questions have lower response rates than close-ended

questions because open-ended questions test people's ability to articulate a response. However, Geer's (1988) study of surveys from the National Election Studies in the United States showed that the few individuals who did not respond to the surveys' open-ended question(s) appeared uninterested in the question posed; it was not because they lacked the eloquence to respond. Scholars have also observed that respondents may give superficial or casual responses to open-ended questions that do not represent salient attitudes (Geer, 1991). However, Geer (1991) showed by experiment that this criticism may be overstated and that open-ended questions can accurately measure the attitudes of respondents. Despite these valid criticisms, our view, echoing that of Roberts et al. (2014), is that the open-ended questions crucially contribute to researchers' understandings of respondents' experiences.

Structural Topic Modeling

In the era of big data analytics and rapid text digitalization tools, researchers can more efficiently analyze large volumes of texts and documents from sources such as the Internet, social media, electronic patient records, and digitized governmental records. Because conventional human coding is often not practical for analyzing large amounts of data (DiMaggio et al., 2013), researchers need an efficient, effective, and reproducible method to (a) simplify the data and (b) produce valid interpretations of that data (Allahyari et al., 2017; DiMaggio et al., 2013; Lindstedt, 2019).

For this reason, many social scientists are turning to *text mining*, a computational method for extracting meaningful knowledge from large amounts of textual data (see work by Perron et al., 2019 and also Victor, in press). For instance, Rodriguez & Storer (2020) analyzed tweets using structural topic modeling to understand the reasons why individuals leave or stay in abusive relationships. Topic modeling (e.g., STM) comprises a group of text-mining techniques

for discovering latent topics contained in textual data based on the co-occurrences of words (explained below). Because topic modeling is computer-assisted (Roberts et al., 2014), researchers can analyze large amounts of textual data efficiently and consistently. The process is also transparent, meaning that researchers can replicate their analyses (DiMaggio et al., 2013). In this study, we demonstrate the use of STM which is one approach among different topic modeling techniques.

STM was developed by Roberts et al. (2013). It builds on the Latent Dirichlet Allocation (LDA) topic modeling algorithm, which inductively identifies latent topics among a large collection of documents and assigns individual documents to one or more topical categories (Blei et al., 2003). In topic modeling, a “topic” refers to a group of words associated with a “semantically interpretable theme” (Roberts et al., 2014, p. 3). For instance, an STM analysis of President Barack Obama’s political speeches may identify latent topics including gun laws, universal health care, and racial inequities. STM also allows each document (e.g., one speech) to consist of multiple topics (i.e., mixed membership models; Roberts et al., 2014) with varying degrees of membership between documents.

STM takes a relational approach to meaning, in that the co-occurrences of words are important in defining topics (Roberts et al., 2014). For instance, the co-occurrence of words like “coaches,” “players,” “soccer,” and “team” in a corpus of newspaper articles would be codified as a “sports” topic. Due to their emphasis on relationality, topic models are able to account for polysemy (i.e., a word’s different contextually based meanings). For example, the term “bat” can co-occur with words like “mammal” and “flying” in one topic, and “baseball” and “pitcher” in another topic: two very different meanings for the same word.

To identify the topics underlying a corpus of documents, STM builds on the LDA algorithm to incorporate a generative statistical model which assumes that the observed documents are produced from a mixture of latent topics (Blei, 2012; Blei et al., 2003). These latent topics (the number of which is defined by the researcher beforehand) then generate associated words based on their probability of co-occurrence. Thus, given a set of texts and a preestablished number of topics, the algorithm attempts to identify what combination of unknown topics could generate those documents (Roberts et al., 2014). Figure 1 provides a visual representation of the STM framework.

[Insert Figure 1 here]

A key feature of STM is its ability to use document-level information or covariates to explain differences in topical prevalence and content across documents. In other words, STM allows researchers to study how the prevalence and content of topics change in relation to information about the document (e.g., metadata such as year or source) or the respondent (e.g., sociodemographic). For instance, the prevalence of topics across President Trump's speeches may differ depending on the year or external events (e.g., COVID-19 in the year 2020).

Collectively, these features of STM make it a suitable method for analyzing textual data from open-ended survey questions. Social science researchers are often interested in discovering the latent qualitative themes of a set of open responses (Roberts et al., 2014), and STM provides a method to do so. Further, because STM is an unsupervised method, researchers do not need to assign topics by hand-coding the textual data based on pre-determined themes (Roberts et al., 2014). However, though STM relies on computer-assisted processes, it still requires the researcher to judge and interpret the model and its results. Next, we discuss a study where we

acquired textual data from parents' responses to an open-ended survey question and analyzed it using STM.

Study of Parenting During Circuit-Breaker Social Distancing in Singapore

To combat the spread of the coronavirus (COVID-19), from April to May 2020 the Singaporean government implemented social distancing and lockdown measures, known as “Circuit-breaker” measures. Under Circuit-breaker, schools and workplaces were closed and parents had to balance telecommuting with parenting responsibilities at home. Coupled with the high degree of economic uncertainty and reduced community contact during the pandemic, researchers hypothesized that parental stress would increase under lock-downs (Brown et al., 2020; Authors, 2020a). Other researchers further warned that increased parenting stress would lead to an increased risk of family violence and child abuse during the period when families remained at home with reduced community contact and social support (Authors, 2020b). Indeed, according to the Parenting Stress Model (Abidin, 1992), parenting stress is a key determinant of dysfunctional parenting behaviors and outcomes (Deater-Deckard, 1998).

To understand the contributing effects of parenting stress among Singaporean parents, our earlier study (Authors, 2020a) demonstrated how the impact of COVID-19 reduced parent-child relationship closeness and increased harsh parenting behaviors through the mediating effects of parenting stress. To further parse out this finding, we also studied the coping methods used by parents to reduce parenting stress during Circuit-breaker. Coping strategies play a demonstrably important role in reducing parenting stress (Deater-Deckard, 1998). Adaptive coping strategies (e.g., emotional and instrumental support from spouses and friends; positive parenting beliefs; problem-focused solving strategies) can help reduce parenting stress (Deater-Deckard, 1998). We wanted to know what kinds of coping strategies Singaporean parents used to

cope with parenting stress generated by the unprecedented Circuit-breaker measures. Given the lack of prior research to guide our inquiry, we used an open-ended survey question that prompted these parents to describe their coping methods in their own words.

Aims of the Present Study

In the present study, we used STM to analyze the textual data obtained from the open-ended responses of 199 Singaporean parents. Our main aim was to demonstrate that STM is an efficient, useful, and replicable method for discovering the latent topics underlying a corpus of open-ended responses. We hope that this demonstration of STM's utility will encourage social work researchers to use more open-ended survey questions in their research and use text mining methods as an alternative or a complement to human-driven analyses of textual data.

Method

Data and Sample

We analyzed data from an online survey that we created and disseminated to parents in Singapore from April 22, 2020, to May 5, 2020, the period when Circuit-breaker measures were in effect. To be eligible for the study, respondents had to be at least 18 years old, live in Singapore with at least one child 12 years old or younger, and be Singaporean citizens or permanent residents. We focused on parents with young children because schools and childcare facilities were ordered to close during Circuit-breaker, significantly affecting parents with children from this age group. Only one respondent from each household completed the survey.

The online survey was disseminated using a website link hosted on a Qualtrics server. We reached potential respondents via advertisements on Facebook with a link to the online survey and via Facebook pages of family-related organizations and online social groups (e.g. parenting support groups). We also approached community organizations (e.g., Centre for

Fathering) promoting family wellbeing in Singapore to help disseminate the online survey to their databases of parent clients. We chose these methods to disseminate the surveys because these organizations or groups provide information and services that are open to all parents in Singapore, greatly expanding our potential recruitment pool. Before conducting the survey, we provided each respondent with a detailed information sheet explaining the study and obtained consent from all respondents. This study has been approved by the Institutional Review Board at the [blinded for review]. No incentives or compensations were provided to respondents for participating in the survey.

In total, 268 respondents participated in the survey, of whom 208 provided a response to the open-ended survey question. Because of the present study's focus on parents, we excluded caregivers who were not parents ($n = 1$). Because the vast majority of respondents were married, we excluded unmarried respondents ($n = 2$). We also excluded one-word responses ($n = 6$) in the preprocessing stage, leaving a final analytical sample of 199 responses. Table 1 displays the sociodemographic characteristics of this analytical sample. Most parents were mothers (68%), Chinese (84%), from 36 to 40 years old (33%), had at least a university degree (85%), and had a monthly household income greater than 8000 Singapore dollars (58%). Most parents had only one young child at home (46%); most children were younger than six years old (64%).

[Table 1 here]

Measures

The online survey was written in English, consisted of 50 questions (close-ended and open-ended questions), and took about 12 minutes to complete. Survey questions were related to parenting, the impact of COVID-19, and demographic information.

Coping With Parenting

TEXT MINING OPEN-ENDED SURVEY RESPONSES

Respondents were asked to respond to the open-ended question, “What has helped you to cope with parenting in the past weeks?” by typing their responses into a text box. The average number of words for each response was 20.92 ($SD = 23.39$). All responses were in English.

Covariates

We use two covariates to examine topical prevalence across responses: (a) parenting stress and (b) parent gender.

We employed the Parental Stress Scale (PSS) to measure respondents’ perceptions and feelings of stress directly associated with parenting (Berry & Jones, 1995). Earlier validation studies have found that the PSS scale has strong psychometric properties (Berry & Jones, 1995), including an internal reliability of Cronbach’s $\alpha = .89$ in a validation study with a Hong Kong-based sample (Cheung, 2000), and strong criterion validity with established parenting stress scales such as the Parenting Stress Index (Berry & Jones, 1995). Parents responded to statements about their parenting over the past weeks using a 4-point Likert scale (1 = *Never* to 4 = *Often*). Examples of these statements included “Caring for my children take more time and energy than I have to give” and “I sometimes worry whether I am doing enough for my child(ren).” The Cronbach α was .74. A composite score was created by averaging the score across the survey items ($M = 2.39$; $SD = .46$), with higher scores indicating higher parenting stress.

Additionally, respondents reported if they were mothers or fathers and responses were coded as a binary variable (0 = mother; 1 = father).

Structural Topic Model Workflow

As noted earlier, STM is a semi-automated method that allows researchers to uncover latent topics within a collection of documents (Roberts et al., 2019). In the present study, we referred to each respondent’s response to the open-ended question as a “document” (the unit of

analysis) and we referred to the full set of open-ended responses as a “corpus.” Statistically, in the STM framework, a topic is a probability distribution over words where each word has a probability of belonging to a topic. And a document is a probability distribution over topics (a single document can be composed of multiple topics; Roberts et al., 2019). For a technical discussion of the statistical model employed by STM, we refer readers to Roberts et al. (2016).

Although the primary goal of textual analysis may be to discover topics within a document or corpus, social work researchers often have additional information about the documents (e.g., publication year) or the respondents (e.g. sociodemographic data). The relationships between these observed covariates and the latent topics may have scientific importance for researchers. STM can incorporate these document-level covariates into the model to investigate (a) the proportion of a document devoted to a topic (i.e., topical prevalence); and (b) the word used in discussing a topic (i.e., topical content). If researchers believe that the covariate will affect how much the respondent will discuss a particular topic (or the saliency of the topic), this refers to topical prevalence. If the covariate is believed to influence the words that a respondent uses to discuss a certain topic, this refers to topical content.

In this study, we examined whether and how topical prevalence differed according to levels of parenting stress (continuous variable) and between fathers and mothers (binary variable). Specifically, we sought to determine if particular coping methods were more salient at higher parenting stress levels and more or less prevalent among fathers or mothers.

Roberts et al. (2019) have provided a helpful workflow for STM that we also used in this study. First, in the preprocessing stage, documents are processed and prepared for analysis. Next, a series of STM candidate models (each with a different number of topics) is estimated and evaluated. Researchers then select the most suitable model and interpret and visualize its results.

Notably, though STM is an unsupervised method, it emphasizes rather than sidelines the researcher's active involvement in interpreting the results and engaging with the documents (Grimmer & Stewart, 2013; Roberts et al., 2014). Since the aim of our paper is to demonstrate the use of STM so as to develop researchers' interest in using the method, we highlighted steps in the workflow which requires the researcher's judgments and interpretation.

Preprocessing the Data

STM analysis can be performed in the R statistical software by using the *stm* package (Roberts et al., 2019). Researchers often use spreadsheets (e.g., .csv format files) to store text data and records of the covariates related to that text. When the text data is read into a data frame in R, the functions “textProcessor” and “prepDocuments” in the *stm* package in R will convert and process the data to prepare it for analysis.

The “textProcessor” function also performs other preprocessing steps to prepare the data, including removing punctuation and numbers, converting all letters to lower case, and stemming words by removing all prefixes and suffixes (e.g., “talking” and “talks” become “talk”). Stop words (e.g., “and,” “the,” and “is”) are also removed from the corpus because they do not contribute to words' contextual meanings or the identification of topics. Researchers may include additional preprocessing protocols according to their experiences or domain expertise. For instance, in the present study, we removed the word “Circuit-breaker” from our corpus because it describes a shared context for all respondents and thus does not contribute to our topic interpretation. We also converted all synonyms of “children” (e.g. “kids”) to “children” to avoid creating redundant topics.

We suggest using the preText package in R developed by Denny and Spirling (2018), which can help researchers to determine the best preprocessing steps using an empirical

approach. Specifically, the preText package allows researchers to examine the sensitivity of findings under different preprocessing choices based on a single metric (i.e., a pretext score). Researchers can then make preprocessing decisions using this metric along with their expertise and substantive knowledge.

In summary, preprocessing steps are important for STM because they prepare the data for computer analysis and increase the interpretability of the latent topics in that data.

Model Estimation and Search

Model estimation comprises an iterative process of running multiple models with k number of topics in order for researchers to determine the optimal model for further analysis. Model estimation can be performed using the “searchK” function in the stm package, which will estimate candidate models with a range of different k -values (specified by the analyst) and compute diagnostic values for each model.

In this initial step of model estimation, the choice of a range for k will depend on the documents’ length and focus. For a shorter, focused corpus (i.e., a corpus including a hundred to a few thousand documents) an initial choice between five and 50 topics is best. For larger, unfocused corpora (i.e., those including tens of thousands of documents), research indicates that a range between 60 and 100 topics is best (Lindstedt, 2019). In comparison to speeches or interviews, open-ended survey responses are focused and shorter (Roberts et al., 2014).

In the present study, we ran a total of 19 models (i.e., from $k = 2$ topics to $k = 20$ topics). Because the results of this model estimation may differ depending on starting values, we used spectral initialization as recommended by Roberts et al. (2019) so that the same results could be generated each run and replicated.

Model Selection

The key task of model selection is to select from among the various candidate models a model that holds semantically interpretable and useful topics. There is no one correct number of topics for a given corpus (Grimmer & Stewart, 2013) because varying the k number of topics also changes the “level of granularity of the [model’s and researcher’s] view into the data” (Roberts et al., 2014, p. 6). Determining the appropriate number of topics requires a combination of diagnostic measures and the researcher’s judgment and expertise (Roberts et al., 2014; an additional step that we do not discuss in this paper is Mimno & Lee, 2014's algorithm to find a starting number of topics. However, this algorithm does not provide a statistical guarantee of a starting value for k , so we move forward without employing it).

One way to make this judgment is to evaluate characteristics of the topic-word distribution across models. Specifically, researchers look for a model characterized by high *semantic coherence* and *semantic exclusivity* (Roberts et al., 2014). A topic is semantically coherent if words with a high probability of appearing in a given topic tend to co-occur within documents (high internal consistency). Using our earlier example, newspaper articles with a sport topic would be defined by words (e.g., “coaches,” “players,” and “team”) that are more likely to occur together in those articles (i.e, documents). More importantly, when a set of words are highly likely to co-occur with high probability in a given topic, but also unlikely to co-occur in other topics, the given topic is considered semantically exclusive. For instance, newspaper articles about sports would have different topic words found in articles about gardening. In STM, “a topic that is both cohesive and exclusive is more likely to be semantically useful” (Roberts et al., 2014). These two qualities also align with Gerring's (1999) “consistency” and “differentiation” criteria for concepts in empirical social science.

The stm package is able to compute semantic coherence and exclusivity using interpretable metrics. We recommend that researchers plot the exclusivity and semantic coherence of the candidate models and then select a model on the semantic coherence-exclusivity frontier: that is, where no model clearly dominates another in terms of both semantic coherence and exclusivity (we demonstrate this later). However, these measures cannot replace the researcher’s insights and judgments, and we recommend that researchers conduct close readings of exemplary documents from each topic in order to assess the quality of a topic model (Roberts et al., 2014). Helpfully, researchers can use the “findThoughts” function in the stm package to print documents highly associated with each topic.

As noted, in the present study, given our corpus of 199 documents, we ran a total of 19 candidate models with $k = 2$ to $k = 20$ topics. Figure 2 shows the semantic coherence-exclusivity plot of candidate models. Researchers should choose among models with high scores along both the semantic coherence and exclusivity dimensions (i.e., models that are nearer the upper right-hand quadrant of the plot: the coherence-exclusivity frontier). As seen in Figure 2, our coherence-exclusivity frontier included the 18-topic model and the 20-topic model. Compared to the 20-topic model, the 18-topic model had lower semantic coherence but higher exclusivity scores, indicating that its topics were more distinct. To choose between these two models, we also calculated and plotted their respective residual, held-out likelihood, and lower bound scores (Figure 3; see Table S1 for diagnostic scores). A lower residual score implies better model fit because a higher residual score may indicate that “more topics are needed to soak up some of the extra variance” (Roberts et al., 2019, p. 38). The held-out likelihood, similar to cross-validation, assesses the model’s prediction performance (a high held-out likelihood is optimal) while the lower bound refers to the lower bound of the marginal loglikelihood (maximised lower bound is

optimal; Rodriguez & Storer, 2020). Ultimately, given the relatively high residuals in 20-topic model, we determined that the 18-topic model was appropriate for our subsequent analyses.

[Figure 2 and 3 here]

Researchers should also examine the documents and words highly associated with each topic in order to assess the semantic usefulness of the topics generated by each model. The stm package allows researchers to extract exemplary documents associated with the topics in order to conduct close readings. Researchers should also pay special attention to the words most associated with each topic. Researchers can call out the highest probability (PROB) words and frequent and exclusive (FREX) words for each topic. PROB words are words that the model estimates have the highest probability of association with each topic. However, because higher frequency words are likely to have higher probability of being associated with a topic, researchers should also look at FREX words, which are high-probability words weighted by their overall frequency and exclusivity to a given topic.

Table 2 shows the PROB and FREX words for our $k=18$ topic model (see Figure S1 supplementary for the $k=20$ model). When we examined the topic words in the $k=20$ model, we noted that several words such as “spouse” and “parent” were highly associated with multiple topics that were close in meaning, which explained this model’s lower exclusivity scores compared to the $k=18$ model. As such, we were affirmed in our decision that the $k=18$ model was a more appropriate model based on our quantitative metrics and our evaluation of the documents and topics.

[Table 2 here]

Model Interpretation and Visualization

After choosing the appropriate model, the researcher has to interpret the model.

Researchers can interpret the meaning of the topics in one of two ways. The first way is to use the set of words highly associated with each topic (i.e, PROB and FREX words) to interpret and name the topics. The second way is to read in full the documents that are highly associated with each topic.

For instance, in the present study, we found that Topic 12 was highly associated with the topic words “spouse,” “support,” and “involvement.” By extracting and reviewing the documents highly associated with Topic 12 (see Table 3), we found that respondents mentioned that spousal support, quality of communication, and high spousal involvement helped participants to cope with parenting stresses resulting from lockdown measures. Accordingly, we classified Topic 12 as a topic on “Spousal Support.” The validity of our classification was further supported by prior literature on social support and parenting alliance as sources for coping with stress among parents (Abidin, 1992).

[Table 3 here]

Similarly, our examination of documents and topic words (i.e., “schedule,” “routine,” and “stay”) associated with Topic 8 (see Table 3) showed that this topic involved sticking to routines, planning ahead, and scheduling. We classified Topic 8 as the “Routines/Schedules” topic, an interpretation suggesting that creating routines, plans, and schedules was one coping method which respondents used to cope with parenting stress. Evidence supports this interpretation. One qualitative study of mothers in the United Kingdom who worked from home found that developing daily and/or weekly work timetables or schedules helped these mothers cope with stress arising from the competing demands of work and parenting (Crosbie & Moore, 2004).

Though this process generated other topical interpretations supported by literature on risk and protective factors for parenting stress, we do not describe them all here. Table S2 supplementary provides a sample of representative documents for each of the remaining topics.

When in interpreting the model, researchers may also gain insights from assessing the prevalence of topics in the corpus. STM calculates each topic's proportional prevalence (i.e., its percentage of all occurrences of all topics) across the corpus of documents. Topics with higher percentages show higher salience across the responses. As shown by Table 3, topic percentages for our corpus ($N = 199$) ranged from 3% (Topic 18) to 11% (Topic 12). The five topics with highest salience were (a) spousal support, (b) routines/schedules, (c) managing expectations, (d) sharing with other parents, (e) faith. Exercising was one of the topics with the lowest prevalence (4%), possibly due to the difficulty of going outdoors during a pandemic lock-down.

STM can also estimate correlations between topics. Positive correlations between topics indicate that these topics are likely to be discussed together in a document (see Figure S2 supplementary for positive correlations across topics). For example, Topic 9 (Faith) was likely to be discussed together with Topic 1 (Friends). One participant's open-ended response offered a possible explanation of this association: they noted that they were "*able to share struggles with Christian friends (cell group) and to pray for each other.*" In other words, respondents' friend groups often overlapped with their faith communities.

STM can also identify and estimate the strength of relationships between observed covariates and topics. Using documents as units, the "estimateEffect" function in the stm package will estimate a regression that shows the proportion of each document about each topic, with covariates for different document metadata (e.g., year document was created). In the present study, we used "estimateEffect" to examine how topical prevalence differed according to

parenting stress levels (continuous variable) and parent gender (binary variable). We found that parents with higher parenting stress wrote more in their open-ended survey question responses about the following topics: (a) Topic 1, “Friends” ($b = .02; p < .01$); (b) Topic 6, “Focus” ($b = .03; p < .01$); and (c) Topic 17, “Activities” ($b = .04; p < .01$; see Figure S3 supplementary for plots of expected topic proportion with parenting stress). Compared to mothers, fathers tended to talk more about exercise ($b = .07; p < .001$; see Figure S4 supplementary). However, given that STM is best used at the exploratory stage of research (e.g., prior to large-scale studies or human coding), these findings should be confirmed with subsequent studies.

Model Validation

Researchers must “validate, validate, validate” their results (Grimmer & Stewart, 2013, p. 5). Demonstrating the validity of the STM-generated topic categories requires establishing their semantic validity (i.e., the extent to which the model identified coherent sets of topics that were internally consistent and distinctive from other topics; Grimmer & Stewart, 2013). Our present study employed a combination of statistical evidence (e.g., semantic coherence and exclusivity metrics) and substantive evidence (i.e., human-checking topic words and reviewing documents highly associated with each topic) to support the validity of our model.

Discussion

Social work researchers typically use close-ended questions to analyze respondents’ attitudes or beliefs because (a) they require less effort from respondents and (b) they yield data that is easier to analyze statistically than the unstructured data generated by open-ended questions. Nevertheless, to understand and faithfully represent the experiences of research participants, social work researchers need to gather data via open-ended questions, too (Thyer, 2012). Encouragingly, new rapid document digitalization tools are speeding up the process of

TEXT MINING OPEN-ENDED SURVEY RESPONSES

generating textual data and new computational methods are available to analyze that data efficiently. Researchers can leverage these technologies when applying text mining methods to textual data collected from open-ended survey questions.

The present study showcases STM's utility for our analysis of parents' open-ended responses to a survey question about how they cope with parenting during a lockdown in Singapore. This study was a follow-up to our earlier study that found that increased parenting stress contributed to parents' harsher disciplinary responses to their children. Given the (understandable) dearth of studies on parents' parenting stress and coping strategies during a pandemic, our study sought to identify parents' major coping strategies for parenting stress during COVID-19.

Our study demonstrated STM's ability to uncover latent topics within a corpus of text documents (i.e., responses to open-ended survey questions). STM enabled us to identify a variety of coping-related topics discussed by our respondents (e.g., spousal support, daily routines and schedules, managing expectations, and faith) and examine co-occurrences (i.e., correlation) between the topics within documents. For instance, we found that when a parent discussed "faith" as a form of coping with parenting stress, they were also likely to talk about "friends" (possibly because their friends were from their faith community). STM also allowed us to input data on respondent-level covariates (e.g., sociodemographic factors; psychosocial functioning) in order to determine whether the prevalence of a topic changed with these covariates. For instance, we found that respondents with higher parenting stress were more likely to talk about topics related to "friends" and keeping a "focused" mindset. We also found that fathers were more likely than mothers to talk about exercising as a coping method for parenting stress during COVID-19. Moreover, as a computer-assisted process, STM identified these topics and

TEXT MINING OPEN-ENDED SURVEY RESPONSES

correlations while avoiding the biases and inconsistencies that researchers may have brought to a human-powered analysis of the same corpus of documents. STM's payoff in terms of time, resources, and coding consistency would be even more evident in an analysis of a larger corpus (i.e., thousands to more than hundreds of thousands) of open-ended responses.

For social work research, STM offers an efficient, valid, and replicable way to analyze textual data. STM bridges qualitative and quantitative research, allowing qualitative data such as open-ended responses to be analyzed with quantitative methods (Thyer, 2012). The researcher's knowledge, skills, and judgment are complemented rather than sidelined by STM's processing power and computer-assisted pattern recognition (Nelson, 2020). Indeed, the researcher's role is central to the model selection stage of the STM workflow. By deciding the appropriate number of topics for the final model based on researcher's substantive knowledge, the researcher can significantly shape the findings derived from the model. In addition to this paper's earlier discussion of tools to help in the researcher's model selection, we also recommend that researchers refer to Roberts et al. (2019), who provide an extensive and in-depth discussion of additional model selection options as well as issues that can arise from using STM (e.g., alternative model specifications; random model selection; false positives).

Notably, STM is not exclusively a standalone process: it can also be used to support other analyses (e.g., clarifying findings; complementing observational data collection methods) or serve as part of a multistage research process or mixed method strategy (Rodriguez & Storer, 2020). For instance, STM can be used both as a coding mechanism to identify themes within a corpus of text documents and as a tool to check the validity of themes already generated by the research team (Rodriguez & Storer, 2020). STM also easily integrates into the computational grounded theory framework (Nelson, 2020), which offers a methodologically rigorous approach

to analyzing textual data. In this three-step framework, the first step, pattern detection, involves an inductive computational exploration of the textual corpus using techniques such as STM. In the second step, pattern refinement, researchers return to the data via a structured qualitative analysis to confirm the validity of the identified topics. The third step, pattern confirmation, assesses whether the results are generalizable to the rest of the corpus or other corpora.

Our study introduces social work researchers to STM and demonstrates its usefulness in analyzing textual data from open-ended survey questions (though it is also suitable for other social work texts such as case notes, policy documents, and interviews). With STM, social work researchers can begin to gather more qualitative data (e.g., open-ended survey questions) in their research without fearing a lengthy and labor-intensive process of analysis. We hope that our study will encourage social work researchers to incorporate more open-ended survey questions in survey methods and embrace new text mining methods such as STM.

References

- Abidin, R. R. (1992). The Determinants of Parenting Behavior. *Journal of Clinical Child Psychology*, 21(4), 407–412. https://doi.org/10.1207/s15374424jccp2104_12
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *ArXiv:1707.02919 [Cs]*. <http://arxiv.org/abs/1707.02919>
- Berry, J. O., & Jones, W. H. (1995). The Parental Stress Scale: Initial Psychometric Evidence. *Journal of Social and Personal Relationships*, 12(3), 463–472. <https://doi.org/10.1177/0265407595123009>
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. <https://doi.org/10.1145/2133806.2133826>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Brown, S. M., Doom, J., Watamura, S., Lechuga-Pena, S., & Koppels, T. (2020). *Stress and Parenting during the Global COVID-19 Pandemic* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/ucezm>
- Burg, M. A., Adorno, G., Lopez, E. D. S., Loerzel, V., Stein, K., Wallace, C., & Sharma, D. K. B. (2015). Current unmet needs of cancer survivors: Analysis of open-ended responses to the American Cancer Society Study of Cancer Survivors II. *Cancer*, 121(4), 623–630. <https://doi.org/10.1002/cncr.28951>
- Cheung, S.-K. (2000). Psychometric properties of the Chinese version of the Parental Stress Scale. *Psychologia: An International Journal of Psychology in the Orient*, 43, 253–261.

Authors (2020). *Associations Between Work-Family Balance, Parenting Stress, and Marital Conflicts During COVID-19 Pandemic in Singapore* [Preprint].

<https://doi.org/10.31219/osf.io/nz9s8>

Authors. Mediating Effects of Parental Stress on Harsh Parenting and Parent-Child Relationship during Coronavirus (COVID-19) Pandemic in Singapore. *Journal of Family Violence*.

<https://doi.org/10.1007/s10896-020-00200-1>

Crosbie, T., & Moore, J. (2004). Work–life Balance and Working from Home. *Social Policy and Society*, 3(3), 223–233. <https://doi.org/10.1017/S1474746404001733>

Deater-Deckard, K. (1998). Parenting Stress and Child Adjustment: Some Old Hypotheses and New Questions. *Clinical Psychology: Science and Practice*, 5(3), 314–332.

<https://doi.org/10.1111/j.1468-2850.1998.tb00152.x>

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, 26(2), 168–189. <https://doi.org/10.1017/pan.2017.44>

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606.

<https://doi.org/10.1016/j.poetic.2013.08.004>

Geer, J. G. (1988). What Do Open-Ended Questions Measure? *Public Opinion Quarterly*, 52(3), 365. <https://doi.org/10.1086/269113>

Geer, J. G. (1991). “Do Open-Ended Questions Measure ‘Salient’ Issues?” *Public Opinion Quarterly*, 55(3), 360–370. <https://doi.org/10.1086/269268>

Gerring, J. (1999). What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences. *Polity*, 31(3), 357–393. JSTOR.

<https://doi.org/10.2307/3235246>

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

<https://doi.org/10.1093/pan/mps028>

Jackson, K. M., & Trochim, W. M. K. (2002). Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses. *Organizational Research Methods*, 5(4), 307–336. <https://doi.org/10.1177/109442802237114>

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, 50(1), 537–567.

<https://doi.org/10.1146/annurev.psych.50.1.537>

Lindstedt, N. C. (2019). Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017. *Social Currents*, 6(4), 307–318.

<https://doi.org/10.1177/2329496519846505>

Mimno, D., & Lee, M. (2014). Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1319–1328. <https://doi.org/10.3115/v1/D14-1138>

Nelson, L. K. (2020). Computational Grounded Theory: A Methodological Framework.

Sociological Methods & Research, 49(1), 3–42.

<https://doi.org/10.1177/0049124117729703>

Patton, D. U., Blandfort, P., Frey, W. R., Schifanella, R., McGregor, K., & Chang, S.-F. U.

(2020). VATAS: An Open-Source Web Platform for Visual and Textual Analysis of

- Social Media. *Journal of the Society for Social Work and Research*, 11(1), 133–155.
<https://doi.org/10.1086/707667>
- Perron, B. E., Victor, B. G., Bushman, G., Moore, A., Ryan, J. P., Lu, A. J., & Piellusch, E. K. (2019). Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning. *Child Abuse & Neglect*, 98, 104180. <https://doi.org/10.1016/j.chiabu.2019.104180>
- Pietsch, A.-S., & Lessmann, S. (2018). Topic modeling for analyzing open-ended survey responses. *Journal of Business Analytics*, 1(2), 93–116.
<https://doi.org/10.1080/2573234X.2019.1590131>
- Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111(515), 988–1003. <https://doi.org/10.1080/01621459.2016.1141684>
- 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>
- Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). *The Structural Topic Model and Applied Social Science*. Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation, Cambridge, MA.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4), 1064–1082.
<https://doi.org/10.1111/ajps.12103>

- Rodriguez, M. Y., & Storer, H. (2020). A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data. *Journal of Technology in Human Services*, 38(1), 54–86.
<https://doi.org/10.1080/15228835.2019.1616350>
- Rubin, A., & Babbie, R. (2013). *Research Methods for Social Work* (8th ed.). Brooks/Cole.
- Rubin, Allen. (2013). Survey Research. In Allen Rubin, *Encyclopedia of Social Work*. NASW Press and Oxford University Press.
<https://doi.org/10.1093/acrefore/9780199975839.013.387>
- Thyer, B. A. (2012). The scientific value of qualitative research for social work. *Qualitative Social Work*, 11(2), 115–125. <https://doi.org/10.1177/1473325011433928>
- Victor, B.G., Perron, B.E., Sokol, R.L., Fedina, L., & Ryan, J.P. (in press). Automated identification of domestic violence in written child welfare records: Leveraging text mining and machine learning to enhance social work research and evaluation. *Journal of the Society for Social Work and Research*.

TEXT MINING OPEN-ENDED SURVEY RESPONSES

Table 1.

Sample Characteristics and Descriptive Statistics of Study Variables (n = 199)

	<i>N</i>	<i>%</i>	<i>M</i>	<i>SD</i>	<i>Range</i>
Parent (natural/adoptive/stepparent)					
Mothers	136	68%			
Fathers	63	32%			
Parent Race					
Chinese	167	84%			
Malay	18	9%			
Indian	6	3%			
More than one race/Others	8	4%			
Parent Age (in years)					
26-30	11	6%			
31-35	55	28%			
36-40	65	33%			
41-45	39	20%			
46-50	23	12%			
51-55	6	3%			
Parent Education					
Non-University	29	15%			
University	170	85%			
Monthly Household Income (Singapore \$)					
< \$1900	6	3%			
\$1901-\$3500	9	5%			
\$3501-\$5000	27	14%			
\$5001 - \$8000	42	21%			
> \$8000	115	58%			
Youngest Child Age (in years)					
0-1	48	24%			
3-Feb	40	20%			
5-Apr	39	20%			
7-Jun	22	11%			
9-Aug	24	12%			
12-Oct	26	13%			
# of children at home (age 12 or younger)					
1	91	46%			
2	78	39%			
3 or more	30	15%			
Parental Stress			2.39	0.46	2.5

TEXT MINING OPEN-ENDED SURVEY RESPONSES

Table 2.

Topics resulting from structural topic modeling (k = 18)

Topic ID	Topic Label	Highly Associated Terms (PROB)	Frex	Topic Proportion
12	Spousal support	spouse, support, involvement	spouse, support, involvement	11%
8	Routine/schedule	schedule, routine, stay	schedule, routine, stay	8%
4	Managing expectations	manage, expect, adjust	manage, expect, adjust	8%
2	Sharing With Other Parents	parent, share, response	parent,share, listen	7%
9	Faith	faith, law, trust	faith, law, cellgroup	6%
1	Friends	friend, talk, online	friend, talk, online	6%
3	Engage children	plan, engage, rule	engage, child, screen	6%
10	Shopping	shop, parent, positive	shop, monitor, groceries	6%
14	Timetable Duty	timetable, duty, care	balance, duty, calm	5%
6	Focus	focus, live, close	focus, moment, live	5%
7	Knowing I am not alone	parent, caregiver, know	caregiver, know, worker	5%
11	Domestic helper	helper, domestic, communicat	helper, discipline, domestic	4%
15	Understanding	understand, discuss, daily	understand, discuss, daily	4%
17	Activities	activit, child, create	activit, child, create	4%
5	Exercise	exercise, enjoy, game	exercise, indoor, pet	4%
13	Laugh	love, laugh, smile	love, laugh, boss	4%
16	Patience	patient, marriage, parent	patient, marriage, parent	4%
18	Bonding	space, bond, idea	bond, space, idea	3%

TEXT MINING OPEN-ENDED SURVEY RESPONSES

Table 3.

Top Two Topics with Highest Topic Proportion and Exemplar Documents

Topic 12 "Spousal Support"

"A lot busier with work during covid 19 as new jobs arise. Thankful for the support helper and spouse give."

"Open communication and support from spouse"

"Supportive and involved spouse"

"My husband's support"

"Support from my spouse and understanding from co-workers"

Topic 8 "Routine/Schedule"

"Pre-planned routine and schedule for the children, personal timeouts/time for myself, exercise, Netflix subscription, Gov.SG Whatsapp Chat/messages on Covid-19, alone time with my spouse."

"Working from home allows me to spend more time with my children which I am grateful for. Having a routine or a schedule also helps."

"Chat with the kids often and stick to routine."

"Setting up a routine for myself and kids"

"Setting routines and planning ahead for activities for my children"

Figure 1. A Visual Representation of Structural Topic Modeling. Adapted from "Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education" by Chen, Zou, Cheng, Xie, 2020, *Computers & Education*, 151, p.5. Copyright 2020 by Elsevier

TEXT MINING OPEN-ENDED SURVEY RESPONSES

Table 3.

Top Two Topics with Highest Topic Proportion and Exemplar Documents

Topic 12 "Spousal Support"

"A lot busier with work during covid 19 as new jobs arise. Thankful for the support helper and spouse give."

"Open communication and support from spouse"

"Supportive and involved spouse"

"My husband's support"

"Support from my spouse and understanding from co-workers"

Topic 8 "Routine/Schedule"

"Pre-planned routine and schedule for the children, personal timeouts/time for myself, exercise, Netflix subscription, Gov.SG Whatsapp Chat/messages on Covid-19, alone time with my spouse."

"Working from home allows me to spend more time with my children which I am grateful for. Having a routine or a schedule also helps."

"Chat with the kids often and stick to routine."

"Setting up a routine for myself and kids"

"Setting routines and planning ahead for activities for my children"

Figure 2. Exclusivity-Semantic Coherence Plot. This figure illustrates the level of exclusivity and semantic coherence scores. Optimal models are those with high scores along both the semantic coherence and exclusivity dimensions

Figure 3. This figure illustrates the relative goodness of fit for each number of topics. Optimal results would demonstrate relatively high semantic coherence, low residuals, a maximized lower bound, and a high held-out likelihood