The use of Topic Modeling to Analyze Open-Ended Survey Items

W. Holmes Finch Maria E. Hernández Finch Constance E. McIntosh Claire Braun

Ball State University

Open ended survey items

- Researchers making use of surveys for data collection purposes often include both closed format items (e.g., likert type) as well as open ended items for which respondents are asked to generate responses.

- For example: “Please explain the lines of communication that exist at your school among support personnel, for students engaging in non-suicidal self-injury.”

Open ended survey items

- Although open ended survey items can provide useful information, they are often problematic to code.
- It can be difficult to categorize responses into meaningful groupings.
- In addition, relating responses on open ended items to responses on other items can prove challenging.
- Text mining methods (e.g., Topic Modeling) may allow researchers to investigate relationships among such open ended items and closed ended (e.g., Likert-type) items in ways that heretofore have not been possible.

Topic Modeling

- Topic modeling (TM) is a statistical methodology designed to identify underlying themes in text.
- It is very similar to cluster analysis, such that a (hopefully) small set of topics is identified based upon co-occurrence of word usage in a set of texts.
- Each topic is characterized by a mixture of words that appear frequently together.
- In other words, topics are simply collections of frequently co-occurring words.

Topic Modeling

- When conducting TM, researchers analyze data from a set of documents known as a corpus.
- Each document is assumed to contain multiple topics, which themselves contain multiple words.
- A document will be classified as belonging to the topic which is most represented in it based on the document's word mix.
- TM yields information about two parameters: (1) the probability of specific words appearing in the topic (β), and (2) the probability of specific topics appearing in a document (γ).

Latent Dirichlet Allocation

- There exist a number of statistical tools for identifying topics from among documents in a corpus.
- One of the more popular of these is Latent Dirichlet Allocation (LDA).
- LDA provides estimates of both β and γ.
- Under LDA it is assumed that these parameters are distributed as follows:
  - $\beta = Dirichlet(\delta)$
  - $\gamma = Dirichlet(\alpha)$
  - Where δ and α are vectors of probabilities associated with words in topics, and topics in documents, respectively.

Latent Dirichlet Allocation

- The TM parameters can be estimated using maximum likelihood by maximizing the following function:

$$l(\alpha, \beta) = ln\big(p(w|\alpha, \beta)\big)$$

Where

w =Observed mixture of words within documents

α =Dirichlet parameter for topics in corpus

β =Probability of a given word occurring in a given topic

Determining the number of topics to retain

- Perhaps the most important decision a researcher must make when using TM is the number of topics to retain.

- Much as with cluster analysis, or exploratory factor analysis, this decision should be made using both statistical tools and an analysis of the content of the topics (i.e., do the topics make sense).

- There exist a number of statistical tools designed to help with this process.

Determining the number of topics to retain

- One of the most well proven methods for determining the number of topics to retain is based upon a density estimator described in Cao, et al., (2009).
- This approach uses an iterative algorithm in which the distances among pairs of topics are calculated.
- Next, the density for each topic is calculated, where density is based upon the number of clusters within a prespecified distance of a topic.
- The optimal number of topics is the one for which the average density across topics is minimized; i.e., the topics are most independent/separated from one another.

Goals of this study

- The primary goal of this study is to demonstrate the use of TM to identify topics in a corpus of open ended item responses.

- Once these topics are identified and individual respondents assigned to them, relationships among the topics and responses to other items on the scale were investigated.

Methodology

- Participants – 620 individuals were sampled from across the United States, working in school settings as either psychologists, nurses, counselors, or social workers.

- Of these 620 individuals, 184 provided responses to the target open ended item, which will be discussed next.

Profession	Frequency (Percent)
School Nurse	45 (24.5%)
School Counselor	41 (22.2%)
School Psychologist	48 (26.1%)
School Social Worker	50 (27.2%)

Methodology

- Respondents were given a survey that included a number of likert-type items, as well as several open ended questions.

- The target open ended question for this study was:

  "Please provide information regarding your school's policies and procedures regarding the identification of and intervention with students engaging in non-suicidal self-injury."

- Thus, the corpus consisted of 184 written responses to this item.

Methodology

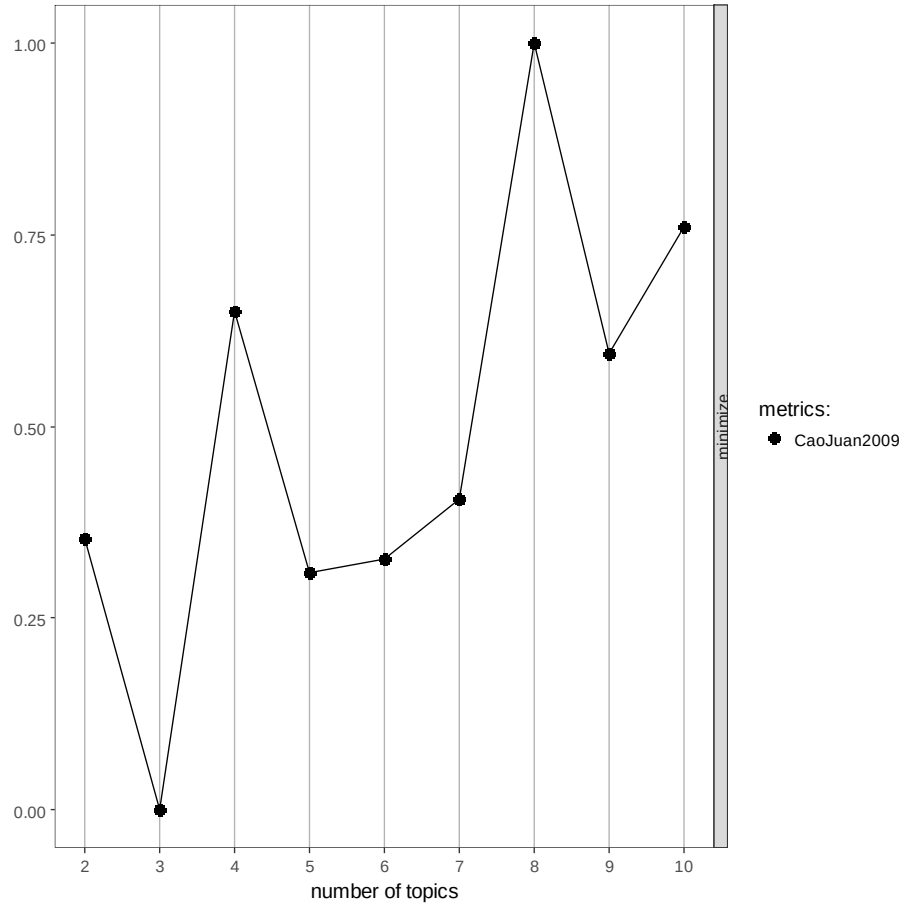
- The data were preprocessed so as to remove nuisance words (e.g., the, a, and), capitalization, suffixes, prefixes, digits, and punctuation.
- TM was then conducted on the processed corpus using LDA.
- The optimal number of topics to be retained was determined based on the density based statistic of Cao, et al., (2009), as well as a content review of the words within the topics to ensure their conceptual coherence.
- Each open ended item response was then classified as belonging to the topic for which it had the highest probability, based upon its word content.

Methodology

- Probabilities of each word being generated by the individual topics (β) were calculated.

- Relationships between the topics and responses to the likert-type survey items were then investigated using cross-tabulations, the Chi-square test of association, measures of association for categorical variables, and the Mantel-Haenszel test.

Results



- The optimal number of topics occurs where the Cao, et al, (2009) statistic is minimized.
- For this dataset, the minimum occurred for 3 topics.

Results: Most Commonly Occurring Words by Topic, and Probability of each Word Being Generated by the Topic (β)

Topic 1 (Role of school nurse)	Topic 2 (Lack of school policy)	Topic3 (Role of Mental Health Professionals)
Student (β=0.012)	Policy (β=0.092)	Counselor (β=0.049)
Contact (β=0.015)	Need (β=0.040)	Psychologist (β=0.028)
Nurse (β=0.066)	Not (β=0.050)	Social (β=0.045)
Parent (β=0.023)	Have (β=0.050)	Injurious (β=0.030)
Refer (β=0.010)	Injurious (β=0.031)	Behavior (β=0.030)
Teacher (β=0.028)	Suicide (β=0.031)	Trained (β=0.028)

- The three topics, along with the 6 most common words in each, appear in the Table.
- Topic 1 – Role of the school nurse vis-à-vis teachers and parents.
- Topic 2 – Lack of school policy with respect to self-injurious behavior.
- Topic 3 – Role of trained school mental health professionals in dealing with self-injurious behavior.

Results: Comparison of word frequency between topic pairs

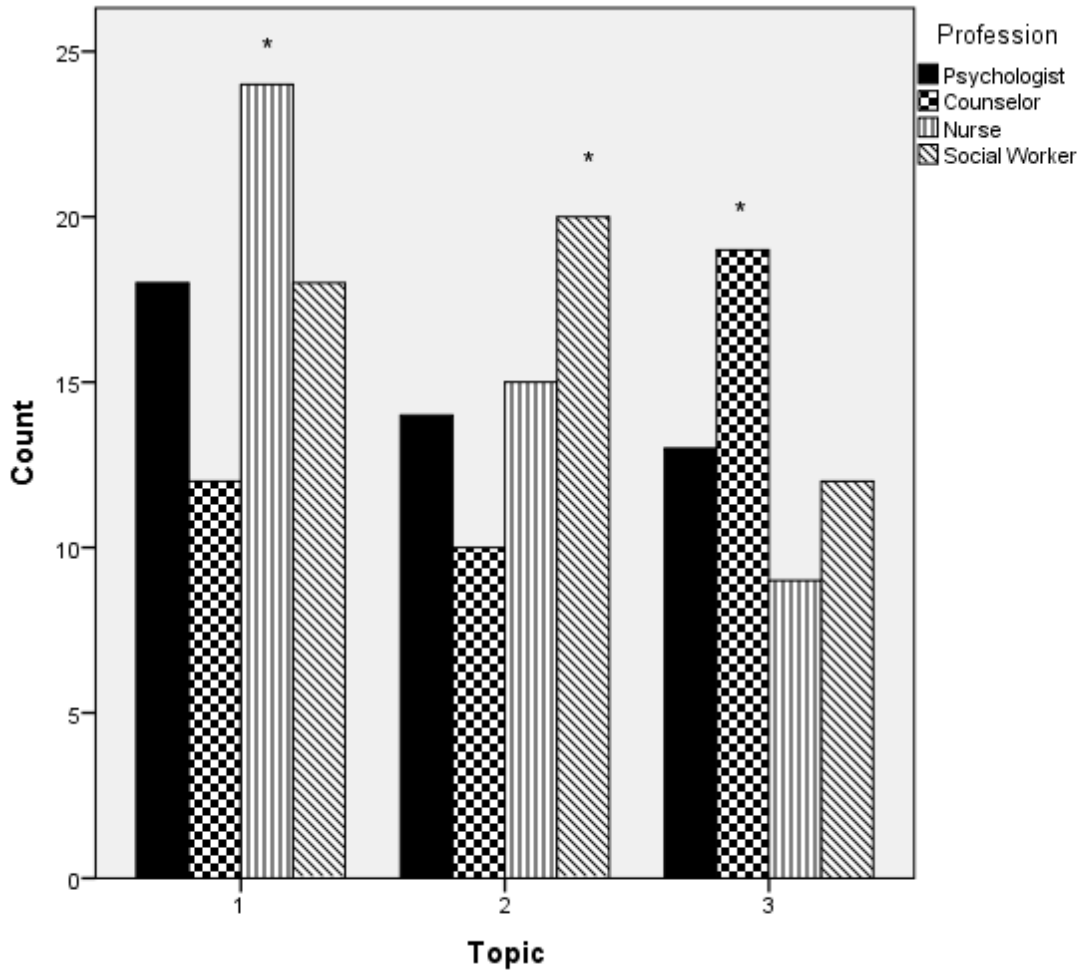
Topic 1 versus Topic 3			
Term	Topic 1	Topic 3	Log Ratio
Communication	0.0001	0.061	8.92
Social	0.0001	0.045	8.48
Psychologist	0.0001	0.028	7.80
Meet	0.0001	0.034	8.08
Team	0.0001	0.033	8.03
Topic 1 versus Topic 2			
Term	Topic 1	Topic 2	Log Ratio
Nurse	0.066	0.0001	-9.04
Have	0.0001	0.050	8.63
Not	0.0001	0.050	8.63
Parent	0.023	0.0001	-7.52
Need	0.0001	0.040	8.31
Topic 2 versus Topic 3			
Term	Topic 2	Topic 3	Log Ratio
Counselor	0.0.049	0.0001	-8.61
Have	0.0001	0.050	8.63
Not	0.0001	0.050	8.63
Need	0.0001	0.040	8.31
Support	0.0.026	0.0001	-7.67

- Respondents in topic 3 were more likely than those in topic 1 to mention mental health professionals, communication, and teamwork.
- Respondents in topic 3 were more likely to mention counselors and support than those in topic 2, and less likely to mention not having a policy.
- Respondents in Topic 1 were more likely than those in topic 2 to mention nurses and parents, and less likely to mention not having a policy.

Results: Relationship between respondent profession and topic

- There was a statistically significant relationship between respondent profession and topic (p=0.001, Cramer's V=0.243).

- Nurses were more likely to be represented in topic 1 than expected by chance.
- Social workers were more likely to be represented in in topic 2 than expected.
- Counselors were more likely to be represented in topic 3 than expected.

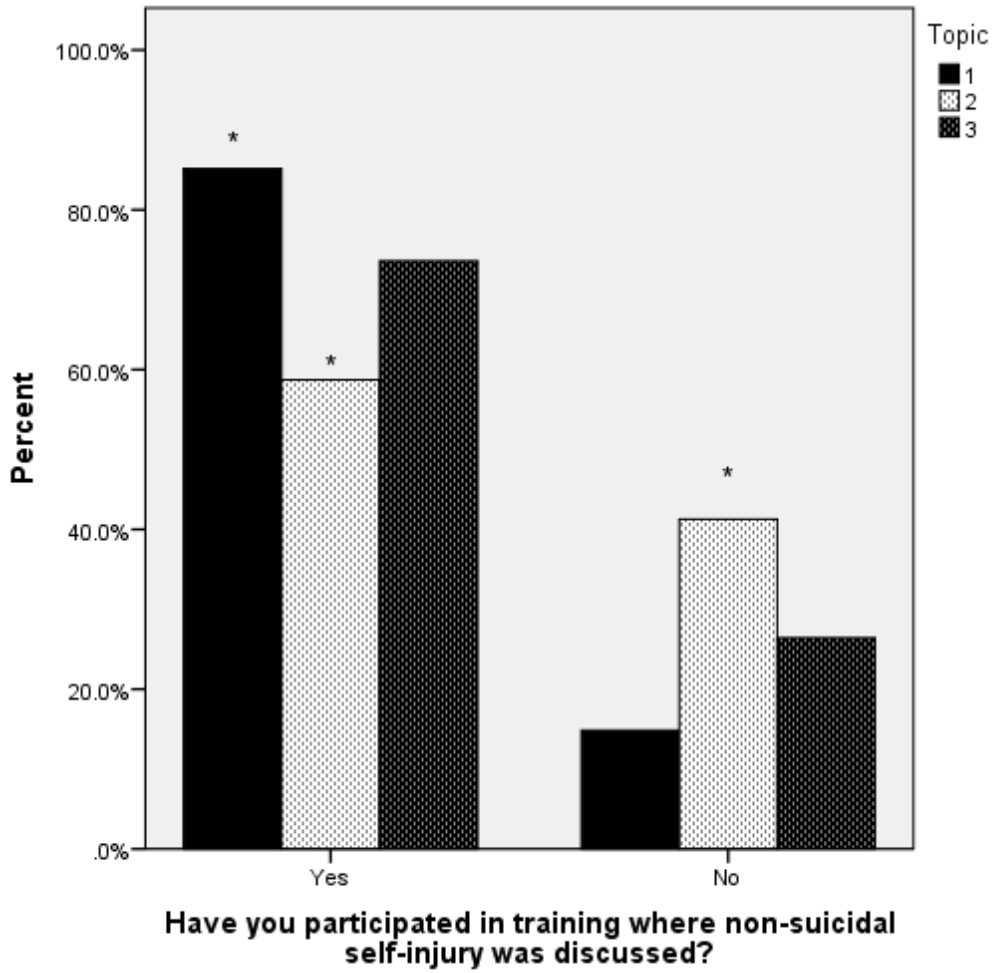


*=Absolute value of the adjusted standardized residual greater than or equal to 2
Topic 1 = Roll of school nurse
Topic 2 = No School Policy
Topic 3 = Roll of Mental Health Professionals

Results: Topic by Participating in Training where Non-Suicidal Self-Injury was Discussed

- There was a statistically significant relationship between training participation and topic (p=0.004, Cramer's V=0.252).

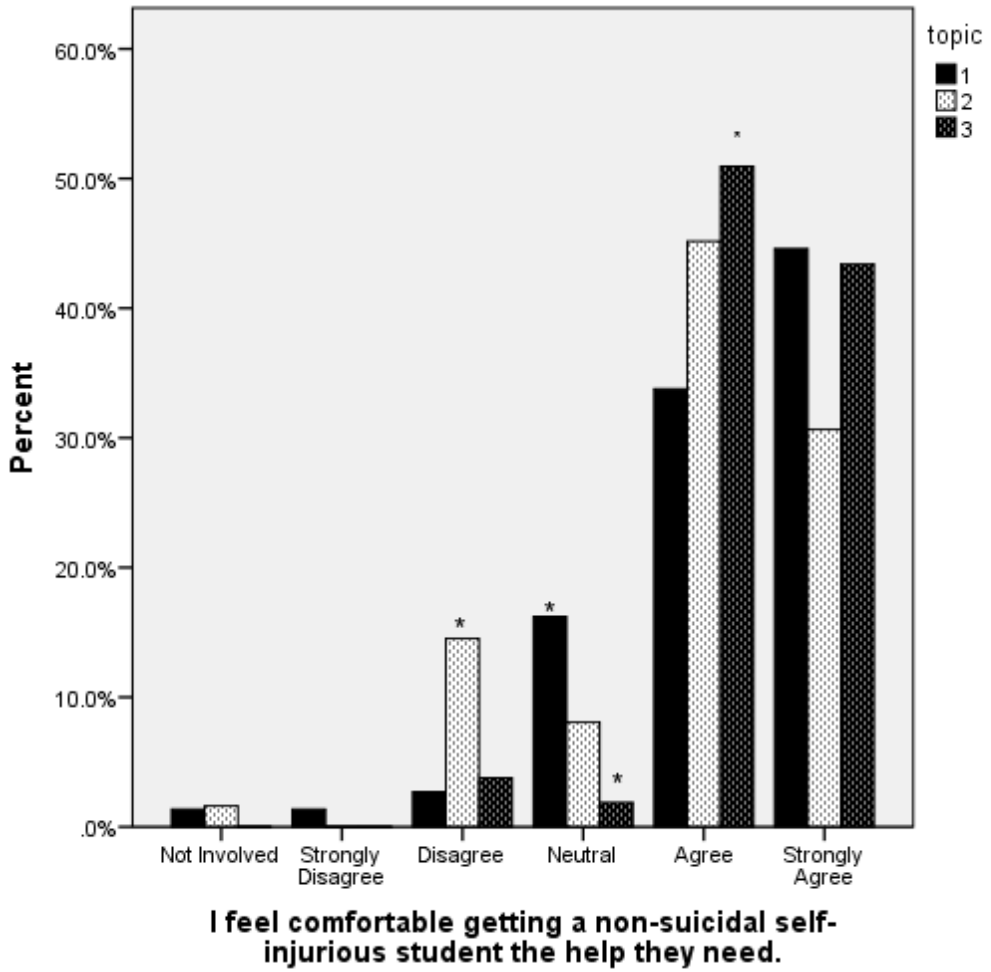
- Respondents in topic 1 were more likely than expected to have participated in training, whereas those in topic 2 were less likely to have participated.



*=Absolute value of the adjusted standardized residual greater than or equal to 2
Topic 1 = Roll of school nurse
Topic 2 = No School Policy
Topic 3 = Roll of Mental Health Professionals

Results: Feel Comfortable Getting Self-Injurious students Help They Need

- There was a statistically significant relationship between comfort getting students help and topic (p=0.012, Cramer's V=0.238).
- Respondents in topic 2 were more likely than expected to disagree that they were comfortable.
- Those in topic 1 were more likely than expected to be neutral.
- Respondents in topic 3 were less likely than expected to be neutral and more likely than expected to agree that they would be comfortable getting students help.



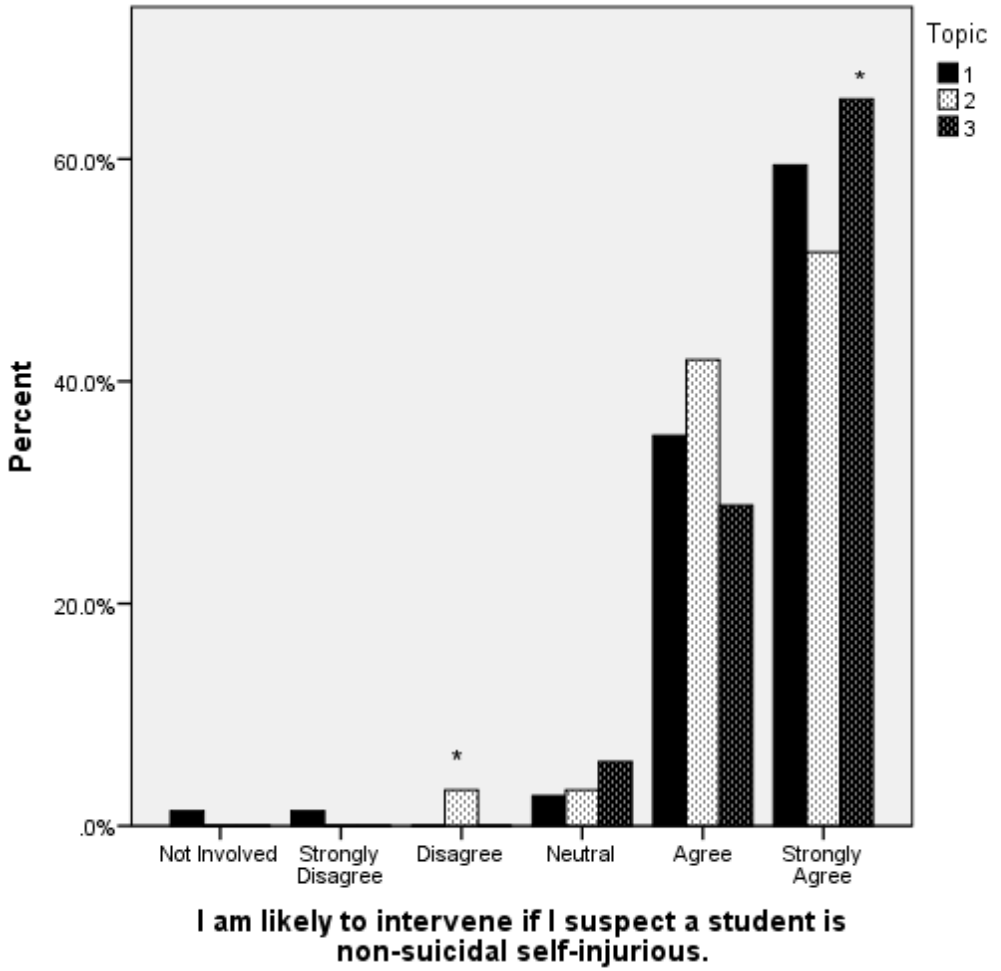
*=Absolute value of the adjusted standardized residual greater than or equal to 2
Topic 1 = Roll of school nurse
Topic 2 = No School Policy
Topic 3 = Roll of Mental Health Professionals

Results: Likelihood of intervening with self-injurious students

- There was a statistically significant relationship between comfort getting students help and topic (p=0.001, Cramer's V=0.288).

- Respondents in topic 2 were more likely than expected to disagree that they are likely to intervene.

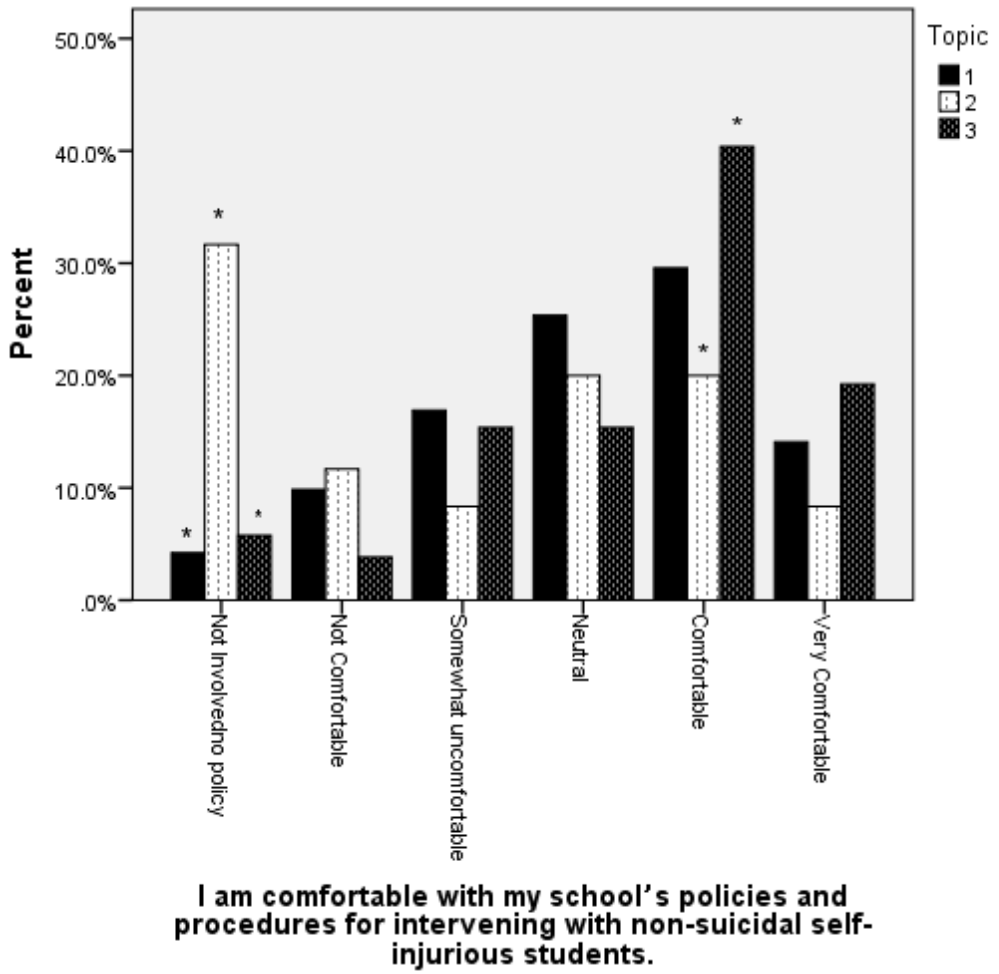
- Respondents in topic 3 were more likely than expected to strongly agree that they are likely to intervene.



*=Absolute value of the adjusted standardized residual greater than or equal to 2
Topic 1 = Roll of school nurse
Topic 2 = No School Policy
Topic 3 = Roll of Mental Health Professionals

Results: Comfortable with School's Policies for Intervention with Self-Injurious Students

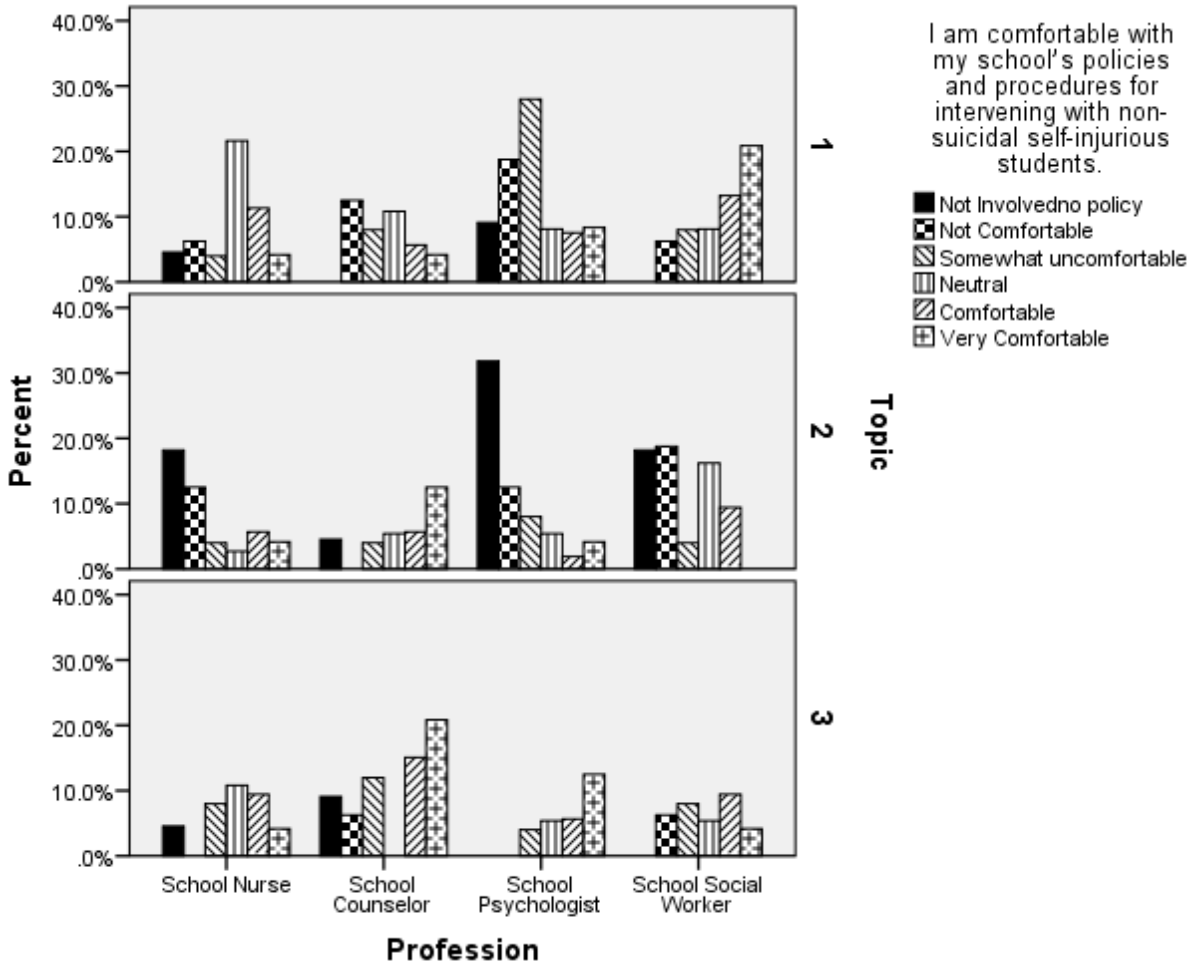
- There was a statistically significant relationship between comfort getting students help and topic (p<0.001, Cramer's V=0.301).
- Respondents in topic 2 were more likely than expected to not be involved in policy, and less likely to be comfortable.
- Respondents in topics 1 and 3 were less likely to not be involved than would be expected.
- Respondents in topic 3 were more likely than expected to be comfortable with school policies on intervening with self-injurious students.
- Respondents in topic 1 were less likely to be comfortable with school's policy than would be expected.



*=Absolute value of the adjusted standardized residual greater than or equal to 2
Topic 1 = Roll of school nurse
Topic 2 = No School Policy
Topic 3 = Roll of Mental Health Professionals

Results: Interaction Among Profession, Topic, and Comfort with School Policies

- There was a statistically significant relationship among topic, profession, and comfort with school policies (p=0.006).
- Nurses and social workers who mentioned nurses, parents, and teachers were generally more comfortable with policies than were counselors and psychologists in this topic.
- Members of all professions who discussed the absence of a school policy were generally less comfortable, except for counselors.
- Counselors and psychologists who mentioned the role of mental health professionals were generally comfortable with school policies, whereas nurses and social workers in this topic were somewhat less comfortable.



Topic 1 = Roll of school nurse
Topic 2 = No School Policy
Topic 3 = Roll of Mental Health Professionals

Conclusions

- Text mining offers survey researchers a very useful tool for identifying themes (topics) within open ended items.
- These themes can then be related back to other items on the survey, thereby providing insights into respondents' thoughts and/or behaviors.
- Using text mining, the themes identified in open ended responses are not subject to the vagaries of individual coders, but rather can be arrived at in a more (though not totally) objective manner.

Conclusions

- In the current study, 3 topics were identified for the item “Please provide information regarding your school’s policies and procedures regarding the identification of and intervention with students engaging in non-suicidal self-injury.”

- These topics reflected respondent concerns about no policy existing, a discussion of the role played by mental health professionals, and contact between school nurses, parents, and teachers.

Conclusions

- Respondents who discussed the absence of a policy were more likely to be employed as social workers.

- In addition, individuals who discussed the absence of policy were less likely to have received training, and were less involved with school policy for intervening with self-injurious students.

- Those who discussed the role of nurses with parents and teachers were themselves more likely to be nurses (though psychologists were also likely to be represented in this topic), to have received training, and to be relatively comfortable with school policies, getting students help, and intervening with self-injurious students.

- Individuals who discussed the role of mental health professionals were more likely to be school counselors, to have had training, and were generally the most comfortable with school policies, with intervening, and with getting self-injurious students help.

Conclusions

- Interestingly, mention of parents and teachers was more likely to co-occur with mention of nurses, rather than with mental health professionals.
- Conversely, mention of teamwork was more commonly associated with mental health professionals rather than nurses.
- In addition, the topics appear to show a bifurcation between the perceived role of nurses and of other mental health professionals in dealing with self-injurious behavior.
- Finally, counselors were the most likely to be comfortable with school policies, unless they also mentioned the role of nurses, in which case their comfort level was lower than nurses or social workers.