403 – Analytics, Attrition, and Words

Code-frame Development for Verbatim Responses Using a 3-stage Approach

Sponsor: Section on Statistical Learning and Data Mining
Keywords: text analysis, Data mining, Survey Research

Bernard L. Dugoni

NORC at the University of Chicago

Kevin Brown

NORC at the University of Chicago

In a study of the priorities of educators in the midwest, a three step process was used to assess open-ended responses, combining the use of text-analysis software with researcher review techniques for a tailored solution that is computer assisted but retains the human element. The first stage used the software "QDAMiner" developed by Provalis Research to conduct content analysis based on word frequency. In the second stage, related words were combined into phrases by using "keyword-in-context" analysis, which assesses whether the same words or phrases share an single concept or mean different things to different respondents. At this stage, responses were combined by the program into a matrix of co-occurrence, which allows the analysis of cases where two or more of the keywords are used in a given response. The software tools used at the initial stages do not capture nuance or ambiuity, so it is important that post-processing be done by the researcher. The third stage involved an analysis of internal consistency and review by multiple raters to achieve useful qualitative categories which were combined for further analysis with quantitative variables also collected in the project.

Code-frame Development for Verbatim Responses Using a 3-stage Approach

Bernard L. Dugoni, Ph.D. and Kevin Brown, Ph.D.
NORC at the University of Chicago

Abstract. In conjunction with a report prepared for a project related to the Regional Education Laboratories, Midwest, NORC developed a procedure for analyzing open-ended responses to questionnaire items using a combination of computer assisted techniques and direct examination of the text responses. In this procedure, we used software developed by Provalis Research (WordStat and SimStat) to conduct the initial content analysis based on word frequency. This was further refined by combining related words and phrases and conducting what is referred to as keyword-in-context analysis which assesses whether the same words or phrases share an over-arching concept or are used by different respondents in different unrelated concepts. This software further allowed us to transfer the resulting keyword and word frequency statistics to SPSS for further assessment to clarify, expand, and statistically evaluate our observations.

Introduction

In this study, NORC employed a 3-stage process using a combination of computer-assisted techniques and direct examination to develop a frame for coding open-ended responses to survey responses of educators who had been asked to identify high priority issues for their school district. *Overal*l, there were 279 valid (non-blank) responses to the open-ended item of interest out of the 603 educators in the Midwest who completed an online survey fielded in the spring of 2013. We removed the 45 respondents who only gave simple declarative replies ("yes," "no," and variations thereof), leaving a sample of 234 usable responses. Our primary goal in this process was to extract from these 234 responses a concise set of shared high priority educational issues.

Procedure

Stage 1. We began with a content analysis based on word frequency that employed the text mining software "WordStat" and "SimStat" developed by Provalis Research. In order to capture issues that were shared by several respondents, the software was set to disregard words that occurred in fewer than five responses. It is important to note, however, that the program employs a built-in thesaurus so that synonyms are combined into a single keyword. For example, the words "funding," "revenue," "money," and "funds" were all counted and combined into the single, most commonly used keyword "funding." The results of this frequency analysis are shown in Table 1.

Table 1. Initial Keyword Analysis

Keyword	Frequency
BASED	9
COMMON	13
CORE	14
CURRICULUM	18
EDUCATION	13
FUNDING	30
GAP	10
INCREASING	12
LEARNING	17
QUALITY	9
SCHOOL	20
STAFF	10
STANDARDS	12
STUDENT	45
TEACHERS	19
TECHNOLOGY	15
Total	266

Stage 2. The next stage of the process refined the results shown in Table 1 by combining related words into phrases and conducting what is commonly referred to as keyword-in-context analysis, which assesses whether the same words or phrases share an over-arching concept or mean different things to different respondents. At this stage, responses are combined by the WordStat analysis program into a matrix of co-occurrence, which presents cases where two or more of the keywords are used in a given response. Table 2 below is an excerpt of this matrix that illustrates part of the process. The diagonal elements of this matrix show the total number of times an individual keyword exists in the verbatim responses (also shown in Table 1). The off-diagonal elements show how often two words are used together. For example, the keyword "based" occurs a total of 9 times, twice co-occurring with the keyword "curriculum." Similarly, the word "core" occurs with the word "common" 12 times and with the word "standards" 7 times. Note that the second most commonly used word, "funding," was rarely used in conjunction with other keywords, so this category could not be further refined until the next stage.

The analyst reviewed this matrix to highlight cases where keywords were used in combination and conducted a preliminary direct examination of responses to ensure that the co-occurrence of words was being used in a meaningful way. Of course this "data reduction" is only a step in developing a scheme for coding the entire set of responses, many of which will be captured even without including any of the keywords in Table 1 as long as the response conveys the same meaning.

Table 2. Matrix for Keyword-in-Context Analysis

	BASED	COMMON	CORE	CURRICULUM	...	FUNDING	...	STANDARDS	...
BASED	9								
COMMON	0	13							
CORE	0	12	14						
CURRICULUM	2	1	0	18					
...	...	...	...	...	...				
FUNDING	0	0	0	0	...	30			
...	...	...	...	...	...	...	...		
STANDARDS	2	6	7	1	...	0	...	12	
...	...	...	...	...	...	...	...	...	...

Stage 3. While helpful for an initial pass through text responses, the software tools used at the initial stages do not always capture nuance or ambiguous wording very well, so it is important that additional post-processing be done by the researcher.

The final stage of the process was designed as a quality control check on results from the previous stages and provided a further refinement of the codeframe by matching the results from the prior stages with more specific issues of known interest to educators in the region and across the country. To accomplish this, both the analyst and the project director applied the basic categories obtained from the computer-assisted analyses to the set of verbatim responses identified in Stage 1 in order to (1) assess whether the codeframe was capturing issues not included in the 34 closed-ended items in Q1 and (2) determine whether the categories from Stage 2 could be elaborated based upon our understanding of priority issues facing educators more generally. Table 3 shows the final set of categories from this joint review of the responses.

Table 3. Final Categories for Coding Q2

Conferenced Codeframe
The Common Core
Curriculum Standards and Quality
Learning Gaps/Diversity Issues
Standards for Teacher Training/Assessment
Funding: (1) Fair allocation of existing resources (2) Funding of legislative mandates
Technology: (1) Availability of equipment/resources (2) Training/support for teachers (3) Development of programs/curricula which use technology

As a final step in stage 3, we applied a series of statistical assessments to evaluate the categories derived from the combination of the techniques described above as well as scale development evaluated with Cronbach's Alpha coefficient.

Using the items identified in the earlier stages, simple scales were developed coding the items as 'present' or 'absent' for a given respondent and summing the items coded as present. Using SPSS Scale, Cronbach's Alpha was computed and the value recomputed as items were eliminated in order to allow us to compare the combinations of items identified by the programs to the combinations suggested by the raters.

Scales including all variables identified by the computer-assisted analyses in the earlier stages resulted in Cronbach's Alpha coefficients between .61 and .75, whereas the scales created in line with the expert reviewed/refined groupings further evaluated by the SPSS Scale procedure yielded alpha coefficients between .83 and .93.

Future work will explore further evaluation of data derived from this approach using cluster analysis to derive loadings to assess the classification structure derived from the stages of this approach.

Conclusions

New software approaches provide useful tools for assisting and streamlining text analysis, however, they do not eliminate the need for human review of results. Statistical procedures provide additional clarity where results may be ambiguous, and multiple tests can be helpful in providing stronger basis for clarification.