# Investigating Teachers' Understanding Through Topic Modeling: A Promising Approach to Studying Teachers' Knowledge


## Abstract

Examining teachers' knowledge on a large scale involves addressing substantial measurement and logistical issues; thus, existing teacher knowledge assessments have mainly consisted of selected-response items because of their ease of scoring. Although open-ended responses could capture a more complex understanding of and provide further insights into teachers' thinking, scoring these responses is expensive and time consuming, which limits their use in large-scale studies. In this study, we investigated whether a novel statistical approach, topic modeling, could be used to score teachers' open-ended responses and if so, whether these scores would capture nuances of teachers' understanding. To test this hypothesis, we used topic modeling to analyze teachers' responses to a proportional reasoning task and examined the associations of the topics identified through this method with categories identified by a separate qualitative analysis of the same data as well as teachers' performance on a measure of ratios and proportional relationships. Our findings suggest that topic modeling seemed to capture nuances of teachers' responses and that such nuances differentiated teachers' performance on the same concept. We discuss the implications of this study for education research.

# Investigating Teachers' Understanding Through Topic Modeling: A Promising Approach to Studying Teachers' Knowledge

## Introduction

Teachers play a significant role in students' academic outcomes (e.g., Aaronson et al., 2007; Gordon et al., 2006; Hill et al., 2005; Kersting et al., 2012; Nye et al., 2004; Ottmar et al., 2015; Rockoff et al., 2011). An important contributor to teachers' impact on students' academic outcomes is teachers' knowledge of the subject matter and how it is used in their teaching (Ball et al., 2008; Shulman, 1986). Prior studies have provided evidence that teachers' content knowledge for teaching is related to the learning environment they create for their students (e.g., Blazar, 2015; Borko et al., 1992; Copur-Gencturk, 2015; Hill et al., 2008; Kersting et al., 2012) and student learning (e.g., Baumert et al., 2010; Hill et al., 2005; Kersting et al., 2012)[1]. Thus, not surprisingly, a substantial number of teacher education and professional development programs have been devoted to enhancing teachers' content knowledge for teaching (e.g., Copur-Gencturk, 2015; Copur-Gencturk, Plowman et al., 2019; Copur-Gencturk & Thacker, 2021).

Yet assessing teacher knowledge on a large scale—whether for the purpose of linking it to teacher practices or student learning or to detect changes in teacher knowledge—is not an easy task. Qualitative analysis can provide rich and useful insights into teachers' knowledge; however, coding data from hundreds of teachers requires substantial time and resources. Measuring teacher knowledge through selected-response items can overcome some of these cost- and time-related issues, but such data sets provide relatively little direct information on teachers' thinking and reasoning. More recently, researchers have begun using open-ended items to

[1] Although empirical evidence supporting the role of teachers' knowledge in student learning is weak, we argue that the weakness of this association is related to several methodological and measurement issues (Copur-Gencturk, Jacobson et al., 2021).

capture teachers' knowledge and then scoring teachers' responses (Kersting et al., 2012). Indeed, when teachers' knowledge has been accessed through open-ended items, the association between their knowledge and students' learning has seemed stronger (Baumert et al., 2010; Kersting et al., 2012). However, the data gathered through this approach have typically been used to quantify teachers' performance rather than to investigate potential differences in teachers' reasoning or thinking. Thus, useful information contained in teachers' responses to constructed-response items that reflects their understanding may not be reflected in their scores alone.

We acknowledge that rigorous, large-scale studies have been conducted to capture qualitative differences in teachers' responses and understanding (e.g., Copur-Gencturk, 2021a; Copur-Gencturk, 2021b; Copur-Gencturk & Doleck, 2021; Copur-Gencturk & Olmez, 2021; Goulding et al., 2002; Tatto, 2013), yet limitations of time and resources have minimized the frequent use of these approaches. In this study, we aimed to explore whether new statistical methods could be used to analyze teachers' responses to constructed-response items as a means of detecting the nuances in their understanding. Although we do not claim that such a method could be used solely or in place of qualitative analyses, if such an approach were successful in capturing the nuances in teachers' responses, it could be used along with qualitative analyses to provide researchers with a tool to use when conducting studies with larger samples of participants. Such an approach would be useful only if it could indeed capture the nuances in teachers' responses. Thus, in this study, we tested the possibility of using statistical topic modeling by analyzing teachers' responses to a constructed-response item that was designed to capture teachers' content knowledge, specifically their understanding of proportional reasoning. We chose this particular item and the teachers' responses because this problem has been used in the math methods course for years and seemed to have the potential to capture nuances in

preservice teachers' understanding of proportional reasoning. Therefore, we wanted to investigate whether nuances we have noticed over the years would be detected to some extent by this new statistical approach.

## Measuring Teachers' Knowledge

Capturing the role of teachers and teaching in students' academic outcomes has long been a topic of interest to many scholars and educators. In particular, the failure of proxy measures of teachers and teaching, such as teachers' years of teaching experience and the number of university-level content courses taken (e.g., Begle, 1979; Monk, 1994), has led scholars to shift their efforts toward more accurately identifying the characteristics of effective teachers and their teaching (e.g., Kane et al., 2013).

In this regard, teachers' content knowledge for teaching has been shown to be an important area of research. In the last three decades, numerous assessments have been developed to measure teachers' knowledge for teaching mathematics (Diagnostic Teacher Assessment in Mathematics and Science [DTAMS], 2020; Learning Mathematics for Teaching [LMT], 2004; Tatto et al., 2008). The majority of these assessments have consisted mainly of selected-response items and have been designed to capture teachers' content knowledge and pedagogical content knowledge. Although teachers' scores on these measures have been used to examine the role of their knowledge in teaching and student learning (e.g., Copur-Gencturk, 2015; Baumert et al., 2010; Blazar, 2015; Hill et al., 2005; Hill et al., 2008; Kersting et al., 2012) or the impact of a professional development program on their knowledge gain (Copur-Gencturk, Plowman et al., 2019; Copur-Gencturk & Thacker, 2021), this item format may limit researchers' ability to capture the nuances in teachers' understanding of the subject matter (Lane & Stone, 2006). Specifically, assessments using solely multiple-choice items may not distinguish teachers based

on the nuances in their understanding, which are considered important manifestations of their robust understanding of mathematics (National Research Council, 2001).

Prior research studies in which teachers' knowledge was measured by using different item formats have provided mixed evidence on the nature of teachers' mathematical knowledge or the role of teachers' knowledge in teaching and student learning[2] (e.g., Baumert et al., 2010; Charalambous et al., 2020; Hill et al., 2005; Kersting et al., 2012). As an example, studies investigating the nature of the knowledge needed for teaching mathematics that used only multiple-choice items to measure such constructs have indicated that teachers' content and pedagogical content knowledge is a single construct (e.g., Copur-Gencturk, Tolar et al, 2019; Charalambous et al., 2021). However, studies measuring teacher knowledge by using items that included constructed-response items have suggested that content and pedagogical content knowledge are empirically distinct from one another (Copur-Gencturk & Tolar (under review); Blömeke et al., 2014; Kleickmann et al., 2015; Krauss et al., 2008). We argue that these conflicting results may be related to the fact that constructed-response items were able to capture more nuances in teachers' knowledge that may not have been captured by multiple-choice items alone, which in turn has affected the observed outcomes. Consider, for instance, two groups of teachers who have different levels of understanding of a concept. Although these two groups of teachers may select the same answer on a multiple-choice assessment, the methods they use to arrive at the same answer could be completely different. Yet according to their scores on the multiple-choice assessment, teachers in these two groups would not differ even though their reasoning processes were different.

[2] For more details on these studies, see (Copur-Gencturk & Tolar, under review; Charalambous et al., 2020)

The growing recognition of the challenges of assessing teachers' knowledge more comprehensively has led to the development of teacher knowledge assessments with different item formats. Some research has addressed this issue either by creating assessments with only constructed-response items or by incorporating constructed-response items along with multiple-choice items (Blömeke et al., 2016; DTAMS, 2020; Kersting et al., 2012). Studies including constructed-response items have found a stronger association with teachers' knowledge and student outcomes (e.g., Baumert et al., 2010; Kersting et al., 2012) or have shown that teachers' knowledge is a multidimensional construct, as theorized (e.g., Krauss et al., 2008).

Yet the use of these assessments to measure the impact of a PD program or to investigate the role of teachers' knowledge in teaching and learning is not common, mainly because of the resources needed to train raters and code such a large set of data in a relatively short time. To this end, the purpose of the work reported here is to explore the possibility of using statistical topic modeling to score teachers' responses to an open-ended question and capture the potential nuances in their responses. We first provide an overview of topic modeling and how it is being used in education research and then present our findings.

## Topic Modeling in Education Research

Topic models are statistical models that are designed to analyze a corpus of text and extract the latent themes or topics (Blei, 2012; Blei et al., 2003; Griffiths & Steyvers, 2004). By identifying which words are commonly used by others, the topic model detects the set of latent clusters, referred to as topics, that best fits the data. Thus, the choice of words teachers use to answer a question allows the researcher to identify different constructs underlying the thinking behind their particular choice of words. As a simple example, let us assume that we asked teachers about the nature of the relationship between the two measurable quantities. Let us

assume some teachers focused on the absolute difference between the measurable quantities by noting how more much there was of one quantity than the other (e.g., if they were given the prices of two items, they would notice that the first item was three dollars more expensive than the second item). The first group commonly used words such as *more*, *expensive*, or *difference* to describe the relationship between the two given items. Let us assume that another group focused on the relative difference by noticing how many times as much as there was of one quantity in relation to the other (e.g., the first item was twice as expensive as the second item). They used words such as *times*, *expensive*, *dividing*, and *multiplying* together. The analysis of these people's responses to the same question could then reveal topics that would capture latent, unobservable differences in their characterizations of the nature of the relationship, such as one focusing on the absolute differences between the quantities and the other focusing on relative differences in terms of how one quantity stood in relation to the other. It is important to note that the topics are latent in the corpus; that is, they are not evident upon direct inspection, meaning that all teachers may be using the words *expensive*, *answer*, and *problem* or the numbers presented in the problems and units (e.g., dollars) in their responses, but how these words were associated with each topic could be different. Returning to the previous example, all participants may have used the word *expensive* or the numbers given in a problem in their responses; however, if more people used absolute thinking than relative thinking, then the word *expensive* would be more closely associated with the topic capturing the latent construct of absolute thinking. In fact, all the words in the corpus have a separate probability of occurring with each of the topics in the model. One of the simplest members of the family of topic models is latent Dirichlet allocation (LDA; Blei et al., 2003). LDA is a clustering algorithm that focuses on detecting the latent clusters of terms that co-occur in the corpus. The terms in the corpus are typically words, but they can also be other things, such as mathematical equations or expressions, chemical formulae, images, and so forth. In this study, the

terms on which we focused (referred to as tokens in the natural language processing literature) were the words, operations, equations and numbers teachers used in their responses to a question that required them to explain their reasons for selecting a certain option among the choices in a question. The terms or tokens form the basic building blocks for topic modeling, and each of the clusters of terms is referred to as a topic. An important assumption in the LDA model is that the order of the terms and the grammatical structure in the corpus are not important (Blei, 2012). The LDA model assumes the terms in the collection of documents are essentially an unordered collection (referred to as the "bag of words" assumption). This differs from other kinds of topic models that consider topics based on co-occurrence among tokens plus syntactic similarity in the corpus (e.g. Deerwester et al., 1990). In the present study, LDA was used to detect the latent clusters of terms that co-occurred in a teacher's response to a question designed to measure proportional reasoning. In this way, the resulting topic model captures the co-occurrences of terms that were present in the response, enabling us to interpret these co-occurrences as reflecting the kinds and levels of reasoning present in the response rather the correctness of the response.

LDA has been used in multiple studies to investigate the latent themes in a variety of areas, including detecting voting themes in politics (e.g., Grimmer, 2010; Lau et al., 2012; Lauderdale & Clark, 2014) and themes in Twitter messages (Rhody, 2012). A few studies have attempted to use this technique to analyze educational data. For example, Ramesh et al. (2014) used LDA to analyze latent topics in students' texting to help predict student retention in a MOOC (Massive Open Online Course). Their results suggested that concerns over course logistics or expressions of negative sentiments were predictive of student dropout. Kim et al. (2017) used LDA to detect latent topics in middle grade students' answers to constructed-response test questions. Kim et al. found that changes in students' use of latent topics from pretest to posttest indicated that students decreased their use of everyday language and increased their use of academic language and discipline-specific language when describing the process of scientific inquiry.

In this study, we used supervised LDA (sLDA; Blei & McAuliffe, 2007) which is an extended LDA model that assigns a label or outcome variable to a response, such as a score or grade given to a constructed-response item. Supervised LDA has been shown to improve upon unsupervised LDA in its ability to detect latent topics useful for prediction (Blei & McAuliffe, 2007). In sLDA, the written response and the outcome variable associated with the written response are jointly modeled to detect latent topics that best predict the outcome variable.

Thus, based on the existing literature, we expect that topic modeling could reveal insights regarding teachers' reasoning. In this study, we used sLDA to analyze teachers' responses to a task that asked them to select an answer and explain their reasoning. We tested whether the topic modeling was able to capture nuances in teachers' responses. We further hypothesized that if topic modeling analysis could capture nuances in teachers' responses, then the latent topics identified by this approach might be associated with the qualitative differences in teachers' responses we identified through a separate qualitative analysis as well as through teachers' differential performance on a measure designed to capture their knowledge of the same content. To test this additional hypothesis, we examined the extent to which the latent topics were associated with teachers' overall performance on an assessment measuring the same construct. By using data collected from 240 teachers, we aimed to answer the following questions:

1. What are the characteristics of latent topics detected by topic modeling in teachers' responses?
2. What is the relationship between the reasoning level identified by a qualitative analysis of teachers' responses and the latent topics detected by topic modeling?
3. What is the relationship between teachers' overall performance and the use of the latent topics detected by topic modeling?

# Methods

## Study Context

We used data collected for a project funded by the National Science Foundation to investigate the development of pedagogical content knowledge among mathematics teachers in Grades 3–7. We partnered with an education research company that offers services such as providing the contact and background information of teachers. In this way, we would be able to collect data from teachers in different states and our findings would not be bound to teachers who work in the same school or district.[3] Participants who were eligible for the study (i.e., mathematics teachers in Grades 3–7) completed an online assessment that included questions on ratios and proportional relationships along with other mathematical questions and background questions regarding their years of teaching experience and the number of mathematics courses taken. The mathematics problems in the assessment were presented to each teacher in a randomized order to avoid item order effects. Teachers were not allowed to move on to the next question until they had provided an answer to the question they were on so that we could gather data on the items with which they were struggling.

## Analytic Sample

The analytic sample consisted of 240 teachers who had completed all the questions and answered the task used in topic modeling analyses. Teachers in the sample came from 21 different states in the United States; 84% of the participating teachers were female, and 68% were White. As presented in Table 1, 25 % of the teachers had a master's degree. In addition, 19% held a certification for teaching mathematics, 69% held a certification for teaching multiple

[3] A few organizations in the United States maintain databases that contain information about teachers, including such variables as their email addresses, the subjects taught, and the grade level at which they are currently teaching, among others. These companies provide access to this information for a fee.

subjects, and the remainder held a credential in another field, such as teaching special education. Approximately 71% of the teachers in the sample had earned their teaching credentials through traditional teacher education programs, whereas 20% had entered teaching through alternative programs. During data collection, 24% of them were teaching mathematics in Grades 6 or 7.

**Table 1**

*Descriptive Statistics for Teacher-Level Variables*

| Variable | Sample (%) |
|---|---|
| Teacher background | |
| Gender (female) | 84.0 |
| Ethnicity (White) | 68.1 |
| Master's degree (yes) | 25.2 |
| Teaching level | |
| Elementary school (Grades 3–5) | 75.6 |
| Middle school (Grades 6 & 7) | 24.4 |
| Professional background | |
| Traditional certification | 70.6 |
| Credential in mathematics | 19.3 |
| Credential in multiple subjects | 68.5 |

*Note*. $N = 238$[4].

## Measures

**Proportional Reasoning Task.** To test the possibility of using topic modeling to reveal nuances in teachers' reasoning, we used the following task, in which teachers were asked to select an option and explain their answer.

The Science Club has four separate rectangular plots for experiments with plants. Which rectangle(s) looks more like a square? Explain your answer.

a. 1 foot by 4 feet b. 17 feet by 20 feet

[4] We have missing background data on two participants.

c. 7 feet by 10 feet d. 27 feet by 30 feet

Note that the ratio of the length to width of a rectangle determines how it looks. This ratio is 1:1 for any square. Thus, we expected teachers to notice the relationship between the dimensions of each rectangle and select the one with the ratio closest to 1 as the one that looked the most similar to a square in terms of appearance. We purposefully selected the dimensions of each rectangle because even though the dimensions of the four rectangles differed, each had the same constant unit difference between the length and width. Prior work (e.g., Lamon, 1993; Misailidou & Williams, 2003) has suggested that noticing the constant difference between quantities is commonly used to characterize proportional situations. Thus, this task allowed us to examine whether teachers might be focusing on the three-unit difference between the length and width across the four given rectangles or on the changing ratio of length to width for each given rectangle.

**Ratios and Proportional Relationships Assessment.** We expected that if the topic modeling approach were able to capture nuances in teachers' responses, then latent topics identified based on the proportional reasoning task could be associated with different scores on an assessment capturing the same topic. To develop this assessment, we adapted problems used in the existing literature (e.g., Beckmann, 2017; Schoenfeld, 2015; Van de Walle et al., 2010). The problems were aimed at capturing both theoretically important concepts, such as identifying proportional situations (e.g., Izsák, & Jacobson, 2017; Van Dooren et al., 2005), and solving ratio problems and representing a proportional relationship graphically (Common Core Standards, 2010). The assessment included seven mathematics problems with several subquestions in different formats, ranging from computational problems, multiple-choice

problems (including justifying the selection), evaluation of mathematical situations, missing-value problems, and word problems.

Each constructed-response item was scored by two raters to capture the correctness of the final answer and the accuracy of the reasons or strategies (see Copur-Gencturk, Baek et al (under review) for details). The interrater reliability, measured as the percentage of exact agreement, was greater than 90% for each item. To create a score that would capture teachers' overall performance on this assessment, we calculated a factor score by applying a confirmatory factor analysis (CFA) using Mplus (Muthén & Muthén, 1998–2017). The overall fit of the CFA and the reliability of the assessment were good (RMSEA = 0.107, CF I= 0.966; Cronbach's alpha = 0.71).

**Data Analysis**

To investigate the characteristics of latent topics in teachers' responses (i.e., Research Question 1), we analyzed teachers' written answers to the proportional reasoning task by using an sLDA statistical topic model. Before applying the topic model, teachers' written answers first needed to be preprocessed. This is a standard data-cleaning step used in topic modeling. It is designed to improve the interpretability of the subsequent results as well as to increase the potential for words with the same or similar meanings to cluster together (Schofield, Magnusson, & Mimno, 2017).

The data-cleaning process included the following steps. First, it required stemming: (1) punctuation was removed from all responses, (2) words in something other than first person were changed to first person, (3) all verb tenses were changed to the present tense, (4) all words were reduced to their root form (e.g., plurals were changed to singular form), and (5) typos evident in the answer were corrected. Next, we removed stop words. These are high-frequency but low-

information words, such as *a*, *the*, *that*, *it*, *be*, and *or*. The rationale behind this decision was that these stop words could intrude on the topic modeling results, thereby interfering with the extraction of interpretable latent clusters.

There are three ways to create the stop word list: (1) to use software; (2) to calculate term frequency-inverse document frequency (TF-IDF)[5] values and set a cut point; and (3) to manually create a list using researchers' expert opinions. We used the third option because context can have a significant role in deciding whether a word is a stop word. Thus, we first created a term matrix that included all the terms in the corpus with their frequency. We stemmed words and fixed typos. We then manually selected stop words and created a stop word list. Once we had the stop word list, we wrote an R code to remove the words on the list from the documents. This process was repeated until we obtained a clean pool of terms for the analysis (see Table 2 for the original responses and the cleaned responses used in the reported data analysis as well as the stop words used in the corresponding responses). After stemming and removing stop words, 3,504 total words remained in the set of responses from the 240 teachers in the sample. There were 422 unique words in the set, with a mean number of 14.6 words in each answer and a standard deviation of 12.45.

**Table 2.** Illustration of Cleaned Responses and Stop Words

| Original response | Stop words | Cleaned response |
|---|---|---|
| 27x30, because 27/30 is greater than all the other proportions. | the, is,becuase | twentysevenbythirty twentysevenoverthirty greater than all other proportion |

[5] TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how important a word is to a document in a corpus. This is calculated by using two components: (1) how many times a word appears in a document, and (2) the frequency of the word across a set of documents.

| | | |
|---|---|---|
| D is the most square because the difference in feet is less noticeable as you get a larger rectangle. (3 feet makes a significant size difference for A, and a much smaller difference for D.) | the, is, a, because | d most square difference ft less notice as get larger rectangle three ft make significant size difference a for much smaller difference for d a |
| 27 by 30. If I turn the rectangles into fractions and change the denominator to 1, 27 by 30 provides a numerator closest to 1. | I, and, to, the, a | twentysevenbythirty if turn rectangle into fraction change denominator one twentysevenbythirty provide numerator closest one |

*Note*. We retained the letter "a" when it was referring to option a in the problem.

The next step was to determine how many latent topics appeared in the data. Previous research with the LDA model on answers to open-ended test items suggested from two to five topics was typical (e.g., Choi et al., 2017; Kim et al., 2017; Kwak et al., 2017; Xiong et al., 2019). Therefore, we first conducted an exploratory analysis to determine the best fitting topic model. To do this, we used LDA to estimate topic models containing from two to five topics as candidate models. At present, no single best method is known for determining the best fitting topic model. The following indices are among those that have typically been used to help determine the best fitting topic model: Kullback–Leibler divergence (Arun et al., 2010), cosine similarity (Cao et al., 2009), and the harmonic mean of posterior log-likelihoods (Griffiths & Steyvers, 2004). All three of these indices were computed in this study to inform the selection of the best fitting LDA topic models (i.e., models with two to five topics). In addition, the interpretability of the solutions for each LDA model was considered.

After determining the best fitting LDA candidate models, we added teachers' selection of the correct option on the task to the best fitting LDA model. This resulted in an sLDA topic model that provided the probability of each of the words in the corpus of teachers' response data

for each latent topic, along with a vector of the probabilities of their answers for individual topics. The R package 'lda' (Chang, 2015) was used to estimate each of the four candidate LDA and sLDA topic models. The *lda* program basically works in the following way to estimate the LDA model: Once the data have been cleaned, the program counts how many times each word appears in each response. These word counts are then analyzed, not the actual words themselves. The full set of all these word counts per item in the corpus is then analyzed to determine the probability with which each word co-occurs with each of the other words. These co-occurrences form clusters based on these probabilities. The topics consist of the different latent clusters detected by the software. Because the number of clusters is typically unknown a priori, the *lda* program is run multiple times to estimate the different numbers of these clusters. The best fitting of these candidate models is taken as the solution for the given data. This LDA solution is referred to as an *unsupervised model* because the only criterion for estimating the model is the probable co-occurrence of each of the words.

The sLDA solution is also available from this program. It differs from the LDA solution only in that, in addition to the words in the corpus, additional information about the item is included to estimate the word probabilities. In this study, the correct answer to the proportional reasoning task was used as the outcome variable to guide the estimation of the model in a linear regression. This regression was done simultaneously with the estimation of the topics in an sLDA model. This is called a *supervised model* because the regression on the outcome variable helps guide or supervise the estimation of the latent topics in the given model.

To report the characteristics of the latent topics identified in the best fitting model (i.e., Research Question 1), we examined the highest probability words for each topic. As described by Kim et al. (2017), we also examined the complete responses for each topic from the teachers

whose responses contained the highest probability words associated with a given topic. This helped place the highest probability words for each topic in the original context of the teachers' responses. In reporting our results, we included the actual responses of four teachers who used 100% of the words in a given topic, respectively. By doing so, we aimed to depict accurately what each topic encompassed and how one was different from another.

To validate the quality of the nuances captured by the topic models (i.e., Research Question 2), we also compared the topics identified through topic modeling with the qualitative categories we identified based on a separate analysis of the same data. The categories for the qualitative analysis were based on an initial set of categories drawn from the existing literature (Ben Chaim et al., 1998; Cramer & Post, 1993; Lamon, 1993; Parish, 2010) and patterns we noticed when coding a sample of teachers' responses. Through an iterative process of coding (for further details, see Copur-Gencturk, Baek et al. ( under review) for details), we identified four categories of proportional reasoning in teachers' responses.[6] The *incorrect reasoning* category indicated that teachers did not focus on the quantities that would determine the appearance of a rectangle (i.e., length and width). For instance, one teacher responded to this question by stating, "[Rectangle] C because they are closest to being a perfect square at 9 by 9." The second category, *additive reasoning,* included responses in which teachers focused on the correct quantities (i.e., length and width) but attended to the constant difference between the length and width to determine the appearance of the rectangles. One teacher whose response fell in this category explained, "They all have a difference of 3 feet between the length and width. I determined, then, that they would all have the same 'squareness." The third category, *relative reasoning*, captured responses that indicated teachers identified the correct quantities (i.e., length

[6] The first author trained another rater, and they coded the responses separately, reaching 97% exact agreement. All responses were coded by two raters.

and width) and reported the constant differences between length and width for each rectangle. Yet unlike the teachers in the additive reasoning category, they reported the relative impact of the length and width on the appearance of the rectangles when the dimensions changed. One such response in this category was, "Because, since the side lengths are longer than all of the other rectangles, the 3 feet difference is less noticeable." Finally, the fourth category, *proportional reasoning*, consisted of responses that involved attending to the quotient of the length and the width to determine which rectangle looked more like a square. As one teacher whose response fell into this category stated,

> Rectangle D because 1 to 4 is just 25% as long as the other side, 7 to 10 is 70% as long as the other side, 17 to 20 is 85% as long as the other side, and 27 to 30 is around 90% as long as the other side.

To investigate the extent to which topic modeling captured nuances in teachers' responses, we reported the percentage of all responses that used only the highest frequency words in each topic and the category in which they were coded according to the qualitative analysis. To examine the relationship between teachers' overall performance and the use of individual topics (i.e., Research Question 3), we used an ordinary regression analysis on teachers' individual topic scores to predict their overall performance.

## Results

### Characteristics of the Latent Topics Detected by Topic Model

Table 3 presents results for the three fit statistics for the LDA models. We selected the four-topic model because it was suggested by two of the three fit indices, the harmonic mean of log-likelihoods and the cosine similarity. We also considered the interpretability of the four-topic model to finalize our model selection.

**Table 3**

*Summary of Fit Statistics for the Latent Dirichlet Allocation Models*

| Number of topics in the model | Harmonic mean of log-likelihoods | Cosine similarity | Kullback–Leibler divergence |
|---|---|---|---|
| 2 | −17635 | 0.493 | 201.96 |
| 3 | −17537 | 0.486 | 190.69 |
| 4 | **−17506** | **0.413** | 179.44 |
| 5 | −17564 | 0.473 | **175.84** |

*Note*. For the harmonic mean of log-likelihoods, the larger the value, the better the fit; for the cosine similarity and Kullback–Leibler divergence, a smaller value indicates a better fit. The values in bold for each of the indices indicate the best fit of the four models.

Table 4 shows how the words used by teachers to explain their reasoning about which rectangle was more like a square were associated with each topic. Note that the topics were not categorical; rather, every word in the corpus appeared in each topic but might have a different probability of appearing in each topic. For instance, word *choice D* (*correct answer*) was a high probability word for Topics 2, 3, and 4, but the probability of this word appearing in the Topic 3 construct was almost twice as likely as those appearing in Topics 2 and 4.

**Table 4**

*Top 10 Highest Probability Words for Each Topic*

| Topic 1 | | Topic 2 | | Topic 3 | | Topic 4 | |
|---|---|---|---|---|---|---|---|
| Word | Prob. | Word | Prob. | Word | Prob. | Word | Prob. |
| *square* | .061 | *percent* | .085 | *choice D* | .086 | *ft* | .076 |
| *look* | .037 | *choice D* | .047 | *closest* | .080 | *square* | .072 |
| *answer* | .035 | *as* | .040 | *ratio* | .065 | *choice D* | .045 |
| *so* | .033 | *closer* | .035 | *one* | .058 | *all* | .042 |
| *not* | .033 | *side* | .032 | *square* | .054 | *three* | .041 |
| *think* | .030 | *other* | .032 | *side* | .034 | *difference* | .035 |
| *each* | .026 | *proportion* | .030 | *most* | .032 | *rectangle* | .034 |
| *choice A* | .022 | *90* | .025 | *which* | .031 | *length* | .028 |
| *close* | .022 | *than* | .022 | *27 by 30* | .028 | *most* | .026 |
| *like* | .020 | *number* | .022 | *length* | .021 | *width* | .022 |

*Note*. Prob. = probability of the word falling into that topic.

Given that these topics indicate latent constructs underlying teachers' thinking, we also looked at teachers' actual responses whose responses consisted of using only the words that were closely associated with a particular topic. Specifically, as shown in Table 4, the highest probability words associated with Topic 1 were not specific to words used to describe ratios or proportions. The actual responses of teachers who used only words highly associated with Topic 1 (see the "Teacher Answer" column in Table 5) provided a better glimpse of what Topic 1 captured. Specifically, Topic 1 seemed to capture the responses of teachers who were not using correct reasoning. For instance, one of the responses in this category was, "A is most square because it does not have a remainder or decimal points."

**Table 5**

*Answers from the Teachers with the Highest Probability of Using Each Topic and the Reasoning Level Assigned in the Qualitative Analysis*

| Topic | Prob. | Teacher's answer | Proportional reasoning level according to the qualitative analysis |
|---|---|---|---|
| 1 | 1.00 | A is most square because it does not have a remainder or decimal point. | Incorrect |
| 1 | 1.00 | All of them are proportionate. | Incorrect |
| 1 | 1.00 | C, it's the closest to a square. 7 and 10 are pretty close to each other. | Incorrect |
| 1 | 1.00 | 1 by 4. Its dimensions are the most proportional. | Incorrect |
| 2 | 1.00 | Rectangle D because 1 to 4 is just 25% as long as the other side, 7 to 10 is 70% as long as the other side, 17 to 20 is 85% as long as the other side and 27 to 30 is around 90% as long as the other side. | Proportional |
| 2 | 1.00 | D. Its smaller side is 90% of its larger side. Closer to 100% than any other answer. | Proportional |
| 2 | 1.00 | D, the proportions appear closer the higher the numbers go. | Relative |
| 2 | 1.00 | 27 × 30, because 27/30 is greater than all the other proportions. | Proportional |
| 3 | 1.00 | D is the most square because the ratio is closer to equaling 1, which is what you would want a square to be. | Proportional |
| 3 | 1.00 | D. When dividing, the 27 feet by 30 feet is closer to one whole making it closer to a full square. | Proportional |
| 3 | 1.00 | D. You can find out which fraction is closer to 1, and 27/30 is closer to 1. | Proportional |
| 3 | 1.00 | D. 27 to 30, because the ratio of sides 27/30 is 0.9 and that is the closest ratio to 1. | Proportional |
| 4 | 1.00 | They all are the most square because they have a three foot difference on each. | Additive |
| 4 | 1.00 | Neither because each square has a difference of 3 in the side lengths. | Additive |
| 4 | 1.00 | They are all the same. They are 3 feet away from being square. | Additive |

| | | | |
|---|---|---|---|
| 4 | 1.00 | They will all be the same amount of square because the lengths and widths all have the same difference of 3 ft. | Additive |

*Note*. Prob. = probability of the teacher using each topic in his/her response.

Unlike the words characterizing Topic 1, the highest probability words associated with Topics 2 and 3 included mathematical terms used to describe proportional relationships. The completed responses of teachers who used only the words that were closely associated with Topic 2 suggested that the underlying construct associated with Topic 2 was how one quantity was related to the other. In fact, both of the highest probability words for Topic 2 described how the length and width of each rectangle were related to the other proportionally, such as *ninety*, *eighty-five*, and *percent*, and 100% of the responses in the Topic 2 category included such computations. As an example, one teacher noted,

> Rectangle D because 1 to 4 is just 25% as long as the other side, 7 to 10 is 70% as long as the other side, 17 to 20 is 85% as long as the other side, and 27 to 30 is around 90% as long as the other side.

Words with the highest probability of falling in Topic 3 and teachers' responses using only words highly associated with Topic 3, such as *ratio*, *side*, *closest*, and 1, also provided a glimpse of how this topic captured words used to describe proportions. Indeed, this topic seemed to capture teachers' proportional reasoning and how the ratio of length to width of a rectangle should be closer to 1 to look more like a square. As shown in Table 5, one of the responses from a teacher who used all the high-probability words from Topic 3 was, "D. When dividing, the 27 feet by 30 feet is closer to one whole, making it closer to a full square." Thus, although both Topics 2 and 3 seemed to detect responses that showed evidence of proportional reasoning, these two topics differed in the way the teachers reasoned to find the correct solution.

The highest probability words associated with Topic 4 (e.g., *all*, *three*, *ft*, *difference*, *rectangles*, *square*, and *most*) and the teachers' responses that used all the words highly associated with Topic 4 suggested that this topic captured teachers who noticed the constant

three-unit difference between length and width for each rectangle. For instance, a teacher whose responses used all the words highly associated with Topic 4 justified her or his answer by writing, "They will all be the same amount of square because the lengths and widths all have the same difference of 3 ft."

**Comparison of the Topics and the Qualitative Analysis of Teachers' Responses**

As mentioned in the Methods section, our separate analysis of the same data indicated that teachers' responses could be categorized into four groups: incorrect reasoning, additive reasoning, relative reasoning, and proportional reasoning. Each response was double coded and assigned to one of these four categories based on the qualitative analysis. Our separate analysis of the same data set using the topic modeling approach also resulted in four topics, but unlike the qualitative categories, each response had a different probability of being associated with all four topics. For this reason, we looked at the responses of teachers who were using high-probability words for each topic and how they were categorized according to the qualitative analysis.

We found that the responses that included all the words highly associated with Topic 1 were categorized as incorrect reasoning. Similarly, the responses that used only the highest frequency words in Topic 3 were coded as falling in the proportional reasoning category according to the qualitative analysis. Eighty three percent of the responses of teachers who used only high-probability words from Topic 2 were coded as proportional or relative reasoning according to the qualitative analysis. When we looked at the responses that used primarily high-probability words from Topic 4, we found that 59% these responses were coded as relative or additive thinking according to the qualitative analysis.

**Linking Topics to Teachers' Knowledge of Ratios and Proportional Relationships**

We examined the relationship between the topics and teachers' overall understanding of ratios and proportional relationships to test the premise that if the topics were indeed capturing some important insights regarding teachers' proportional reasoning, then teachers' understanding of ratios and proportional relationships on this measure should vary depending on the use of different topics. Given that Topics 1 and 4 did not seem to capture correct proportional thinking, we expected them to be negatively associated with teachers' overall performance on the ratios and proportional relationships measure. Using the same logic, we expected the use of Topics 2 and 3 to be positively associated with teachers' scores on this measure.

As can be seen in Figure 1, the results suggested that the use of Topics 2 and 3 was associated with higher scores on the ratios and proportional reasoning measure, whereas the use of Topics 1 and 4 was associated with lower total scores on the assessment. The expected mean scores were −0.41 and −0.23 for teachers who used only Topic 1 and 4 words, respectively. Recall that the overall performance on the proportional reasoning task was −0.1 with a standard deviation of 0.79. Thus, this finding indicates that teachers who used words highly associated with Topics 1 and 4 did not perform well on the ratios and proportional relationships assessment.

On the contrary, the use of words in Topics 2 and 3 was positively linked to teachers' overall performance on the ratios and proportional relationships assessment. The expected mean scores of teachers who used only Topic 3 or 4 words on the assessment were 0.32 and 0.11, respectively. Furthermore, the total scores of those who used words highly associated with Topics 2 and 3 were statistically higher than the total scores of those who used words highly associated with Topics 1 and 4 (for Topic 2, $p < .01$ for Topic 1 and $p < .05$ Topic 4; for Topic 3, $p < .05$ for both Topics 1 and 4).

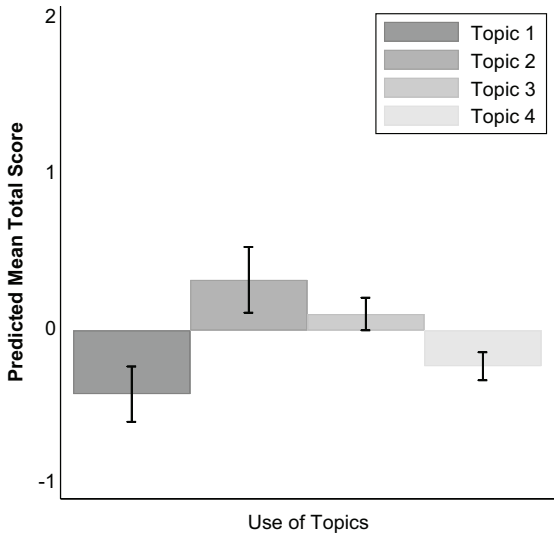



*Figure 1*. Predicted Mean Scores on the Ratios and Proportional Reasoning Assessment by the Use of Each Topic. *Note*. The error bar for each topic indicates teachers' scores within one standard deviation for the corresponding topic.

## Discussion

Prior work on the assessment of teacher knowledge has mainly used multiple-choice items or predetermined coding categories to capture teachers' knowledge. Yet scholars have raised concerns regarding the limitations of such an approach for capturing potential nuances of teachers' knowledge and understanding. In this study, we aimed to investigate the possibility of using open-ended items and a statistical approach, topic modeling, to capture nuances in teachers' responses that could signal differences in their understanding. Our findings indicated that the topics identified through topic modeling were associated with different reasoning levels identified in a separate qualitative analysis as well as with varying performance on the same topic. We believe this is an important finding with implications for using open-ended items in large-scale studies and analyzing them by applying topic modeling to provide insights into teachers' understanding.

Our findings provide initial evidence that topic modeling can capture differences in teachers' understanding. As we reported, the topics and qualitative reasoning categories seemed to be in alignment, and teachers who used Topics 1 and 4 did not show a strong understanding of ratios and proportional relationships. Similarly, teachers who used Topics 2 and 3 appeared to use proportional reasoning and showed a strong understanding of ratios and proportional relationships. Our findings also indicated that the topic modeling approach allowed us to identify teachers based on their thinking and reasoning. For instance, teachers who used Topic 2 seemed to focus more on how one side looked in relation to the others, whereas those who used Topic 3 seemed to focus more on the fact that the closer the ratio of the length to the width was to 1, the more like a square it looked.

We believe this result has important implications for future assessment design by offering the possibility of using topic modeling to analyze constructed-response items in large-scale studies. It is important to note that we do not claim that topic modeling is a stand-alone approach to analyzing the data. Rather, we argue that topic modeling could be used in conjunction with other approaches, such as qualitative analyses. For instance, analyzing data collected from hundreds of teachers qualitatively might take a substantial amount of time; however, using topic modeling could help researchers identify the potentially distinct topics (or groups) into which teachers might be categorized. Analyzing a sample of responses that are most characteristic of each topic (i.e., using only the words highly associated with a given topic) could then allow researchers to characterize the nuances in topics in a relatively short period of time.

Our motivation behind this study was to find a way to the increase the use of rich data in quantitative, large-scale studies, given that constructed-response items have the potential to capture nuances in teachers' understanding that might not be captured through multiple-choice

items (e.g., Copur-Gencturk, 2021a; Copur-Gencturk, 2021b;, Copur-Gencturk & Doleck, 2021;, Copur-Gencturk & Olmez, 2021; Goulding et al., 2002; Tatto, 2013). Because of the aforementioned cost- and time-related issues, the use of constructed-response items in large-scale studies makes analyzing the data in a timely manner difficult. We contend that this problem leads scholars, particularly policy makers, to tend to rely on selected-response questions more often. Yet studies using constructed-response items add to our understanding of teacher knowledge and its role in teaching and student learning in more pronounced ways (e.g., Baumert et al., 2010; Kersting et al., 2012). Therefore, finding ways that will lead to the use of richer data sets to better understand these complex issues has important implications for research and teacher education. Our findings provide initial evidence for the usability of the topic modeling approach for this purpose. As captured by the different topics, teachers who answered the question correctly (incorrectly) could have used different reasoning to arrive at their final answer. And the different reasoning levels captured in the topics had varying degrees of association with their performance on a measure covering the same content. It is possible that teachers who reason differently might create different learning environments for their students, which in turn could influence student learning outcomes. For instance, the teachers who used Topic 2 (i.e., focusing on how one quantity relates to the other) might create a different learning environment than teachers who used Topic 3 (i.e., focusing on the quotient of quantities). Taken altogether, these results suggest the topic modeling approach could be used to investigate how different topics (and therefore underlying constructs) are associated with instruction and student learning as well as with teachers' knowledge development.

In this study, we focused only on teachers' responses to a mathematics question. Further research is needed on the extent to which topic modeling could be used to capture qualitative

differences in other aspects of teachers' professional content knowledge for teaching. Our initial investigations of different types of items, such as solving mathematics problems and analyzing mathematics instruction, suggested that a wide range of items could be used for topic modeling (Hong et al., 2022). On the basis of the prior literature on topic models as well as our experience in using different types of items in our ongoing investigations, the number of items in the sample being analyzed and the average number of words in the set of items seemed to be important quantities in getting a topic model to converge and to clearly reflect the latent thematic structure in the set of items (Mardones Segovia et al., 2021; Wheeler et al., 2020). The larger the number of words in the text being analyzed, the greater the likelihood that the model would converge.

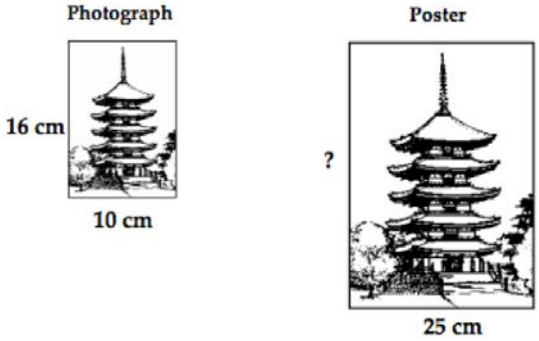
Related to this point, one of the implications for this study pertains to the assessment design. The key issue in getting a topic model to converge is to have enough words in the data set to inform the model, so topic modeling can be used with a smaller number of items that have a sufficient number of words. Thus, asking teachers to provide detailed responses to fewer items could capture important aspects of teachers' knowledge and reasoning and could increase the probability of using open-ended assessment items to capture teachers' knowledge for teaching as well as teachers' learning from professional development. For instance, rather than simply asking teachers to answer a set of multiple-choice items, incorporating a few constructed-response items and applying topic modeling could provide insights into changes in teachers' reasoning.

In conclusion, we argue that teachers' knowledge and understanding could be captured more accurately through constructed-response items. Yet conducting large-scale studies to investigate teachers' knowledge and learning and how a nuanced understanding may relate to student learning and instruction are particularly challenging. Although qualitative work provides insights into teachers' thinking, issues of cost and time have previously limited the possibility of

using open-ended items on assessments in large-scale studies. We believe our study provides evidence that topic modeling could be used in conjunction with qualitative analysis to investigate teachers' responses to open-ended questions in a timely fashion to capture teachers' nuanced understanding at scale.

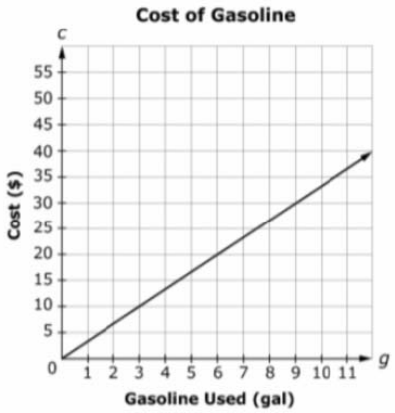
## Appendix

***Ratios and Proportional Relationships Assessment***

| Tasks |
|---|
| A photograph is enlarged to make a poster. The photograph is 10cm wide and 16cm high.<br>  <br><br>• The poster is 25cm wide, how high is the poster? Explain your answer.<br>• If 6ml of paint was needed for the original photograph, how much paint will be needed for the enlarged photo? Explain your answer. |
| Raymond wanted to know the cost of buying different numbers of songs for his MP3 player. The cost of each song is the same. Let s represent the possible number of songs Raymond could buy Let d represent the amount of money, in dollars, Raymond would need to buy the songs. Fill in the table for all missing values of s and d. |

| Number of Songs<br>s | Amount of Money ($)<br>d |
|---|---|
| 2 | 2.50 |
| | 5.00 |
| 7 | |
| | 22.50 |

| |
|---|
| Yasmin went to the store to buy a new purse. The purse she wanted was on sale for 40% off the original price, and the salesperson offered her an additional discount of 15% off the sale price. The salesperson told Yasmin that this was a great deal and that she was getting a discount of 55% off the original price. Do you agree with the salesperson? Please explain. |
| A recipe requires $\frac{1}{4}$ cups of flour for every $\frac{1}{3}$ batch of cookies. How many batches of cookies can be made with $5\frac{1}{2}$ cups of flour? Explain how you solved the problem. |
| This graph shows the relationship between the number of gallons of gasoline used (g) and the total cost of gasoline (c). |




1. How much money will be paid if 12 gallons of gasoline are used?
2. Write an equation to find the cost for any given amount of gasoline used. Explain how you found the equation.

---

Of following word problems, which represent equivalent ratios? Select all that represent equivalent ratios.

- ○ If *a* men paint the outside of a house in *b* minutes, then how many minutes *d* would it take *c* men to paint the same house, if all the men work at the same rate?
- ○ A leaky faucet was dripping water into a bucket. There was already some water in the bucket before Latika started collecting data. She found that there were *a* ounces of water in the bucket after *b* minutes. How many ounces of water *c* will be in the bucket after *d* minutes?
- ○ Bob and Marty run laps together because they run at the same pace. Today, Marty started running before Bob came out of the locker room. Marty had run *a* laps by the time Bob had run *b* laps. How many laps *c* had Marty run by the time that Bob had run *d* laps?

Explain your reasoning.

## References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.

Abt Associates, Inc. (2013, November). *Mathematics and science partnerships: Summary of performance period 2011 annual reports*. Retrieved from http://www.ed-msp.net/images/public_documents/document/annual/MSP%20PP11%20Annual%20Final%20Report.pdf

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010, June). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran and Vikram Pudi (eds.). *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-13657-3_43

Ball, D.L., Thames, M.H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59(5),* 389-407.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133-180.

Beckmann, S. (2017). *Mathematics for elementary teachers with activities*. Pearson.

Begle, E. (1979). *Critical variables in mathematics education: Findings from a survey of the empirical literature*. Washington, DC: Mathematical Association of America and National Council of Teachers of Mathematics

Ben-Chaim, D., Fey, J. T., Fitzgerald, W. M., Benedetto, C., & Miller, J. (1998). Proportional reasoning among 7th grade students with different curricular experiences. *Educational Studies in Mathematics*, *36*(3), 247-273.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16-29.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), pp. 77-84.

Blei, D.M., Ng, A.Y., & Jordan, M., I (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Blömeke, S., Houang, R. T., & Suhl, U. (2014). Diagnosing teacher knowledge by applying multidimensional item response theory and multiple-group models. In S. Blömeke, F. J. Hsieh, G. Kaiser, & W. H. Schmidt (Eds.), *International perspectives on teacher knowledge, beliefs and opportunities to learn: TEDS-M results* (pp. 483-501). Springer Netherlands.

Blömeke, S., Busse, A., Kaiser, G., König, J., & Suhl, U. (2016). The relation between content-specific and general teacher knowledge and skills. *Teaching and Teacher Education*, *56*, 35-46.

Borko, H., Eisenhart, M., Brown, C. A., Underhill, R. G., Jones, D., & Agard, P. C. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? *Journal for Research in Mathematics Education, 23*(3), 194-222.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781.

Chang, J. (2015). lda: Collapsed Gibbs Sampling Methods for Topic Models. R package version 1.4.2, URL http://CRAN.R-project.org/package=lda.

Charalambous, C. Y., Hill, H. C., Chin, M. J., & McGinn, D. (2020). Mathematical content knowledge and knowledge for teaching: exploring their distinguishability and contribution to student learning. *Journal of Mathematics Teacher Education*, *23*(6), 579-613.

Cramer, K., & Post, T. (1993). Making connections: A case for proportionality. *The Arithmetic Teacher*, *40*(6), 342-346.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41* (6), 391–407.

DTAMS (2020, March 28). DTAMS Home. https://louisville.edu/education/centers/crimsted/dtams

Gordon, R. J., Kane, T. J., & Staiger, D. (2006). *Identifying effective teachers using performance on the job* (pp. 2006-01). Washington, DC: Brookings Institution.

Goulding, M., Rowland, T., & Barber, P. (2002). Does it matter? Primary teacher trainees' subject knowledge in mathematics. *British Educational Research Journal*, *28*(5), 689-704.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235. http://doi.org/10.1073/pnas.0307752101

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis, 18*, 1-35.Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430-511.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371-406.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*(4), 372-400.

Hong, M., Choi, H.-J., Mardones-Segovia, C.A., Copur-Gencturk, Y., & Cohen, A.S. (2022, accepted). Two-step approach to topic modeling to incorporate covariates and outcome. Wiberg, M., Molenaar, D., González, J., Kim, J.-S., & Hwang, H. (Eds.) *Quantitative Psychology.* Springer Proceedings in Mathematics & Statistics. Springer, Cham.

Izsák, A., & Jacobson, E. (2017). Preservice teachers' reasoning about relationships that are and are not proportional: A Knowledge-in-Pieces Account. *Journal for Research in Mathematics Education*, *48*(3), 300-339.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49(3),* 568-589.

Kim, S., Kwak, M., Cardozo-Gaibisso, L.A., Buxton, C.A., & Cohen, A.S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics*, *1*, 82-102.

Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., Cheo, M., & Baumert, J. (2015). Content knowledge and pedagogical content knowledge in Taiwanese and German mathematics teachers. *Teaching and Teacher Education*, *46*, 115-126.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, *100*(3), 716.

Lane, S., & Stone, C.A. (2006). Performance testing. In Brennan, R.L. (Ed.), *Educational Measurement, 4th edition*, American Council on Education.

Learning Mathematics for Teaching. (2020, March 20). LMT Project. http://www.umich.edu/~lmtweb/.

Lamon, S. (1993). Ratio and Proportion: Connecting Content and Children's Thinking. *Journal For Research in Mathematics Education*, *24*(1), 41. doi: 10.2307/749385

Lau, J. H., Collier, N., & Baldwin, T. (2012). On-line trend analysis with topic models:# twitter trends detection topic model online. *Proceedings of COLING 2012*, 1519-1534.

Lauderdale, B. E., & Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science, 58*, 754-771.

McAuliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121-128).

Misailidou, C., & Williams, J. (2003). Diagnostic assessment of children's proportional reasoning. *The Journal of Mathematical Behavior*, *22*(3), 335-368. doi: 10.1016/s0732-3123(03)00025-7

Monk, D. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*(2), 125-145.

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide*. Eighth Edition.

National Research Council (2001). *Adding it up: Helping children learn mathematics*. National Academies Press. Washington, DC: National Academy Press.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington, DC: Authors.

Nye, B., Konstantopoulos, S. and Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis (26)3,* 237-257.

Ottmar, E. R., Rimm-Kaufman, S. E., Larsen, R. A., & Berry, R. Q. (2015). Mathematical knowledge for teaching, standards-based mathematics teaching practices, and student achievement in the context of the responsive classroom approach. *American Educational Research Journal*,*52*(4), 787-821.

Parish, L. (2010). Facilitating the Development of Proportional Reasoning through Teaching Ratio. *Mathematics Education Research Group of Australasia*.

Ramesh, A., Goldwasser, D., Huang, B., Daume III, H., & Getoor, L. (2014). Understanding MOOC discussion forums using seeded LDA. In *Proceedings of 9th Workshop on Innovative Use of NLP for Building Educational Applications*. 28-33.

Rhody, L. (2012). Topic modeling and figurative language. *Journal of Digital Humanities*, *2*(1), pp. 19-35.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one?. *Education*, *6*(1), 43-74.

Schoenfeld, A. H. (2015). Summative and formative assessments in mathematics supporting the goals of the common core standards. *Theory Into Practice*, *54*(3), 183-194.

Schofield, A., Magnusson, M., & Mimno, D. (2017). Understanding text pre-processing for latent Dirichlet allocation. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics* (Vol. 2, pp. 432-436).

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15 (5),* 4-14.

Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher education and development study in mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics. Conceptual framework.* East Lansing, MI: Teacher Education and Development International Study Center, College of Education, Michigan State University. Retrieved from https://msu.edu/user/mttatto/documents/TEDS_FrameworkFinal.pdf

Tatto, M. T. (2013). *The Teacher Education and Development Study in Mathematics (TEDS-M): Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries. Technical Report*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

Van de Walle, J. A., Karp, K. S., Bay-Williams, J. M., & Wray, J. (2010). *Elementary and middle school mathematics: Teaching developmentally.* Boston, MA: Pearson

Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., & Verschaffel, L. (2005). Not Everything Is Proportional: Effects of Age and Problem Type on Propensities for Overgeneralization. *Cognition and Instruction*, *23*(1), 57-86. doi: 10.1207/s1532690xci2301_3

Xiong, H., Cheng, Y., Zhao, W., & Liu, J. (2019). Analyzing scientific research topics in manufacturing field using a topic model. *Computers & Industrial Engineering*, *135*, 333-347.