Computers & Education 193 (2023) 104682

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Machine and expert judgments of student perceptions of teaching behavior in secondary education: Added value of topic modeling with big data

Bilge Gencoglu [a,*], Michelle Helms-Lorenz [a], Ridwan Maulana [a], Ellen P.W.A. Jansen [a], Oguzhan Gencoglu [b]

[a] Teacher Education, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, the Netherlands
[b] Top Data Science Ltd. Helsinki, Finland

ARTICLE INFO

Keywords:
Secondary education
Data science applications in education
Topic modeling
Student perceptions of teaching behavior

ABSTRACT

Research shows that effective teaching behavior is important for students' learning and outcomes, and scholars have developed various instruments for measuring effective teaching behavior domains. Although student assessments are frequently used for evaluating teaching behavior, they are mainly in Likert-scale or categorical forms, which precludes students from freely expressing their perceptions of teaching. Drawing on an open-ended questionnaire from large-scale student surveys, this study uses a machine learning tool aiming to extract teaching behavior topics from large-scale students' open-ended answers and to test the convergent validity of the outcomes by comparing them with theory-driven manual coding outcomes based on expert judgments. We applied a latent Dirichlet allocation (LDA) topic modeling analysis, together with a visualization tool (LDAvis), to qualitative data collected from 173,858 secondary education students in the Netherlands. This data-driven machine learning analysis yielded eight topics of teaching behavior domains: *Clear explanation, Student-centered supportive learning climate, Lesson variety, Likable characteristics of the teacher, Evoking interest, Monitoring understanding, Inclusiveness and equity, Lesson objectives and formative assessment.* In addition, we subjected 864 randomly selected student responses from the same dataset to manual coding, and performed theory-driven content analysis, which resulted in nine teaching behavior domains and 19 sub-domains. Results suggest that the relation between machine learning and human analysis is complementary. By comparing the bottom-up (machine learning analysis) and top-down (content analysis), we found that the proposed topic modeling approach reveals unique domains of teaching behavior, and confirmed the validity of the topic modeling outcomes evident from the overlapping topics.

1. Introduction

For decades, researchers have investigated factors influencing student achievement, and effective teaching behavior has emerged as an essential factor (Creemers, 1994). Effective teaching behavior can be conceptualized as teachers' behavior that positively

* Corresponding author. Department of Teacher Education, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712TS, Groningen, the Netherlands.
E-mail address: b.gencoglu@rug.nl (B. Gencoglu).

https://doi.org/10.1016/j.compedu.2022.104682
Received 20 September 2021; Received in revised form 5 September 2022; Accepted 13 November 2022
Available online 17 November 2022

influences students' learning and outcomes (Good et al., 2009). This behavior involves personal and professional characteristics, classroom climate and management, pedagogical knowledge, and teacher–student relationships (e.g., Danielson, 2013; Pianta & Hamre, 2009; van de Grift, 2007).

Teaching behavior is generally viewed as multidimensional (Hattie, 2009; Kyriakides et al., 2009; Muijs et al., 2005). Growing attention has been directed towards identifying components of teaching behavior, i.e., teaching behavior domains, that have a substantial effect on student outcomes (Scheerens, 2016; Seidel & Shavelson, 2007), and promoting teaching quality in learning environments (Smale-Jacobse et al., 2019; Kyriakides et al., 2020; Panayiotou et al., 2014; van de Grift, 2007). Scholars have developed various instruments for measuring these domains, and a frequently used instrument is student questionnaires.

Research using student questionnaires has yielded a great deal of knowledge on student perceptions of teaching behavior. Despite these efforts, however, a limited number of studies use big written text data and apply the machine learning approach to analyze the qualitative data. This novel approach is a promising way to capture what analyses of student questionnaires have not uncovered to date. The present study aims to apply a machine learning approach to use the benefit of big written text data—namely, students' responses to an open-ended questionnaire. Furthermore, we apply manual analysis of a sample of the same open-ended question to further assess the validity of the outcomes of machine learning approach.

One practical reason for the lack of studies using and analyzing big written text data is that manual qualitative coding is expensive, laborious, and time-consuming (Feinerer, 2007; Schonlau & Couper, 2016). Another drawback is that the current qualitative data analysis tools (e.g., NVivo, Altas. ti) are not sufficiently equipped to allow researchers to easily code text data of large samples (Longo, 2020). Although technology can accelerate the speed of the data collection process and allows for the collection of multidimensional information from larger sample sizes, interpreting each comment and subsuming every opinion remain highly challenging when analyzing big text data (Balahadia et al., 2016).

These practical difficulties have limited researchers in using big written text data, even though the added value of large-scale surveys is incontrovertible. Large-scale surveys are usually used for generalizations and monitoring trends. For example, some popular large-scale comparative surveys, such as the Teaching and Learning International Survey (TALIS) and the Programme for International Student Assessment (PISA) of the Organization for Economic Co-operation and Development (OECD), allow researchers to create a general snapshot of teachers' working conditions and student achievement and compare trends to detect differences and similarities across countries and economies, as well as in various demographic subgroups over time. These large-scale survey data represent contemporary opinions at a certain moment in time (snapshot), which are conducive to detecting context- and time-dependent trends. However, these large-scale surveys are mostly based on Likert-type self-reports or binary categories.

Recently, smart and automated text analytic techniques have emerged that allow researchers to cost-efficiently operate large-scale written text data in their findings. Education researchers need to implement this more intelligent solution, which can use elaborative information inherent in the big qualitative data efficiently in a way that minimizes the time and labor demand for qualitative data analysis (Longo, 2020). One potentially promising method that offers a solution to this challenge is the machine learning approach. Thus, this study applies a machine learning tool to the education field as a novel approach to explore students' contemporary perceptions of teaching behavior.

1.1. *Effective teaching behavior domains*

Educational researchers view effective teaching behavior as a complex construct made up of multiple behavior domains. Several studies have integrated various teaching behavior domains to determine effective teaching behavior that promotes student learning (see, e.g., Creemers & Kyriakides, 2008; Pianta & Hamre, 2009). A framework that has gained much attention in recent years suggests that there are three basic domains of teaching behavior that matter most for student outcomes: classroom management, supportive climate, and cognitive activation (Pianta & Hamre, 2009; Praetorius et al., 2018). Although several extant theoretical frameworks on effective teaching behaviors incorporate various domains, the instruments are limited to Likert-scale questionnaires and classroom observations using rating scales.

The International Comparative Analysis of Learning and Teaching (ICALT) proposes an inclusive and comprehensive framework based on the effective teaching behavior model (van de Grift, 2007, 2014). The model identifies six domains of observable effective teaching behavior that lead to positive learning and outcomes (Inda-Caro et al., 2019; van de Grift, 2014): Learning Climate, Classroom Management, Clarity of Instruction, Activating Teaching, Differentiated Instruction, and Teaching Learning Strategies. In this study, we take these six domains as a starting point in theory-driven manual coding to examine the likelihood of unveiling contemporary domains in the iterative process of coding from the student's point of view.

1.2. *Student voice on teaching behavior*

Student perceptions are important for at least three reasons. First, student perceptions of teaching behavior have a direct effect on learning processes and outcomes (Maulana & Helms-Lorenz, 2016; Könings et al., 2005, 2011). What students learn in the classroom depends on how they perceive, interpret, and process the information transmitted by the teacher (Shuell, 1993; 1996a; 1996b). In particular, how students feel about the teaching methods is an essential predictor of their academic performance, learning attitudes, and school life (N. Kim & Son, 2021). Second, to be able to foster students' knowledge and skill development, teachers should be aware of their students' achievement, motivation, and learning preferences (Maulana et al., 2019; Smale-Jacobse et al., 2019). However, teachers often have limited insights into the learning strategies, perceptions, desires, and needs of students in the classroom (Auwarter & Aruguete, 2008; Lynch & Salikhova, 2017; Watkins, 2004). Thus, being aware of students' perceptions of teaching behavior could

help teachers align their teaching behavior with students' learning needs. Third, students' reports of teaching behavior are an important measurement strategy that teachers as well as other stakeholders such as school principals, educators, researchers, and policy makers can use to monitor teachers' professional development at the local, regional, and national levels (Helms-Lorenz et al., 2018). Researchers propose that student feedback can help teachers understand more about students' learning process, preferences, needs, and actual impact of their teaching on learning (Looney, 2011; Mandouit, 2018). In particular, student ratings of teacher effectiveness provide insights for teacher growth (H. Chen et al., 2021), and receiving feedback on their teaching behavior gives insights into their teaching process (Göbel et al., 2021).

Besides student ratings, students' answers to open-ended questions could also benefit teachers' professional development, thus qualitative research that investigates the student voice is substantial. The idea of incorporating the student voice in educational processes upholds practices that offer students space for speaking about their learning (Finefter-Rosenbluh, 2020). Sanchez et al. (2012) claim that investigating student perceptions of teaching methods would deepen understanding of the learning process, and find effective teaching methods to improve student performance. Studies have acknowledged that students are capable of describing teaching behaviors that can help them attain their learning outcomes, though findings vary depending on respondent age and questionnaire characteristics (Alkan, 2013; Bakx et al., 2015; García-Moya et al., 2020; Ida, 2017; Kutnick & Jules, 1993; Murphy et al., 2004).

Research concurs on the commonly identified teaching behavior domains based on students' descriptions of teachers, the most frequently mentioned of which are teachers' personality traits (e.g., kindness, friendliness, fairness, patience) and teachers' instructional ability (e.g., García-Moya et al., 2020; Murphy et al., 2004; Plavšić & Diković, 2016). In this respect, it is important to emphasize that the behavioral aspects of teaching and teacher characteristics intersect. Other studies include teacher–student relationships as a third domain for students' perceptions of good teaching (Aksoy, 1998; Beishuizen et al., 2001; Ida, 2017; Sethi & Scales, 2020). Utilizing qualitative methods, Yu et al. (2018) explore student perceptions of teacher-student relationships and identify two overarching themes that contribute to a positive relationship: teacher noticing and teacher investment. Alkan (2013) elaborates on the categorization of teaching practices, adding three more domains: classroom management, teachers' interaction with parents, and teachers' appearance. A more recent study reveals four main domains that partly overlap with previous research: (1) relating to students as individuals, (2) creating relatedness within the classroom, (3) being responsive to students' academic needs, and (4) reducing school-related stress (Mælan et al., 2020). In a qualitative study that examines the development of student engagement, encouragement, emotional support, and empathy from teachers are emphasized by students (Pineda-Báez et al., 2019). However, although studies using student descriptions of teaching behavior provide rich information, to date they have been limited to a small sample size.

1.3. Topic modeling

The ability to investigate large amounts of data to uncover patterns, correlations, and insights is referred to as big data analytics (Russom, 2011). Researchers variously refer to the technique that extracts information from textual data to categorize and draw inferences as text analytics, text mining, or automated content analysis of a text. This technique was developed as a way to deal with large datasets that would not be possible to analyze using manual coding (Hopkins & King, 2010). Categorizing student responses into topics facilitates exploratory analysis of large texts by extracting the common themes report. Supervised learning methods and unsupervised learning methods are two general approaches to text analytic techniques (Berry et al., 2020).

We use an unsupervised learning method—namely, topic modeling—for extracting topics from written student responses for two reasons. First, in the absence of a training dataset (i.e., one in which the model is trained using these data), an unsupervised method is a natural choice; otherwise, training a classifier is a laborious task (Hujala et al., 2020). Second, using a pre-established training set may cause information loss or prevent the detection of new emerging topics (Hujala et al., 2020). Because written student responses are usually unstructured, informal, and situation-dependent, a possible drawback of the pre-established training set is a loss in data richness. On the other hand, unsupervised learning methods do not require researchers to pre-specify rules or keywords for the underlying themes; instead, they can discover patterns or topics through exploration of the data (Goloshchapova et al., 2019; Wachen, 2017).

Topic modeling is an unsupervised technique for determining the underlying latent topic or themes in large sets of text documents (i.e., data corpus)[1] (Kandula et al., 2011). The intention of topic modeling is a better understanding of a phenomenon in a written text by gathering the frequently co-occurring words to define a topic (Kandula et al., 2011; Ramage et al., 2009). Blei (2012, p. 77), one of the developers of the method, describes topic modeling as "statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time". These models identify patterns of word use in the data corpus. The central feature of topic modeling is that each document exhibits multiple topics, and every document in the corpus exhibits the topics in different proportions (Blei, 2012). In other words, documents have an estimated topic proportion (i.e., topic weight) for each of the topics in the model. In this study, the text document refers to students' open-ended answers, and each topic is assumed to define a single domain of teaching behavior.

Although several variants of topic modeling are available, latent Dirichlet allocation (Blei et al., 2003), one of the original topic models, is one of the most widely used approaches (Kherwa & Bansal, 2019; Vayansky & Kumar, 2020). It is based on the notion that

[1] Text documents refer to students' open-ended answers, and corpus refers to large sets of text data.

words belonging to a topic are more likely to appear in the same document. The application of topic modeling has increased in educational studies to obtain latent topics by statistical models for text analysis. This method has been used in three broad areas in the education field. First, scholars have used LDA for exploration purposes, such as trend evaluation or best prediction and recommendations. For example, recent studies have used LDA to compare course syllabi and curricula (Sekiya et al., 2015), explore the relationship between research and education (Lee et al., 2014), uncover insights into the learning analytics community by analyzing Twitter archives (B. Chen et al., 2015), investigate the public trend in education policy by focusing on teacher evaluation (Moretti et al., 2015), and explore research trends and patterns that can help researchers to obtain an overview of the field and systematic literature reviews (X. Chen et al., 2020, 2021; Lämsä et al., 2021; Liu et al., 2021; Parsons & Khuri, 2020; Wang et al., 2017). Second, researchers have employed LDA to analyze students' forum discussions in online courses. For example, studies have identified and clustered themes and patterns from students' massive open online course (MOOC) forum discussions (Ezen-Can & Boyer, 2013; Nanda et al., 2021; Ramesh et al., 2014; Reich et al., 2015). A recent study presents a Topic Analysis Instant Feedback System that can explore and visualize the topics according to the contents of an online discussion from learners by using LDA (C.-M. Chen et al., 2021). Third, LDA has been implemented to analyze pre-service teachers' and students' written texts (Y. Chen et al., 2016; Gibson & Kitto, 2015; S. Kim et al., 2017; Southavilay et al., 2013). For example, Matsukawa et al. (2019) apply LDA to analyze university course evaluation questionnaires. In a more recent study, student teachers' comments (peer feedback) on video recordings during their teaching are analyzed by using LDA topic modeling to examine novice teachers' development (Bent et al., 2021).

Building on a line of research introduced by Finch et al. (2018), the research field is not only limited to applying LDA analysis for evaluating large data sets on teacher evaluations, but also supports an approach that combines LDA with further analysis, such as thematic analysis, or sentiment analysis. For example, the LDA topic modeling approach is combined with network modeling and visualization interfaces to explore university teachers' self-assessments (Buenano-Fernandez et al., 2020). Hujala et al. (2020) demonstrate the process for analyzing masses of university students' responses to the open-ended feedback questions by using topic modeling. They also present a method of validating the outcomes by using thematic analysis to assign a theme for each topic, and statistical analysis to assess the relationships between the outcomes of topic models and Likert-scale questions. Grönberg et al. (2021) develop a tool (Palaute) for analyzing written feedback of university students to a study program by using topic modeling and emotion analysis. Similarly, LDA topic modeling analysis is associated with sentiment analysis to identify topics that are relevant for students and to identify polarity in students' opinions (Gottipati et al., 2018; Vargas-Calderón et al., 2020).

1.4. Research aims

This study presents an approach to analyze the masses of student responses to an open-ended questionnaire about their teachers. In particular, we apply LDA to extract hidden information from the textual data of a survey through topic modeling. Our aim is not to provide a new method for an individual teacher to extract topics from student responses. We present a meso- and macro-level approach for institutes, study programs, or nationwide evaluations to explore a large amount of written-text data. Eventually, feeding back these topics to teachers could help teachers understand more about the students' learning process, needs, and preferences on learning. Briefly, we aim to explore contemporary student views of their teachers' behavior by applying a cost-effective investigation of student perceptions based on an open-ended questionnaire from large-scale survey.

To our knowledge, this is the first study employing the novel approach using large-scale student evaluations in secondary education. Student assessments of teacher performance can serve different purposes in higher and secondary education. Besides, a novel aspect of this study is the comparison of data-driven topic modeling analysis with theory-driven manual analysis. By applying both analyses, we examine the convergent validity of the outcomes. This comparison is important for two reasons. First, manual analysis can help us evaluate the machine learning topic validity. Second, it allows us to identify the extent to which the ICALT framework for effective teaching behavior, which was developed based on the synthesis of evidence-based teaching effectiveness research, is in line with data-driven student perceptions. Thus, we aim to explore what we can learn from the masses of student responses to an open-ended questionnaire about teaching behavior. The main research question is: How can masses of responses of students to an open-ended questionnaire about teachers be analyzed to provide teaching behavior domains? The main research question is divided further into sub-questions, which are listed as follows:

- What are the main topics identified in secondary school students' perceptions of teachers' effective teaching behavior based on the LDA topic modeling approach?
- To what extent do the topics identified based on LDA topic modeling approach converge with the theory-driven approach?

2. Sample and measure

We drew data from the longitudinal national induction (2014–2019) project collected from secondary education in the Netherlands at six successive time points. Schools were regionally recruited by Teacher Education departments of nine universities. School-based educators of the interested schools recruited beginning teachers and the students of the beginning teachers who voluntarily completed the questionnaire on teaching behavior at 4 partly overlapping timepoints in 3 waves, forming 6 timepoints in total. Students remained anonymous, so student responses over time could not be linked, yet we could track the teacher's professional growth based on student evaluations of teaching behavior over time. For this study, however, this is not necessary because our focus is on student perceptions. For this purpose, we only used the open-ended answers of the students regardless of the longitudinal nature of the data because the focus was on student perceptions rather than teachers' professional development.

3554 teachers and their 189,632 students from 12 provinces participated in the study. More than half of the teachers ($N = 2,081$, 58.6%) were female, and 65.4% of the teachers had less than two years of teaching experience. Of the students, 59,386 (31.3%) were female, and 74,648 (39.4%) students did not report their gender. The distribution by teaching subject was as follows: 62,928 students (33.2%) for Alpha subjects (language track; i.e., English, Dutch, and French); 52,408 students (27.6%) for Beta subjects (science track; i.e., chemistry, mathematics, and biology); 44,727 students (23.6%) for Gamma subjects (social science track; i.e., history, economy, and civics); 9162 students (4.8%) for Physical education, and 9729 students (5.1%) for Artistic subjects (i.e., drawing, drama, and music). Students were distributed by grades as follows: 57,626 (30.4%) in grade 1 (aged 12–13 years); 52,145 (27.5%) in grade 2 (aged 13–14 years); 43,704 (23%) in grade 3 (aged 14–15 years); 24,685 (13%) in grade 4 (aged 15–16 years); 8.322 (4.4%) in grade 5 (aged 16–17 years); and 1458 (0.8%) in grade 6 (aged 17–18 years). Students were distributed by education track as follows: 62,002 (32.7%) in lower secondary vocational education (VMBO); 56,215 (29.6%) in senior general secondary education (HAVO); and 50,212 (26.5%) in academically oriented pre-university education (VWO).

Students filled out the questionnaire on teaching behavior voluntarily. Data were collected through several means: 74.4% of the questionnaires were collected via Qualtrics survey software, 13.2% were collected via a web application, and 12.4% were collected using paper and pencil. The data contained general demographic information about students (e.g., age, gender, education year), teachers (e.g., gender, teaching subject), and schools. The main open-ended question was "*What the teacher does well is …*". This open-ended question allowed students to communicate their needs and interests at that moment in time by reflecting on what they perceived their teachers did well. Studies of what constitutes an "expert" or a "best" teacher are often motivated by the argument that if more is known about what teachers do well in classrooms and how they do these things, a better position can be created to prepare pre-service teachers, in-service teacher development, and curriculum reform (Mullock, 2003). The open-ended question is therefore exploratory, aimed to unravel contemporary student perceptions of teaching behaviors that contribute to developments in education.

3. Analyses

Based on the whole dataset, we followed two paths: *3.1. Machine-based analyses* and *3.2. Manual analyses*. Fig. 1 represents the data analysis processes that include preparation, analysis, validation, and outcome comparison steps of these two paths.

3.1. Machine-based analyses

Machine-based analyses include preparation, analysis and validation steps as seen in the first path of the data analysis process in Fig. 1.

3.1.1. Pre-processing

Text mining techniques require a series of pre-processing steps to prepare the data for modeling analysis. The first step was to exclude any irrelevant answers that could affect the quality of the modeling analysis. This step involves removing emojis and punctuation, converting all words into lowercase, and excluding blank answers and answers corresponding to a blank (e.g., "*geen opmerkingen* (no comments)"). The second step was to remove Dutch stopwords, or language-specific words that neither add meaning nor contribute to the interpretation of the topics (in English, stopword lists include, e.g., "the", "and", "this"). A large portion of the student responses was in Dutch while a negligible proportion of responses was in English and Spanish. The third step involved the stemming process to equate similar terms, in which words were converted to their root form using an open-source Python library.[2] For example, we converted the words "argued", "argues", and "arguing" to the stem "argue". Last, given the specific nature of the Dutch language, an additional data cleaning step was necessary: combining compound verbs (e.g., converting "*les geven*" to "*lesgeven* (to teach)", "*samen werken*" to "*samenwerken* (to cooperate)"). These pre-processing steps are necessary not only to develop a dataset that is clean and formatted properly before conducting topic modeling but also to ease interpretability of the topic modeling output. After all the pre-processing steps, 173,858 student answers were left in the data. We performed the analysis in the Python programming language.

3.1.2. Topic modeling and selection of number of topics

We used LDA with a bag-of-words to group students' answers into topics according to co-occurring words (Blei et al., 2003; Griffiths & Steyvers, 2004). Determining the appropriate number of topics can be difficult (Grimmer & Stewart, 2013): too few topics in the model may result in overlapping terms and failure to identify the distinct topics, but too many topics may reduce semantic validity or fail to uncover the broader and overarching topic (Y. Chen et al., 2016; Nowlin, 2016). It is important to emphasize that no "optimal" number of topics (K) would work for every application. Instead, determining the appropriate number depends on the research question and the iterative process of varying K considering the parameters (Nowlin, 2016; Wachen, 2017).

Therefore, to determine the number of topics, we used automatic hyperparameter optimization for the varying numbers of topics K as 4, 5, 8, 10, 12, and 15. As a common way to evaluate the quality of a topic, we measured each topic's coherence to assess whether the words in a topic are in fact related to each other. Specifically, we computed each topic's quality using a topic coherence metric based on Normalized Pointwise Mutual Information (NPMI), one of the metrics most strongly correlated with human raters (Lau et al., 2014; Röder et al., 2015).

[2] https://spacy.io/models/nl.

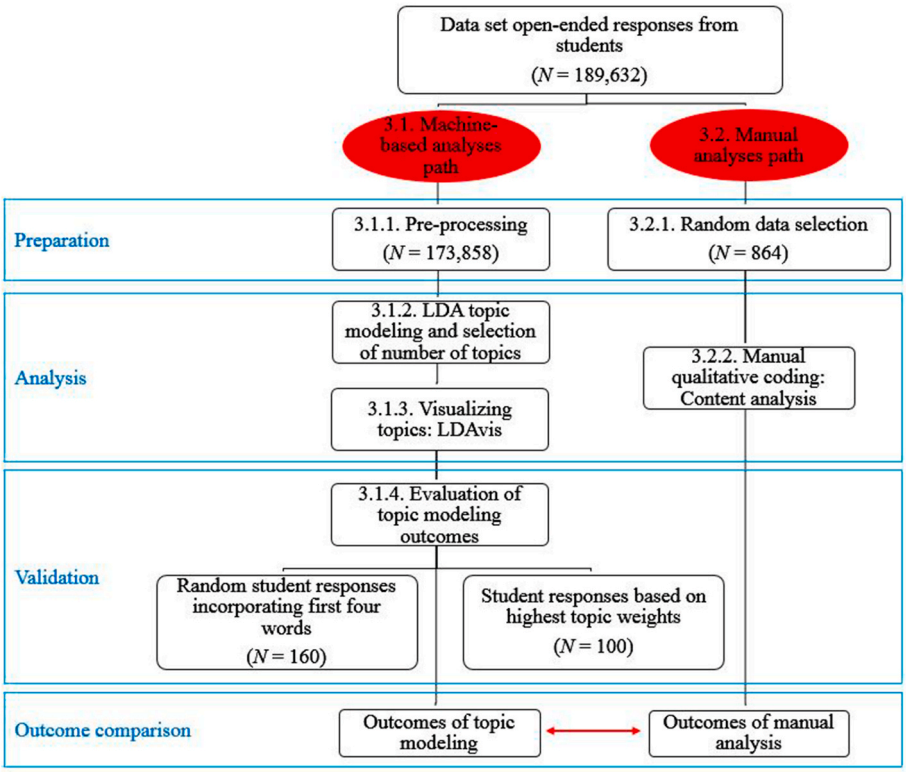

Fig. 1. The data analysis processes for student responses to the open-ended question.

3.1.3. *Visualizing topics: LDAvis*

For the interpretation and labeling process, we used an interactive web-based visualization tool of the LDA topic modeling results (LDAvis[3]; Sievert & Shirley, 2014). This visualization tool eases interpretation and labeling in several ways. First, it graphs the most frequent words in each topic (red bars in Fig. 2A), which indicates which words are highly relevant to the specific topic (Chuang et al., 2012). The top 20 words are considered to interpret and label the theme for each topic. On the right-hand side pane of the visualization tool, the width of the blue bar represents the frequencies of each word in all the responses, and the width of the red bar indicates the frequencies within the topic. A weight parameter λ (where $0 \leq \lambda \leq 1$) can also be used to determine the "relevance" of the word to the topic (Sievert & Shirley, 2014). Setting $\lambda = 1$ identifies the most probable words under each topic, while setting $\lambda = 0$ ranks words solely unique to the specific topic. To illustrate, topic words displayed in Fig. 2A are acquired using $\lambda = 1$, while topic words shown in Fig. 2B resulted from using $\lambda = 0.5$. For example, the bar for words such as "*helpen* (to help)" and "*houden* (to keep)" is fully red, with no blue bar showing (in Fig. 2A), which means that these words are represented exclusively in Topic 2 and thus are highly representative of Topic 2. When using $\lambda = 0.5$ in Fig. 2B, these two terms are the first and second most highly relevant terms representing Topic 2. However, the blue bar is wider than the red bar for the word "*goed* (good)" (in Fig. 2A), which means that this word is not only represented in Topic 2. When using $\lambda = 0.5$ in Fig. 2B, this word shifts into the lower order. Again, note that determining an "optimal" value of λ for topic interpretation is difficult (Sievert & Shirley, 2014). Herein, we adjusted λ to benefit from the additive assistance of this value for topic interpretations and labeling, rather than to determine the optimal λ.

In addition, the visualization tool enables the representation of relationships between topics. The visualization tool depicts each topic as a bubble, where the area of the bubble resembles the prevalence of the topic on the left-hand side of Fig. 2. Inter-topic distances are displayed through multidimensional scaling onto two axes, such that the relative distance and overlap between topics are visualized. This relative distance between topics indicates the semantic distance of topics. The figure also identifies the topic on which each word appears. For example, the word "*duidelijk* (clearly)" is the first word in Topic 3 (Fig. 3A). When this first word is clicked, the left-hand pane shows other topics that include "clearly" in their topic words (Fig. 3B) (e.g., Topic 1 also includes the word "clearly" in topic words).

By using this tool, the most representative 20 topic words are visualized. Each of these computer-generated word clusterings was manually reviewed by four researchers independently to interpret and label each topic. Human-based interpretation and labeling is a common practice although some studies investigate ways for automatic generation of labels (e.g., Lau et al., 2011). However, because data in social sciences are socially constructed, automatic generated labels can lose meaning and value when they are taken out of

[3] https://github.com/bmabey/pyLDAvis.

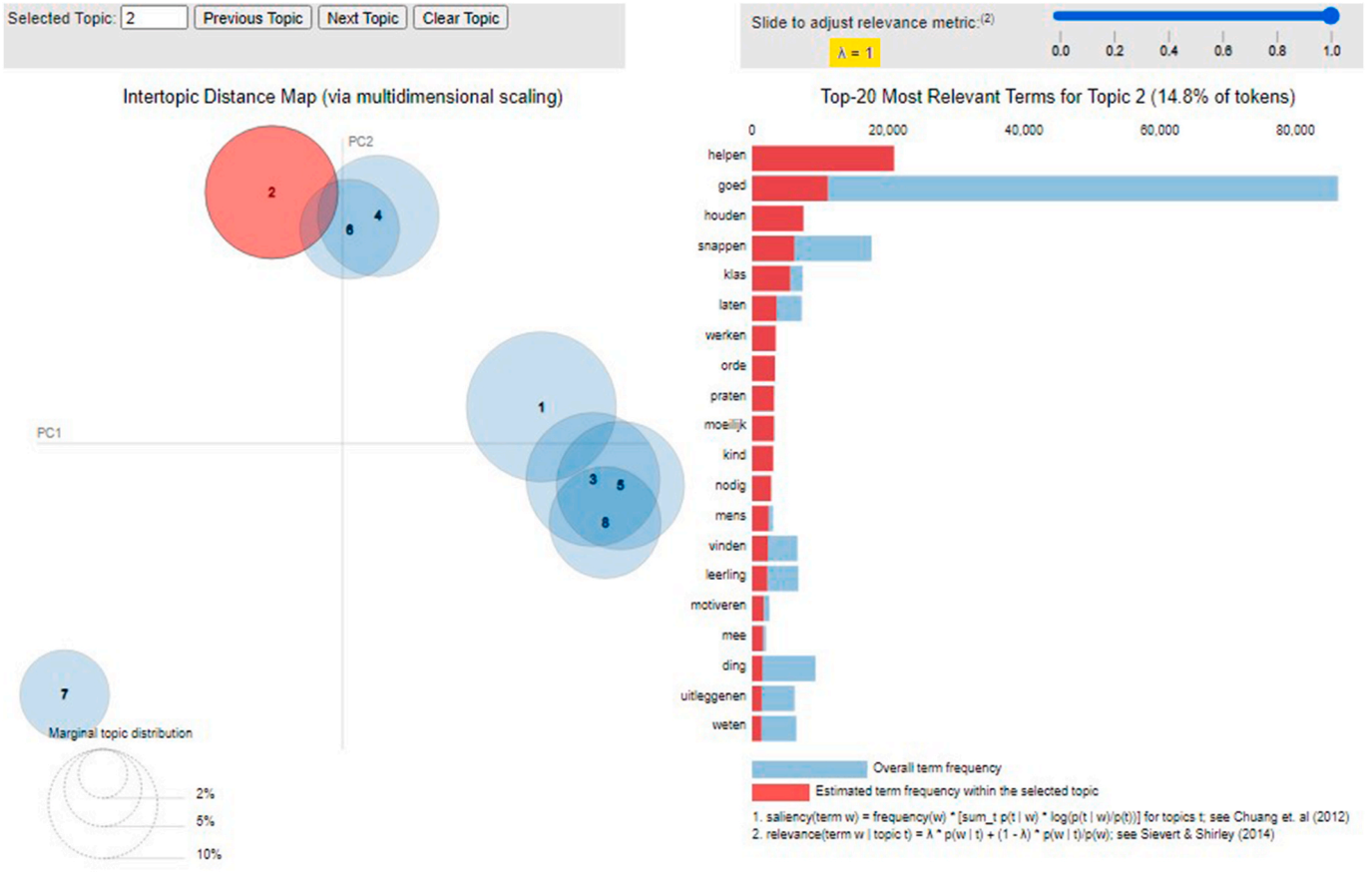

Fig. 2. Visualizing the LDA results of Topic 2 using $\lambda = 1$ and $\lambda = 0.5$, respectively. Note. We generated these visualizations using the LDAvis tool. Table B in the Appendices presents the translated words in English.

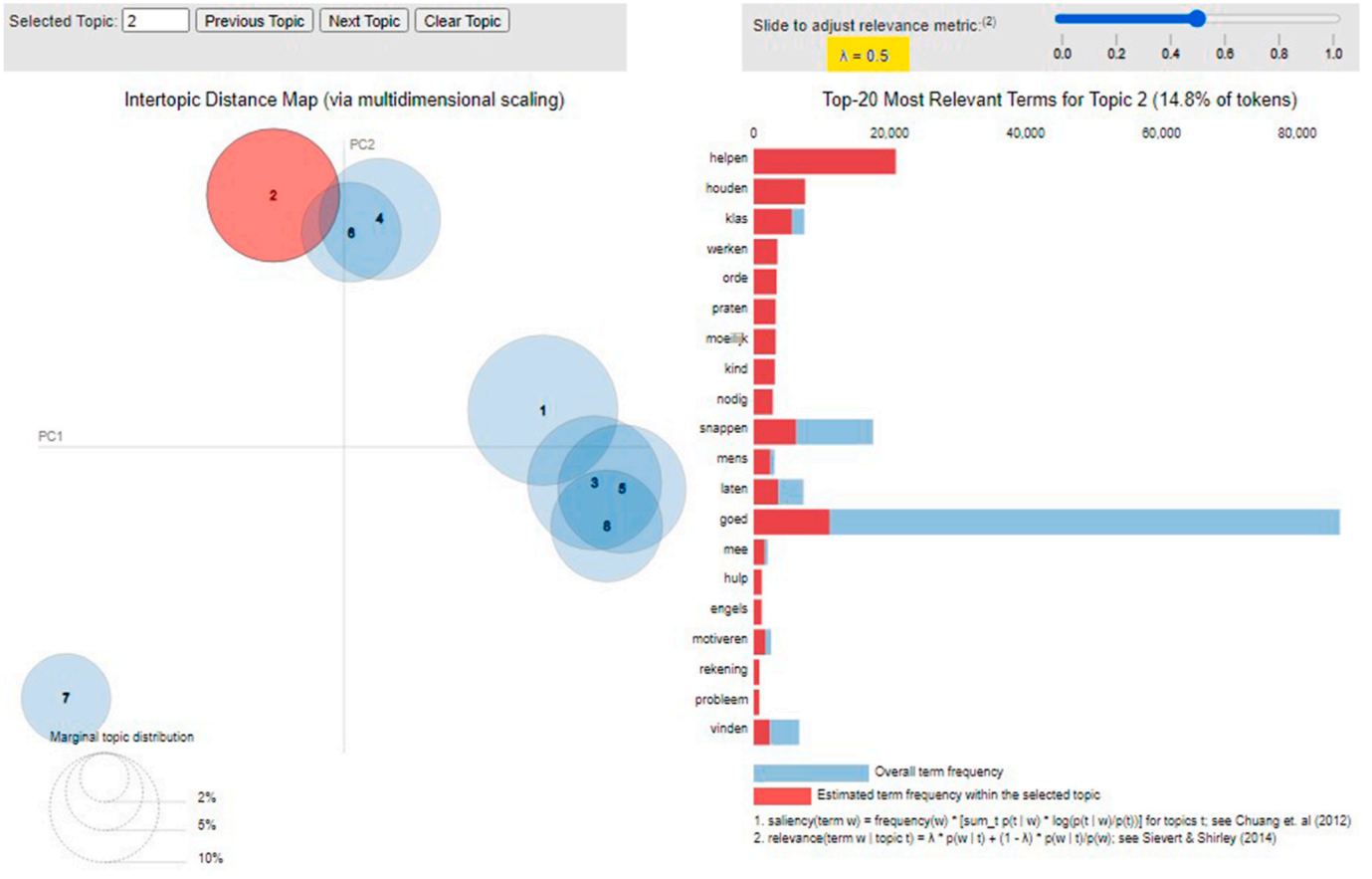

Fig. 2. (*continued*).

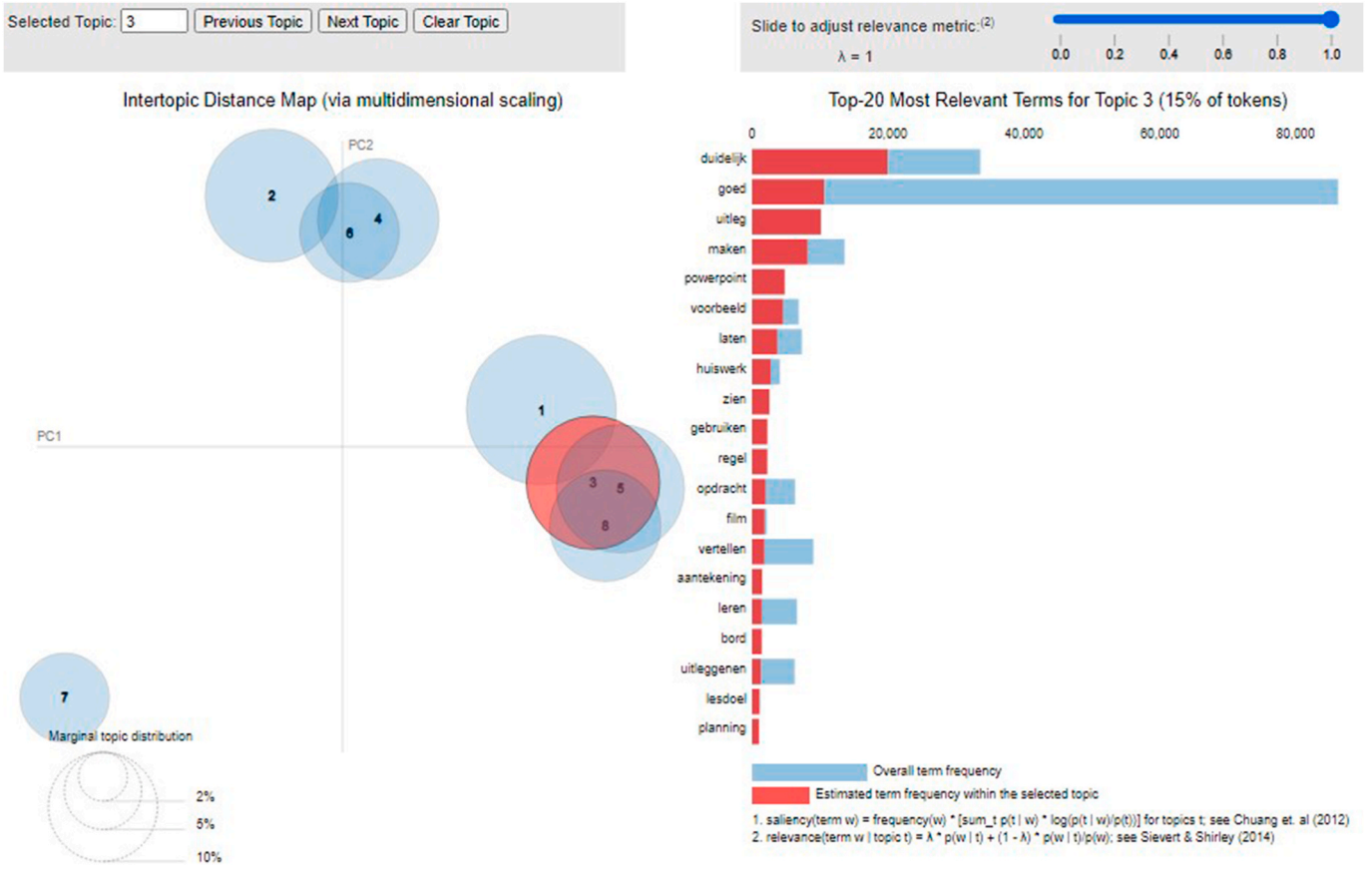

Fig. 3. Visualizing the LDA results of Topic 3 using $\lambda = 1$. Note. We generated the visualizations using the LDAvis tool. Table B in the Appendices presents the translated words in English. Fig. 3B represents the visualization when the word "duidelijk (clearly)" is clicked.

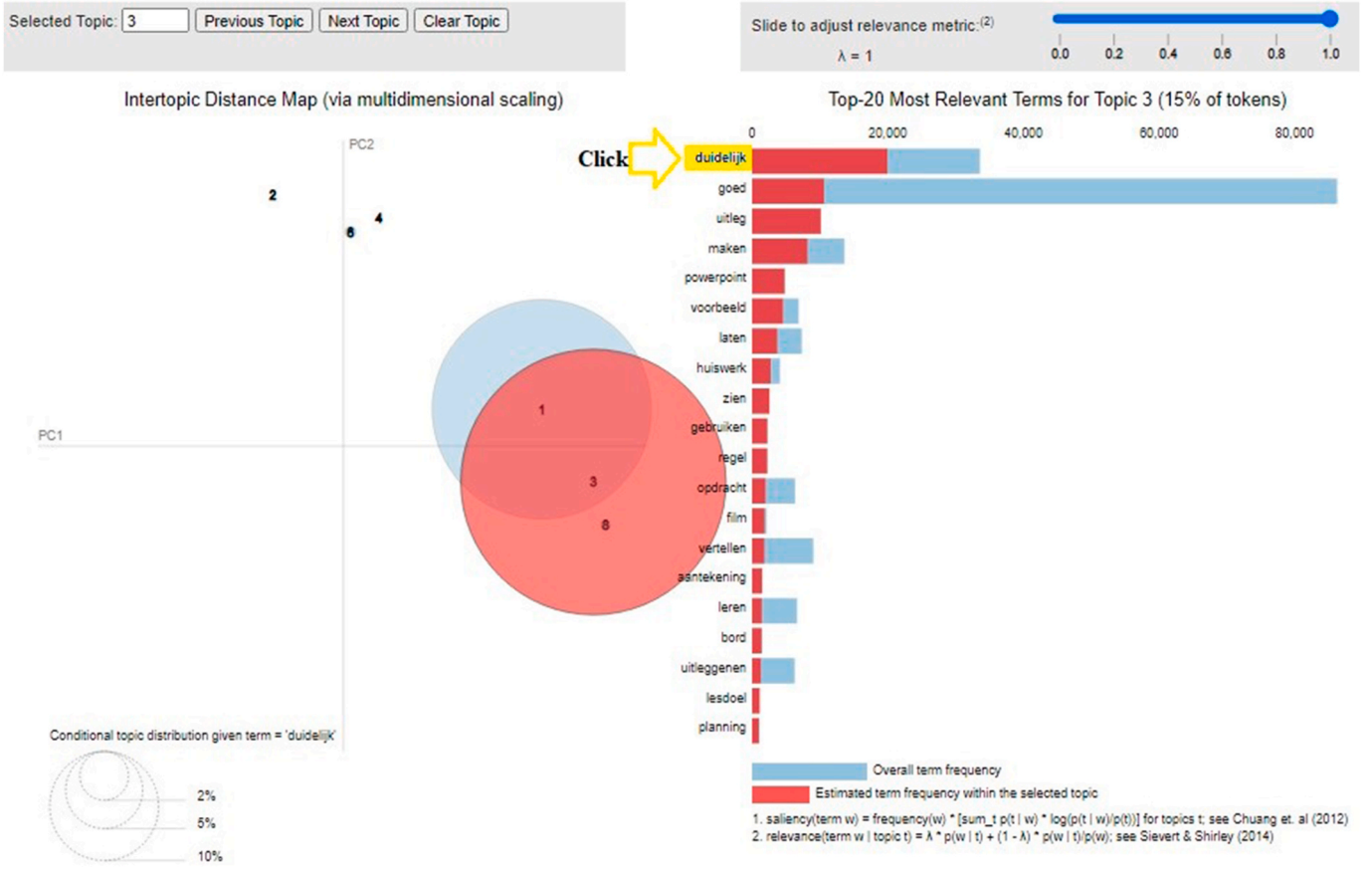

Fig. 3. (*continued*).

Table 1
Domains, sub-domains, and frequencies.

Domains	Sub-domains	Domain frequency (%)	Sub-domain frequency (%)
Safe and stimulating learning climate	Teacher–student (interpersonal) relationship and mutual respect	51 (4.03)	51 (4.03)
Classroom organization	Learning environment and communication	152 (12.00)	18 (1.42)
	Classroom management		32 (2.53)
	Learning time/Management of time		12 (0.95)
	Giving well-structured lessons		90 (7.10)
Clear instruction	Clear explanations and examples	427 (33.70)	358 (28.26)
	Stating the lesson objectives		69 (5.45)
Activating teaching	Behavioral engagement	296 (23.36)	50 (3.95)
	Emotional engagement/Evoking interest/Activating		144 (11.37)
	Developing independent learning skills		29 (2.29)
	Cooperation		9 (0.71)
	Teaching material and lesson variety		64 (5.05)
Differentiation	Addressing all students' levels, learning preferences, and learning profiles	131 (10.34)	15 (1.18)
	Providing help		116 (9.16)
Teaching learning strategies	Explaining how to learn	32 (2.53)	10 (0.79)
	Metacognitive strategies/scaffolding		22 (1.74)
Formative assessment	Monitoring learning	42 (3.31)	15 (1.18)
	Test		14 (1.10)
	Homework		13 (1.03)
Personality characteristics	–	80 (6.31)	80 (6.31)
Everything is good	–	56 (4.42)	56 (4.42)
		Total	1267 (100)

context (Boyd & Crawford, 2012). Because researchers must carefully unpack the meaning behind the topic and label the topic in a given context, in this analysis attention is given to labeling topics that are coherent and reasonable. The aim of labeling is to summarize the content of each topic as well as highlight the distinctiveness of topics from each other. We used labels to determine how the topics are similar to teaching behaviors in the existing theories. The researchers followed three steps for topic labeling: (1) describe each word cluster to reveal generic features, (2) assign a descriptive label that reflects the meaning of the topic in a comprehensive way, (3) after interpreting independently, discuss inconsistent labels and interpretations with each other to reach consensus. As a result, a reconciled set of labels was decided on. With regards to the second step, the topics were labeled by analyzing the most representative words. The labeling and description of each topic were based on the semantics of the words. During interpretation, the incorporation all the words was considered giving the most representative words more weight. For example, Topic 1 was labeled as *clear explanation* based on the most representative topic words "to explain", "good", "clearly", and "teaching". Note that the survey question was general, and thus students might have tended to write the first thing that came to mind, regardless of specific teaching behavior. Therefore, we adopted the bigger picture in the topics with a holistic approach, though each specific word could be connected with a specific teaching

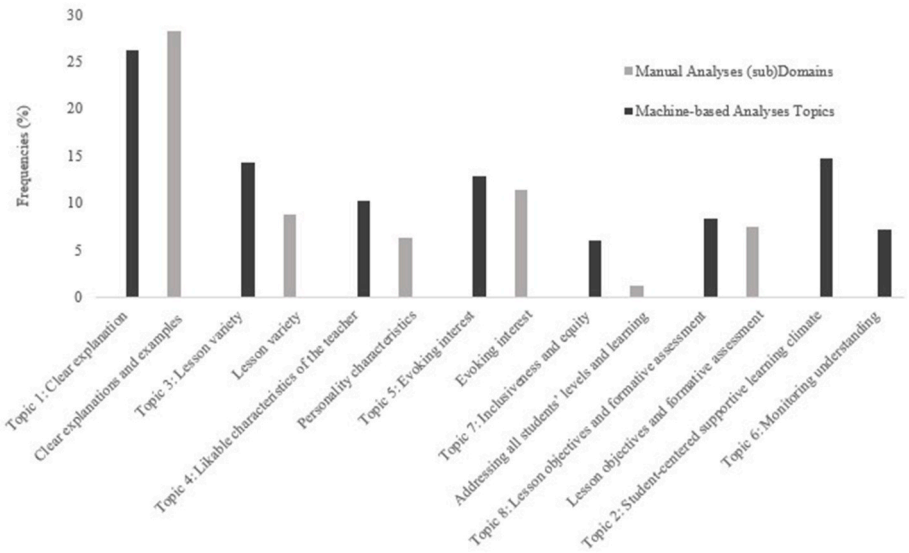

Fig. 4. Frequencies of sub-domains and topics.

behavior domain.

3.1.4. Evaluation of topic modeling outcomes: random student responses

Without linking the results of text analysis to the extant literature, school administrators and educators cannot have confidence that the topics identified have construct (Campbell & Cook, 1979) and interpretive (Maxwell, 1992) validity. In other words, there is no guarantee that the generated topics are relevant to the context because LDA topic modeling is just a probabilistic model in which each topic is defined by a specific set of words that often appear together (Hujala et al., 2020). Thus, it is the researcher's responsibility to evaluate the topics' validity and assign meaning to the content (Hagen, 2018).

To this end, we evaluated the topic interpretations and labeling using three approaches. First, we randomly selected 20 student responses for each topic (in total 160 student responses) to compare the topic content with the topic interpretation. The selection criterion was to incorporate the first four topic words for each topic. The responses that incorporate the first four topic words were manually examined whether the responses corresponded to the topic interpretation and labeling. The topic weights[4] of each response were reviewed. For example, one of the student answers is *"If you have a question, he [the teacher] will explain it well and he will also ask you if you have understood it subsequently. If you still don't get it then he just explains it again",* which has a topic weight of 0.69 for Topic 6 (*Monitoring understanding)*, and incorporates the first four topic words "understanding", "question", "good", and "ask". For each topic, we deemed randomly selected student responses with the highest topic weight as representative examples of the topic (see Table B in the Appendices for the examples of student responses and the topic weights). Unsurprisingly, for the long answers, even if the student response includes the first four words, we observed an equal distribution of topic weights. In other words, including the first four words does not mean that the weight of that topic will be high, and thus might create a new construct to be considered for the topic interpretation. Nevertheless, we observed consistency between the examples of student responses and topic interpretation and labels after slight adjustments. We, therefore, consider this approach stricter than looking at the responses with the highest topic weights of each topic, which was also used as a second approach.

As a second approach, we evaluated topics based on the highest topic weights as described by Nanda et al. (2021). For each topic, the student responses were sorted in order of decreasing weight associated with the topic. We manually checked the 100 student responses that had the highest topic weight associated with that topic. Because these responses had the highest topic weight, these responses were expected to be composed mainly of a single topic. A high level of alignment was detected which indicates that the topic interpretation and labeling were coherent. For example, the student response with the highest topic weight (0.93) for Topic 5 is "make the lesson interesting. She [the teacher] makes a lot of jokes and sometimes tells a nice story, so it is also nice to pay attention in class". The 100th student response when the Topic 5 was ordered by decreasing order of topic weight (0.89) is "he [the teacher] explains well and tells nice creative stories in a fun way". These responses were in alignment with the topic label, which is *Evoking interest.* This alignment makes sense because when the topics are ordered by weight, the student responses mainly compose that part (with the most weight), which clearly is reflected in the topic label. Third, we performed a qualitative evaluation that resembles the human coding process suggested by Hagen (2018) to reveal theory-driven teaching behavior domains, as described in the following section.

3.2. Manual analyses

Manual analyses include preparation and analysis steps as seen in the second path of the data analysis process in Fig. 1.

3.2.1. Random data selection

We randomly selected 864 student responses from the entire dataset to examine the alignment of discovered topics with the students' written responses. To control for the distribution of demographic characteristics, we took several criteria into account during random selection: (1) we selected five classes per education year of VMBO, two classes per education year of HAVO, and one class per education year of VWO; (2) we selected two male and three female teachers per education year of VMBO, one male and one female teacher per education year of HAVO, and three male and three female teachers from VWO; and (3) we tried to keep the gender of the students in each class equal (class size ranged from 16 to 31).

3.2.2. Manual qualitative coding: content analysis

We used mixed methods for the analysis of the student responses to the open-ended question. We followed the standards of content analysis (Elo & Kyngäs, 2008; Mayring, 2010) to explore what and how students respond to the open-ended question. The purpose is to generate a categorical coding system that groups together the given responses. Two coders conducted the qualitative content analyses following the subsequent steps: (1) They familiarized themselves with the data and started coding considering the six effective teaching behavior domains (van de Grift, 2007). (2) They created a codebook based on the dataset and discussed the codebook with an external researcher. (3) Based on an iterative coding process, the codebook was subjected to revision by changing and discussing the emerging codes as the successive responses brought in new ideas. In total, 84% of the data was coded by both coders independently. After the coding phase, to ensure the accuracy of the coding procedure, we enlisted two independent expert researchers to revise the codebook

[4] A topic weight refers to the proportion of this topic in a student response. In other words, it indicates how frequently the topic was present in a student response. The sum of all topic weights for a student response is one. Therefore, student responses composed of multiple topics are expected to be assigned smaller weights for multiple topics, and student responses composed of a single topic are expected to have a high weight associated for that topic.

and discrepancies among codes. In the case of incongruent coding, the researchers discussed the codes and revised discrepancies until reaching mutual consent. The codebook was inspected to determine the expressed similarities and how to map onto teaching behaviors in the existing theories. Thus, we grouped codes into larger themes, which created teaching behavior domains and sub-domains.

4. Results

4.1. Topic modeling, topic interpretations, and topic clusters

The results show that the model with eight topics provides the relatively highest coherence score (0.237), where the lowest score was −0.778. Thus, we identified the topic model with the best hyperparameters, which has eight topics, and conducted the interpretation and labeling process. Overall, the most salient words are "to explain", "lesson", "to help", "clearly", "very", "understanding, "fun", "everybody", "to care", "explanation", "to teach", "to go", "to make", "explain", "to hold", "pupil", "question", "to prepare", "nice", and "very" (see the Appendices Table A for the original Dutch words).

LDAvis reveals that the eight topics create three clusters. These results depict topics 2, 4, and 6 and topics 1, 3, 5, and 8 as relatively close to each other, while topic 7 is located relatively far away from the other topics. Table B in the Appendices summaries eight topics with the 20 most frequent words, the marginal distribution of the topics, suggested topic labels, descriptions of each topic, and examples of student responses.

4.2. Manual qualitative coding: content analysis

The content analysis resulted in 87 codes, which we grouped into nine domains and 19 sub-domains (Table 1). Of the nine domains, six domains are in line with the ICALT framework for effective teaching behavior domains (van de Grift, 2007), and the other three are unique. These unique domains emerged during the iterative process of manual coding: the first concerns test- and homework-related practices and monitoring learning, the second is related to the teacher's personality characteristics, and the last contains answers that do not provide extra information about teaching behavior. Frequency rates showed that respondents mentioned clear instruction most frequently, with the highest frequency percentage occurring in the clear explanations and examples sub-domain, followed by evoking interest. By contrast, teaching learning strategies has the lowest frequency rate.

4.3. Comparison of machine-based analyses and manual analyses

We observed an overlap between topics and (sub-)domains. The semantic overlap can also be observed in the frequency distribution of topics and sub-domains (see Fig. 4 for frequency comparisons). Topic 1, Clear explanation, overlaps with the sub-domain of clear explanations and examples, and both have a high frequency rate. Two sub-domains of activating learning, Lesson variety and Evoking interest, correspond to Topics 3 and 5, respectively. Topic 4, Likable characteristics of the teacher, encompasses the unique domain of personality characteristics. Topic 7, *Inclusiveness and equity*, corresponds to the sub-domain of differentiation, which involves addressing all students' levels, learning preferences, and learning profiles. Topic 8, *Lesson objectives and formative assessment*, corresponds to the formative assessment domain and the sub-domain of stating the lesson objectives. Topic 2, *Student-centered supportive learning climate*, and Topic 6, *Monitoring understanding*, are unique topics that do not hold commonality with manual coding.

5. Discussion

The big data that can be gleaned from students' responses to surveys provide rich information; thus, analyzing this abundant information is worthwhile. To answer the main research question, we apply a machine learning tool to the education field to explore students' responses eliciting their contemporary perceptions of teaching behavior. Using LDA topic modeling, we identified representative words that constitute teaching behavior domains hidden in students' responses. These representative words allowed us to make interpretations. We found that the extracted topics were interpretable, words in each topic could form a meaningful topic together, and the randomly generated examples for each topic corresponded to the relevant topic. With the help of topic modeling analysis, together with a visualization tool, it was possible and manageable to analyze students' open-ended answers meaningfully. Therefore, we argue that using a machine learning approach to analyze big data of student responses is applicable, coherent, and semantically meaningful.

Our machine-based analyses extracted eight data-driven topics that are relevant at a certain point in time. Manual coding analyses using a theory-driven approach revealed nine domains. We grouped the domains and sub-domains according to the effective teaching behavior model (i.e., ICALT; van de Grift, 2007). Comparing these analyses provides theoretical insights into student perception. Although they employ different approaches, the distribution of topics and the corresponding sub-domains showed overlapping frequency rates, adding insights to the validity of the extracted data-driven topics. We identified unique topics and unique domains. Unique topics refer to the topics that are not found by manual coding while overlapping topics refer to the topics that are also found by manual coding. Unique domains refer to teaching behavior domains that are found by manual coding and are not in line with the six teaching behavior domains in the ICALT framework. The unique and overlapping topics are discussed below.

5.1. Reflecting on the overlapping topics

Unsurprisingly, the first topic, *Clear explanation*, which corresponds to the teacher's instructional clarity, constitutes the largest part of the marginal topic distribution. The topic coincides with a sub-domain in manual analysis that also constitutes the highest frequency among other sub-domains. This finding is also consistent with the existing framework on effective teaching behavior domains. Research originating from constructivism, behaviorism, and direct instruction paradigms shows that clear explanation is related to students' learning outcomes (Maulana et al., 2015b; Maulana et al., 2017). Instructions are highly valued if they are direct or explicit (Maulana et al., 2015b), and students do not learn as well if instructions are unclear (Maulana et al., 2017). Given that in this study, 69.1% of the teachers had less than two years of experience, and given the focus of instruction clarity in teacher education curriculum, it is not surprising that this topic constitutes the largest part of the marginal topic distribution.

Topic 3, *Lesson variety*, emphasizes the use of various supportive illustrations and teaching materials. Words in Topic 3 such as "PowerPoint" and "movie" illustrate that the teacher supports clear explanations by making them more visible and tangible. These words overlap with codes in the corresponding sub-domain, such as "provide a lot of variety" and "visualization (pictures in PowerPoint), movies". In the literature, while some frameworks incorporate this teaching domain under the domain of activating learning (e.g., van de Grift, 2007), others consider it a distinct domain to emphasize the significance of the quality of subject-matter representation (Bell et al., 2019). Concordantly, technological skills (e.g., "make great PowerPoints for us [students] to learn", "play with Smart Board") were common responses from students asked to identify good teaching characteristics (Bullock, 2015, p. 9). Indeed, in practice, almost every classroom has multiple visual aids and eye-catching cues that adorn the walls to increase the richness of the representations. Visual aids, together with technology-based tools such as PowerPoint, movies, and interactive whiteboards, allow teachers to work through ideas together with students, use these visual resources to explore and make sense of ideas, actively engage students to detect and correct misconceptions, break down concepts into chunks, and reach a common understanding (Reedy, 2008). Moreover, considering visual and audio materials' ability to hold students' attention, it is unsurprising that students value lessons that are varied in visual and audio materials (Bullock, 2015).

Topic 4, *Likable characteristics of the teacher*, refers to positive affective teacher characteristics. Students often mentioned characteristics that make someone likable and socially attractive, such as being kind and friendly. Words such as "good", "nice", "cheerful", and "calm" were clustered in this topic together with words referring to the teacher. In the manual analysis, this topic coincides with the unique domain of personality characteristics. The codes in the corresponding domain, such as "nice", "stay patient/calm", and "friendly", explain the comparable frequency rates. This domain is unique; it only became apparent during the iterative process of the manual analysis and thus represents a novel domain not present in the six teaching behavior domains. This domain is in line with research that incorporates students' voices (e.g., Mælan et al., 2020). By contrast, teacher effectiveness literature mostly defines teaching behavior as observable behavioral aspects of teaching practices, which has led to a lack of focus on teacher characteristics such as personality traits and affective characteristics. Given that both the machine learning analysis and iterative process of manual analysis identified personality characteristics as important, we conclude that students notice not only behavioral aspects of teaching but also teachers' personality traits and affective characteristics. It is important to note that teaching behavior and teacher characteristics intersect.

Topic 5, Evoking interest, considers the teacher a performer/entertainer who provides a "fun", "interesting", and "fascinating" lesson and one who prepares the lesson well. This topic has a comparable frequency rate with the corresponding sub-domain in the manual analyses. The evoking interest sub-domain pertains to teaching behavior that facilitates students' active learning by attracting their attention, making the lesson interesting, involving them in the lesson, and motivating them (Maulana & Helms-Lorenz, 2016; Maulana et al., 2015a). The interaction among students and between teacher and student is related to the domain of activating teaching (Meeuwisse et al., 2010). Furthermore, we found that students value teachers who take an active role in getting attention rather than simply involving students in group work or interactions; in other words, the activation is teacher-initiated. This domain corresponds to creating relatedness within and beyond the classroom, which facilitates engagement and collaboration (Mælan et al., 2020).

Topic 7, *Inclusiveness and equity*, refers to teachers being respectful and inclusive to everyone while treating every student equally. Students mentioned that they value teachers who provide equal access to learning activities and supply the same opportunities to all students. This finding might seem to oppose the effective teaching behavior domain, which emphasizes the importance of differentiated instruction (Smale-Jacobse et al., 2019; van de Grift, 2007) because differentiation implies adapting the instruction to student characteristics to better meet the learning needs of diverse students (Tomlinson et al., 2003). However, from this topic, it is difficult to conclude whether students appreciate differentiated instruction that leads them to achieve equal learning outcomes or appreciate the same instruction to keep the equity in the classroom. Student perceptions might also depend on the extent to which teachers perceive diversity, inclusiveness, and equity and how teacher perceptions manifest in their teaching behavior in the classroom. How teachers consider diversity, equity, inclusiveness, and differentiation in their teaching behaviors might therefore be decisive for choosing to treat all students equally or differentiate instructions (Civitillo et al., 2016). We conclude that the relationship between equity and differentiation is complex and requires further investigation.

Topic 8, *Lesson objectives and formative assessment*, is related to teachers preparing students for tests and ensuring that goals are achieved: the teacher states the lesson objectives, explaining what students will do in the lecture, what they will learn, and what they need for the exam. Because this topic combines two sub-domains, it shows a different pattern than the effective teaching behavior domains. Although the sub-domain of stating lesson objectives is clustered under the domain of clear instruction, Topic 8 incorporates it with formative assessment, possibly because students often associate the purpose of their lesson and lesson objectives with exams. The literature on goal orientation distinguishes learning with a goal of understanding and mastery as intrinsic goal orientation and

learning with a goal of higher grades, rewards, or approval from others as extrinsic goal orientation (Pintrich & Schrauben, 1992). These goals indicate how students define success and failure, their affective reactions, and their subsequent behaviors (Farsani et al., 2014). Broadly defined, achievement goal orientation reflects the reasons and purposes for students to engage in achievement tasks (Sins et al., 2008). Indeed, understanding students' achievement goals can motivate teachers to develop suitable classroom practices to facilitate learning (Farsani et al., 2014).

5.2. Reflecting on the unique topics

Notably, LDA topic modeling generated unique topics that were not specifically detected by manual analysis, which may have occurred for three reasons. First, the uniqueness of the topic might indicate that the two analyses have different approaches. In the manual analysis, the student responses were precisely addressed such that each specific response was considered a specific code, which may have impeded the researchers' understanding of the holistic meaning behind the answer. LDA topic modeling addresses each response severally. But besides that, it handles each response to detect meaningful word combinations that form a new meaning. These word combinations might be difficult to detect with human-based coding. For example, the unique topics describe a general atmosphere in the classroom, which might be difficult to detect with the manual codes. It might become meaningful when relevant words form a topic that is derived from data, as conducted with the LDA topic modeling analyses. Second, owing to the universality and generalizability concerns of a theory-based approach, the manual analysis might have overlooked time- and context-dependent responses. Consequently, the unique topics might represent teaching behavior domains that are relevant at a certain point in time; for example, students with specific needs at a certain time might emphasize these teaching behavior domains. Third, although simple random sampling is intended to be an unbiased approach, sample selection bias could have occurred. The small sample of the larger dataset might not be representative enough, and the representation of the larger dataset may require additional sample techniques. If several randomly selected samples or larger samples could be selected for the manual analyses, complementary topics could be found. These unique topics require further evaluation to establish full validity.

5.3. Insights of the visualization tool outcomes: topic clusters

In addition, LDA topic modeling together with the visualization tool revealed that the eight topics created three clusters. The first cluster contains Topics 2, 4, and 6, which are Student-centered supportive learning climate, Likable characteristics of the teacher, and Monitoring understanding, respectively. The commonality in this cluster is that they represent the first-person point of view, in this case, the student-oriented viewpoint. This cluster depicts how students feel and experience the learning environment and the relationship with their teacher. For example, in Topic 2, the teacher supports and keeps an eye on students, helps them, and actively follows their development. In Topic 4, the affective characteristics of the teacher are emphasized, which means that students like and have an interpersonal bond with their teacher. In Topic 6, the focus is on whether students can understand the lesson. The second cluster contains Topics 1, 3, 5, and 8, which are Clear explanation, Lesson variety, Evoking interest, and Lesson objectives and formative assessment, respectively. This cluster shows a teacher-oriented viewpoint. In each topic, the teacher is required to initiate an action and be the leader of the actions. As a third cluster, Topic 7, Inclusiveness and equity, stands alone, which could be because this topic does not specifically concern particular individuals, neither the student nor the teacher, but instead encompasses everyone in the classroom.

5.4. Limitations

Although the machine-based analyses provide promising findings, we acknowledge several limitations and challenges of the study. A known challenge is determining the optimal number of topics, which makes the approach more exploratory and poses some degree of subjectivity. Using an iterative process of varying K with respect to semantic validity or relying on parameters as determinant factors may influence further topic modeling analyses. Another limitation is that, due to the nature of our data, the students' answers were relatively short, which means that words belonging to the same topic may co-occur less frequently than in longer texts. This limitation may have led to fewer or less meaningful topics (Tang et al., 2014). Furthermore, due to the nature of the data-driven topic modeling analysis, the extracted topics are sample-dependent. With a different group of students and teachers (e.g., student age groups, education tracks, subject areas), different topics might emerge, despite the large sample size. For example, student responses within different subject areas (i.e., Alpha subjects, Beta subjects, Gamma subjects) might have specific features, which might lead to unique topics. In future research, the analysis should be replicated in different cultural and educational contexts with different student and teacher subgroups and improved to accommodate multilingual open-ended feedback. We also applied the proposed process to analyze student perceptions of teaching behavior, which limits the scope of the results, but this is easily adaptable to other written text surveys. The process that we present in this study is novel and requires further evaluation to establish its full validity across various contexts and conditions. To this end, we also recommend evaluating the effectiveness aspect of the topics and comparing them with a large sample of Likert-scale data.

6. Conclusion

Spurred by the availability of digital texts, researchers have shown increasing interest in the automation of big data analysis. In the field of student responses in education, LDA topic modeling was used to analyze open-ended feedback from learners (e.g., Y. Chen

et al., 2016; Cunningham-Nelson et al., 2019; Grönberg et al., 2021; Hujala et al., 2020; Matsukawa et al., 2019) and online course evaluations and forums (Nanda et al., 2018, 2021; Unankard & Nadee, 2019). These studies have found the approach of determining the themes of topics generated by LDA using qualitative analysis, to be effective for analyzing open-ended survey responses.

Building on these findings, the current study contributes to the field in several ways. First, the current study applies LDA modeling to an underrepresented context (context-specificity). Previous studies used LDA topic modeling mainly in higher education contexts. Studies applying this modeling in secondary education are scarce. One study was conducted among secondary education students, developed a feedback system based on topic analysis and examined the effects of this system on learners' reasoning performance in online discussions (C.-M. Chen et al., 2021). Another study conducted in secondary education compared two types of linguistic analyses, one of which being LDA topic modeling, using students' answers to a constructed response test (S. Kim et al., 2017). In both studies, LDA topic modeling analysis was used as a methodological tool to examine its utility in understanding students' learning. In the current study, a distinct focus is presented. Students' perceptions of teaching behavior in secondary education were analyzed using LDA topic modeling to unravel contemporary student perceptions.

Second, the current study includes a large sample (at the national level). Compared to previous work with LDA (e.g., C.-M. Chen et al., 2021; Y. Chen et al., 2016; Grönberg et al., 2021; Hujala et al., 2020; S. Kim et al., 2017; Matsukawa et al., 2019), the sample size in this study is substantially larger. For example, in some recent studies using the LDA topic modeling analysis, a sample of 6087 university students' course evaluations was used (Grönberg et al., 2021; Hujala et al., 2020). The sample sizes in studies conducted with secondary education students are even smaller; e.g., 61 students from two classes in a secondary school, where two classes were divided into two groups (i.e., experimental and control group) to compare the feedback-assisted online discussion board with the general discussion board (C.-M. Chen et al., 2021)); and a sample of 252 students' answers for the LDA topic modeling analysis (S. Kim et al., 2017). The sample size is a particularly important issue (Tang et al., 2014) as it is very difficult to draw statistically significant conclusions from a small number of observations. A small number of students and/or teachers leads to less generalizable findings to other contexts. The robustness of the findings in our study was supported by the size of the dataset being examined, as well as by the comparison with small sample student responses and confirmed by manual coding results.

The LDA topic modeling has several advantages over traditional qualitative data analysis methods. First, LDA topic modeling can bring together word combinations that form a new meaning, which is not easily detected by human-based coding. This approach is systematic and repeatable, which enables fast evaluation of results. Second, LDA topic modeling uncovered unique general topics that would not specifically be considered in the theory-driven analysis. These new topics can provide an opportunity to monitor and track students' insights over time, considering the context- and time-dependent aspects of teaching behavior. For example, this approach can be implicated to examine the impact of the COVID-19 pandemic on student perceptions of teaching behavior (e.g., we could hypothesize that students would mention teaching behaviors related to online teaching more frequently due to the prevalence of remote education) (e.g., Aiyanyo et al., 2021). Although students' expectations and preferences are context- and time-dependent, it might also be difficult and time-consuming for theories to pursue and adapt quickly to changes. The chosen theoretical framework might be narrower than contemporary student perceptions of good teaching. This study has theoretical significance, showing a novel approach to detect contemporary teaching behaviors. To sum up, we suggest using LDA topic modeling to easily detect relevant and contemporary topics, such that trends over time can be projected and theories revised.

In addition to the aforementioned scientific importance, this study has important implications for practice as well. This method is of value to teachers because LDA topic modeling analysis allows for extracting meanings from written text without the burden of reading every line of evaluation. If the topics could be fed back to teachers, they could fine-tune their teaching toward contemporary student preferences to scaffold students' learning and outcomes and create a learning environment that matches students' preferences. The novel process demonstrates how to apply the proposed evaluation methods in practice. Yet our aim is not to provide a new method for an individual teacher to extract topics from student feedback given to him or her. Instead, we propose a comprehensive approach for institutions, study programs, or nationwide evaluations to summarize and categorize large amounts of data to bring new value for improvement to the community of teachers, school administrators, and educational leaders. For example, at the degree program level, analyzing large amounts of data on student perception may lead to a progressive deliberation on curricula to ensure student teachers understand and align their teaching behavior with students' learning preferences and needs. School administrators and continuous professional development providers may use the large-scale student data to provide input for teachers to redesign their lessons. Considering policy makers' higher expectations from researchers for information extracted via a large data-driven process, nationwide evaluations may produce indicators of good teaching quality based on student perceptions. In short, this study contributes to the extant literature by developing a streamlined, but rigorous, process for educational institutions to analyze masses of responses to open-ended feedback questions. Nevertheless, the integration of a machine-based assisted analysis should be aligned with a school's or institute's vision and culture (Wang, 2021). It is important for educational leaders to cooperate with teachers, administrators at schools, policymakers, and researchers with regard to whether and how to adopt machine-based analysis, develop machine-based training strategies, and work with machine-based software providers (Wang, 2021).

In conclusion, the automated methods used cannot completely replace human analysis and evaluation of written responses; therefore, human-centered computing is crucial (Wang, 2021). Our study suggests that the relation between machine learning and human analysis is complementary. That said, the meaningfulness and the quality of outcomes produced by machine learning are enhanced by integrating human analysis in concert. Our study also suggests that without the assistance of automated text analysis methods, the large-scale analysis of written text responses is an extremely time-consuming and labor-demanding task to conduct. Employing human analysis solely to the large-scale data analysis is highly impractical. Thus, the automated methods are valuable in assisting in the analysis of students' written text responses. This study is an example of human-computer interaction.

Credit author statement

Bilge Gencoglu: Conceptualization, Project administration, Writing – original draft **Michelle Helms-Lorenz:** Conceptualization, Supervision, Writing – review & editing **Ridwan Maulana:** Conceptualization, Supervision, Writing – review & editing **Ellen P.W.A. Jansen:** Conceptualization, Supervision, Writing – review & editing **Oguzhan Gencoglu:** Conceptualization, Software, Writing – review & editing

Funding source

This project was financed by the Dutch Ministry of Education, Culture and Science: OND/ODB-2013/45916 (project number is 804A0-45835).

Declaration of competing interest

None.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors thank student assistants for helping with conducting the manual analyses and appreciate Peter Moorer for his support with the management of data.

Appendices.

Table A
The eight extracted topics based on topic modeling

Topics	Words in original (Dutch)
Topic 1	uitleggen, goed, duidelijk, lesgeven, leggen, ding, stof, voorbeeld, lesstof, begrijpen, opdracht, grammatica, tekenen, nieuw, leerstof, uitbreiden, rustig, straf, mevrouw, enthousiast
Topic 2	helpen, goed, houden, snappen, klas, laten, werken, orde, praten, moeilijk, kind, nodig, mens, vinden, leerling, motiveren, mee, ding, uitleggenen, weten
Topic 3	duidelijk, goed, uitleg, maken, powerpoint, voorbeeld, laten, huiswerk, zien, gebruiken, regel, opdracht, film, vertellen, aantekening, leren, bord, uitleggenen, lesdoel, planning
Topic 4	heel, goed, aardig, erg, vinden, fijn, streng, leraar, klas, vak, docent, vrolijk, mevrouw, enthousiast, snel, uitleggenen, echt, weten, blijven, rustig
Topic 5	les, leuk, goed, voorbereiden, maken, vertellen, manier, bereid, grap, interessant, boeiend, soms, ding, saai, gezellig, grappig, opdracht, verhaal, boeien, waardoor
Topic 6	snappen, vraag, goed, vragen, keer, weten, leggen, komen, zodat, begrijpen, willen, beantwoorden, antwoord, uitleggenen, stellen, luisteren, fout, uitleggenals, ding, meestal
Topic 7	iedereen, zorgen, leerling, betrekken, aandacht, proberen, ervoor, respect, behandelen, opletten, begrijpen, werk, zetten, blijven, mens, omgaan, denken, elkaar, erbij, lesstof
Topic 8	les, gaan, goed, leren, toets, tijd, zeggen, herhalen, vertellen, bereiden, weten, stof, bespreken, uitleggenwat, begin, huiswerk, nemen, vorig, komen, proefwerk

Table B
The eight extracted topics based on the topic modeling.

Topics	Translated words (English)	Marginal topic distribution (%)	Label	Description	Examples of student responses (topic weight in percent)
Topic 1	to explain, good, clearly, teaching, explaining, thing, teaching, material, example, teaching, material, to understand, assignment, grammar, to draw, new, learning matter, to expand, calm, punishment, Mrs, enthusiastic	26.3	Clear explanation	The provided explanation is good. Teaching is clear, and teaching materials are used. The instruction is teacher-centered.	"Teach well and explain clearly" (.85). "She teaches very clearly and can explain well" (.69). "Teaching well and explaining everything clearly. And he really takes the time to explain it properly" (.59).
Topic 2	to help, good, to hold, understanding, class, let, to work, order, to talk, difficult, child, required, man, find, pupil,	14.7	Student-centered supportive learning climate	The teacher supports and keeps an eye on students, helps them, actively follows their development and learning, and	"She helps if someone does not understand something and can also keep the class under control and make sure that we are quiet

(*continued on next page*)

Table B (*continued*)

Topics	Translated words (English)	Marginal topic distribution (%)	Label	Description	Examples of student responses (topic weight in percent)
	motivate, along, thing, explain, know			ensures that students are on the right track. The support is student-centered.	and continue to work properly" (.94). "She motivates people to do assignments well she helps if you don't understand something she explains the assignment well she knows what she's talking about she repeats it again if you don't understand she can keep us well under control" (.88). "It is always relaxed but works well and she keeps order. She always helps when I don't get something and she keeps going until I get it" (.70).
Topic 3	clearly, good, explanation, to make, PowerPoint, example, let, homework, see, use, rule, order, movie, to tell, note, to learn, board, explain, lesson goal, planning	14.3	Lesson variety	The teacher reinforces clear explanation by making it more visible and tangible, supporting it with diverse illustrations, and using visual and audio materials. The teacher provides various teaching materials and increases the lesson variety—for example, giving independent or group assignments.	"A good and clear explanation about an assignment or chapter. Checks well if we have or have made our homework and stuff" (.81). "He prepares his lessons well. He always has clear powerpoints that are well-organized, and there is a clear schedule for the period. He also makes a fairly good distinction between explanation and making assignments independently/together" (.58). "Makes the lessons very interesting, with powerpoints and small assignments. Gives a good and clear explanation. Also puts the powerpoints online so that you can always see what you had to learn before a test" (.50).
Topic 4	very good, nice, very, find, fine, strict, teacher, class, subject, teacher, cheerful, Mrs, enthusiastic, fast, explain, really, know, stay, calm	10.3	Likable characteristics of the teacher	The teacher is a likable and nice person, which represents the (positive) affective characteristics of the teacher.	"She is very kind and she is very good" (.87). "He explains everything very well and is one of my favorite teachers. I like him very much and he explains well" (.74). "He is very good at his profession/subject. He is a very good teacher because he helps you when you need help and he motivates you and he is very nice" (.65).
Topic 5	lesson, fun, good, to prepare, to make, to tell, way, prepared, joke, interesting, fascinating, sometimes, thing boring, pleasant, funny, assignment, story, captivate, causing	12.8	Evoking interest	The teacher is a performer, which makes the lesson fun, pleasant, and interesting; the teacher prepares the lesson well; and the learning activation is teacher centered.	"Always fun and well-prepared lessons that help me" (.82). "Always prepare his lessons well and explain the material in an understandable and interesting way. He alternates between fun and serious lessons, so his lessons never become monotonous" (.71). "Prepare the lessons well, fun lessons such as learning with games, interesting storytelling, showing videos" (.61).
Topic 6	understanding, question, good, ask, times, know, explain, come, so that, to understand, want, reply, answer, explain, claim, listen, wrong, explain as, thing, mostly	7.2	Monitoring understanding	The teacher is willing to answer questions, repeats regularly, and replies to questions to be sure that students understand the lesson material.	"If you have a question, he will explain it well and he will also ask you if you have understood it subsequently. If you still don't get it then he just explains it again" (.69). "He always pays a lot of attention to everyone. He asks carefully what you do and do not understand and he explains

(*continued on next page*)

Table B (*continued*)

Topics	Translated words (English)	Marginal topic distribution (%)	Label	Description	Examples of student responses (topic weight in percent)
					everything where necessary. Even though I have already asked that question 3 times. He also supervises well at the study center. There he explains everything again and checks if I really understood it" (.53). "She clearly watches everyone, so if someone doesn't participate, she tells them that and they do. She gives many examples that you can practice with. [She] also clearly states what is being asked. If you don't understand the question, she explains it very accurately, so that you understand well what the intention is and what the answer is" (.52).
Topic 7	everybody, to care, pupil, involve, attention, to try, before, respect, to treat, pay attention, to understand, work, put, stay, man, to work through, to think, each other, with it, teaching material	6.0	Inclusiveness and equity	Students in the class are cared for and involved. They are not treated differently. They are treated with equal respect.	"Involving students and ensuring that everyone participates and listens" (.87) "Treat students with respect. She is a committed teacher, takes every student into account and ensures that everyone becomes independent with regard to the subject of dutch" (.50). "She ensures that everyone is involved in the lesson and just chooses students for an assignment (not the same students all the time)" (.41).
Topic 8	lesson, to go, good, to learn, test, time, to say, to repeat, to tell, to prepare, to know, material, to discuss, to explain what, to begin, homework, to take, previous, to come, test	8.4	Lesson objectives and formative assessment	The teacher regularly checks and prepares students for tests to ensure that the goals are achieved. The teacher states the lesson objectives and explains what students will do in the lesson, what they will learn, and what they need for the exam.	"Explain to them well what we are going to cover in class. She also clearly explains what we have to learn for a test" (.81). "Always explain what you need to learn for a test, they also repeat this in the lessons in such a way that we are well prepared for the tests. She also always explains well what we are going to do in class and what you have to do for the homework" (.64). "He always tells us what we will do throughout the lesson. He keeps us well informed with tests and when they are. He also tells us what we have to learn for the test. He can explain well and we often repeat the lesson material at the beginning of the lesson what we did last time" (.63).

References

Aiyanyo, I. D., Samuel, H., & Lim, H. (2021). Effects of the COVID-19 pandemic on classrooms: A case study on foreigners in South Korea using applied machine learning. *Sustainability, 13*(9), 4986. https://doi.org/10.3390/su13094986

Aksoy, N. (1998). *Opinions of upper elementary students about a "good teacher" (Case study in Turkey). Paper presented at the 29th Annual.* Ellenville, NY: Meeting of the Northeastern Educational Research Association. October 28–30.

Alkan, V. (2013). Pupils' voice: "My primary school teacher. *Educational Research and Reviews, 8*(11), 777–784. https://doi.org/10.5897/ERR2013.1422

Auwarter, A. E., & Aruguete, M. S. (2008). Effects of student gender and socioeconomic status on teacher perceptions. *Journal of Educational Research, 101*(4), 242–246. https://doi.org/10.3200/JOER.101.4.243-246

Bakx, A., Koopman, M., De Kruijf, J., & Den Brok, P. (2015). Primary school pupils views of characteristics of good primary school teachers: An exploratory, open approach for investigating pupils perceptions. *Teachers and Teaching: Theory and Practice, 21*(5), 543–564. https://doi.org/10.1080/13540602.2014.995477

Balahadia, F. F., Fernando, M. C. G., & Juanatas, I. C. (2016). Teacher's performance evaluation tool using opinion mining with sentiment analysis. In *IEEE region 10 symposium (TENSYMP)* (pp. 95–98). https://doi.org/10.1109/TENCONSpring.2016.7519384
Beishuizen, J. J., Hof, E., van Putten, C. M., Bouwmeester, S., & Asscher, J. J. (2001). Students' and teachers' cognitions about good teachers. *British Journal of Educational Psychology, 71*, 185–201.
Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement, 30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014
Bent, M., Velazquez-Godinez, E., & Jong, F. De (2021). Becoming an expert teacher: Assessing expertise growth in peer feedback video recordings by lexical analysis. *Education Sciences, 11*(665). https://doi.org/10.3390/educsci11110665
Berry, M. W., Mohamed, A., & Yap, B. W. (2020). In M. W. Berry, A. Mohamed, & B. W. Yap (Eds.), *Supervised and unsupervised learning for data science*. Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2.
Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9
Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878
Buenano-Fernandez, D., Gonzalez, M., Gil, D., & Lujan-Mora, S. (2020). Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach. *IEEE Access, 8*, 35318–35330. https://doi.org/10.1109/ACCESS.2020.2974983
Bullock, M. (2015). What makes a good teacher? Exploring student and teacher beliefs on good teaching. *Rising Tide, 7*, 1–30.
Campbell, D. T., & Cook, T. D. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Boston).
Chen, B., Chen, X., & Xing, W. (2015). Twitter archeology" of learning analytics and knowledge conferences. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge - LAK, '15*, 16–20. https://doi.org/10.1145/2723576.2723584. *-Marc*, 340–349.
Chen, C.-M., Li, M.-C., Chang, W.-C., & Chen, X.-X. (2021). Developing a topic analysis instant feedback system to facilitate asynchronous online discussion effectiveness. *Computers & Education, 163*. https://doi.org/10.1016/j.compedu.2020.104095
Chen, H., Li, M., Ni, X., Zheng, Q., & Li, L. (2021). Teacher effectiveness and teacher growth from student ratings: An action research of school-based teacher evaluation. *Studies In Educational Evaluation, 70*, Article 101010. https://doi.org/10.1016/j.stueduc.2021.101010
Chen, X., Xie, H., Li, Z., & Cheng, G. (2021). Topic analysis and development in knowledge graph research: A bibliometric review on three decades. *Neurocomputing, 461*, 497–515. https://doi.org/10.1016/j.neucom.2021.02.098
Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals. In *ACM international conference proceeding series* (pp. 1–5). https://doi.org/10.1145/2883851.2883951
Chen, X., Zou, D., Cheng, G., & Xie, H. (2020). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of computers & education. *Computers & Education, 151*. https://doi.org/10.1016/J.COMPEDU.2020.103855
Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of International Working Conference on Advanced Visual Interfaces (AVI)*, 74–77.
Civitillo, S., Denessen, E., & Molenaar, I. (2016). How to see the classroom through the eyes of a teacher: Consistency between perceptions on diversity and differentiation practices. *Journal of Research in Special Educational Needs, 16*, 587–591. https://doi.org/10.1111/1471-3802.12190
Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.
Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. Routledge.
Cunningham-Nelson, S., Baktashmotlagh, M., & Boles, W. (2019). Visualizing student opinion through text analysis. *IEEE Transactions on Education, 62*(4), 305–311. https://doi.org/10.1109/TE.2019.2924385
Danielson, C. (2013). *The framework for teaching: Evaluation instrument*. Princeton, NJ: Danielson Group.
Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x
Ezen-Can, A., & Boyer, K. E. (2013). Unsupervised classification of student dialogue acts with query-likelihood clustering. In *Proceedings of the 6th international conference on educational data mining (EDM)* (pp. 20–27).
Farsani, M., Beikmohammadi, M., & Mohebbi, A. (2014). Self-regulated learning, goal-oriented learning, and academic writing performance of undergraduate Iranian EFL learners. *The Electronic Journal for English as a Second Language or Foreign Language, 18*(2), 1–19.
Feinerer, I. (2007). Automated coding of qualitative interviews with latent semantic analysis. *Information Systems Technology and Its Applications—6th International Conference—ISTA*, 66–77.
Finch, W. H., Hernández Finch, M. E., McIntosh, C. E., & Braun, C. (2018). The use of topic modeling with latent Dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science, 4*(4), 403–424. https://doi.org/10.1037/tps0000173
Finefter-Rosenbluh, I. (2020). Try walking in my shoes": Teachers' interpretation of student perception surveys and the role of self-efficacy beliefs, perspective taking and inclusivity in teacher evaluation. *Cambridge Journal of Education, 50*(6), 747–769. https://doi.org/10.1080/0305764X.2020.1770692
García-Moya, I., Brooks, F., & Moreno, C. (2020). Humanizing and conducive to learning: An adolescent students' perspective on the central attributes of positive relationships with teachers. *European Journal of Psychology of Education, 35*(1), 1–20. https://doi.org/10.1007/s10212-019-00413-z
Gibson, A., & Kitto, K. (2015). Analysing reflective text for learning analytics: An approach using anomaly recontextualisation. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge—LAK, '15*, 275–279. https://doi.org/10.1145/2723576.2723635
Göbel, K., Wyss, C., Neuber, K., & Raaflaub, M. (2021). Student feedback as a source for reflection in practical phases of teacher education. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools using student perceptions for the development of teaching and teachers* (pp. 173–189). Cham: Springer.
Goloshchapova, I., Poon, S. H., Pritchard, M., & Reed, P. (2019). Corporate social responsibility reports: Topic analysis and big data approach. *The European Journal of Finance, 25*(17), 1637–1654. https://doi.org/10.1080/1351847X.2019.1572637
Good, T. L., Wiley, C. R. H., & Florez, I. R. (2009). Effective teaching: An emerging synthesis. In L. J. Saha, & A. G. Dworkin (Eds.), *International handbook of research on teachers and teaching*. Springer. https://doi.org/10.1007/978-0-387-73317-3.
Gottipati, S., Shankararaman, V., & Lin, J. R. (2018). Latent dirichlet allocation for textual student feedback analysis. *Proceedings of the 26th International Conference on Computers in Education ICCE*, 220–227, 2018.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(SUPPL. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101
van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research, 49*(2), 127–152. https://doi.org/10.1080/00131880701369651
van de Grift, W. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement, 25*(3), 295–311. https://doi.org/10.1080/09243453.2013.794845
Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028
Grönberg, N., Knutas, A., Hynninen, T., & Hujala, M. (2021). Palaute: An online text mining tool for analyzing written student course feedback. *IEEE Access, 9*, 134518–134529. https://doi.org/10.1109/ACCESS.2021.3116425
Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management, 54*(6), 1292–1307. https://doi.org/10.1016/j.ipm.2018.05.006
Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Helms-Lorenz, M., van de Grift, W., Canrinus, E., Maulana, R., & van Veen, K. (2018). Evaluation of the behavioral and affective outcomes of novice teachers working in professional development schools versus non-professional development schools. *Studies in Educational Evaluation, 56*, 8–20. https://doi.org/10.1016/j.stueduc.2017.10.006

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229–247. https://doi.org/10.1111/j.1540-5907.2009.00428.x

Hujala, M., Knutas, A., Hynninen, T., & Arminen, H. (2020). Improving the quality of teaching by utilising written student feedback: A streamlined process. *Computers & Education, 157*. https://doi.org/10.1016/j.compedu.2020.103965

Ida, S. Z. (2017). What makes a good teacher? *Universal Journal of Educational Research, 5*(1), 141–147. https://doi.org/10.1080/00131725209341464

Inda-Caro, M., Maulana, R., Fernández-García, C. M., Peña-Calvo, J. V., Rodríguez-Menéndez, M. del C., & Helms-Lorenz, M. (2019). Validating a model of effective teaching behaviour and student engagement: perspectives from Spanish students. *Learning Environments Research, 22*(2), 229–251. https://doi.org/10.1007/s10984-018-9275-z

Kandula, S., Curtis, D., Hill, B., & Zeng-Treitler, Q. (2011). Use of topic modeling for recommending relevant education material to diabetic patients. In *Annual symposium proceedings/AMIA symposium* (pp. 674–682).

Kherwa, P., & Bansal, P. (2019). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems, 7*(24), 1–16.

Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & 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*(1), 82–102.

Kim, N., & Son, Y. (2021). Multilevel latent profile analysis of Korean middle school student perceptions of teaching methods. In *Asia pacific education review*. https://doi.org/10.1007/s12564-021-09721-w

Könings, K. D., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2005). Towards more powerful learning environments through combining the perspectives of designers, teachers, and students. *British Journal of Educational Psychology, 75*(4), 645–660. https://doi.org/10.1348/000709905X43616

Könings, K. D., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2011). Participatory instructional redesign by students and teachers in secondary education: Effects on perceptions of instruction. *Instructional Science, 39*(5), 737–762. https://doi.org/10.1007/s11251-010-9152-3

Kutnick, P., & Jules, V. (1993). Pupils' perceptions of a good teacher: A developmental perspective from Trinidad and Tobago. *British Journal of Educational Psychology, 63*(3), 400–413. https://doi.org/10.1111/j.2044-8279.1993.tb01067.x

Kyriakides, L., Anthimou, M., & Panayiotou, A. (2020). Searching for the impact of teacher behavior on promoting students' cognitive and metacognitive skills. *Studies In Educational Evaluation, 64*, Article 100810. https://doi.org/10.1016/j.stueduc.2019.100810

Kyriakides, L., Creemers, B. P. M., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25*(1), 12–23. https://doi.org/10.1016/j.tate.2008.06.001

Lämsä, J., Espinoza, C., Tuhkala, A., & Hämäläinen, R. (2021). Staying at the front line of literature: How can topic modelling help researchers follow recent studies? *Frontline Learning Research, 9*(3), 1–12. https://doi.org/10.14786/flr.v9i3.645

Lau, J. H., Grieser, K., Newman, D., & Baldwin, T. (2011). Automatic labelling of topic models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1536–1545.

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th conference of the European* (pp. 530–539). https://doi.org/10.3115/v1/E14-1056.

Lee, H., Kwak, J., Song, M., & Kim, C. O. (2014). Coherence analysis of research and education using topic modeling. *Scientometrics, 102*(2), 1119–1137. https://doi.org/10.1007/s11192-014-1453-x

Liu, C., Zou, D., Chen, X., Xie, H., & Chan, W. H. (2021). A bibliometric review on latent topics and trends of the empirical MOOC literature (2008–2019). *Asia Pacific Education Review, 22*(3), 515–534. https://doi.org/10.1007/s12564-021-09692-y

Longo, L. (2020). Empowering qualitative research methods in education with artificial intelligence. In A. Costa, L. Reis, & A. Moreira (Eds.), *Advances in intelligent systems and computing: Vol. 1068. Computer supported qualitative research. WCQR 2019*. Springer. https://doi.org/10.1007/978-3-030-31787-4_1.

Looney, J. (2011). Developing high-quality teachers: Teacher evaluation for improvement. *European Journal of Education, 46*(4), 440–455.

Lynch, M., & Salikhova, N. (2017). Teachers' beliefs about the needs of students: Teachers as local experts (A qualitative analysis). *Education and Self Development, 12*(3), 33–43. https://doi.org/10.26907/esd12.3.03

Mælan, E. N., Tjomsland, H. E., Samdal, O., & Thurston, M. (2020). Pupils' perceptions of how teachers' everyday practices support their mental health: A qualitative study of pupils aged 14–15 in Norway. *Scandinavian Journal of Educational Research, 64*(7), 1015–1029. https://doi.org/10.1080/00313831.2019.1639819

Mandouit, L. (2018). Using student feedback to improve teaching. *Educational Action Research, 26*(5), 755–769. https://doi.org/10.1080/09650792.2018.1426470

Matsukawa, H., Oyama, M., Negishi, C., Arai, Y., Iwasaki, C., & Hotta, H. (2019). Analysis of the free descriptions obtained through course evaluation questionnaires using topic modeling. *Educational Technology Research, 41*(1), 125–137. https://doi.org/10.15077/etr.42154

Maulana, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: construct representation and predictive quality. *Learn. Environ. Res., 19*(3), 335–357. https://doi.org/10.1007/s10984-016-9215-8

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015a). A longitudinal study of induction on the acceleration of growth in teaching quality of beginning teachers through the eyes of their students. *Teaching and Teacher Education, 51*, 225–245. https://doi.org/10.1016/j.tate.2015.07.003.

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015b). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*(2), 169–194. https://doi.org/10.1080/09243453.2014.939198

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2017). Validating a model of effective teaching behaviour of pre-service teachers. Teachers and Teaching. *Theory and Practice, 23*(4), 471–493. https://doi.org/10.1080/13540602.2016.1211102

Maulana, R., Smale-Jacobse, A., Helms-Lorenz, M., Chun, S., & Lee, O. (2019). Measuring differentiated instruction in The Netherlands and South Korea: factor structure equivalence, correlates, and complexity level. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-019-00446-4

Maxwell, J. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62*(3), 279–301. https://doi.org/10.17763/haer.62.3.8323320856251826

Mayring, P. H. (2010). *Qualitative inhaltsanalyse (Qualitative content analysis)* (11th ed.) (11th ed., Vol. 1) (Beltz).

Meeuwisse, M., Severiens, S. E., & Born, M. P. (2010). Learning environment, interaction, sense of belonging and study success in ethnically diverse student groups. *Research in Higher Education, 51*(6), 528–545. https://doi.org/10.1007/s11162-010-9168-1

Moretti, A., McKnight, K., & Salleb-Aouissi, A. (2015). Application of sentiment and topic analysis to teacher evaluation policy in the U.S. *Proceedings of the 8th International Conference on Educational Data Mining*, 628–629. http://www.educationaldatamining.org/EDM2015/uploads/papers/paper_310.pdf.

Muijs, D., Campbell, J., Kyriakides, L., & Robinson, W. (2005). Making the case for differentiated teacher effectiveness: An overview of research in four key areas. *School Effectiveness and School Improvement, 16*(1), 51–70. https://doi.org/10.1080/09243450500113985

Mullock, B. (2003). What makes a good teacher? The perceptions of postgraduate TESOL students. *Prospect, 18*(3), 3–24.

Murphy, P. K., Delli, L. A. M., & Edwards, M. N. (2004). The good teacher and good teaching: Comparing beliefs of second-grade students, preservice teachers, and inservice teachers. *The Journal of Experimental Education, 72*(2), 69–92.

Nanda, G., Douglas, A., Waller, K. R., D., E. Merzdorf, H., Goldwasser, D. (2021). Analyzing large collections of open-ended feedback from MOOC learners using LDA topic modeling and qualitative analysis. *IEEE Transactions on Learning Technologies, 14*(2), 146–160. https://doi.org/10.1109/TLT.2021.3064798

Nanda, G., Hicks, N. M., Waller, D. R., Goldwasser, D., & Douglas, K. A. (2018). Understanding learners' opinion about participation certificates in online courses using topic modeling. *Proceedings of the 11th International Conference on Educational Data Mining, EDM*, 376–382, 2018.

Nowlin, M. C. (2016). Modeling issue definitions using quantitative text analysis. *Policy Studies Journal, 44*(3), 309–331. https://doi.org/10.1111/psj.12110

Panayiotou, A., Kyriakides, L., Creemers, B. P. M., McMahon, L., Vanlaar, G., Pfeifer, M., Rekalidou, G., & Bren, M. (2014). Teacher behavior and student outcomes: Results of a European study. *Educational Assessment, Evaluation and Accountability, 26*(1), 73–93. https://doi.org/10.1007/s11092-013-9182-x

Parsons, S., & Khuri, N. (2020). Discovery of research trends in computer science education on ethics using topic modeling. In *International conference on computational science and computational intelligence (CSCI)* (pp. 885–891). https://doi.org/10.1109/CSCI51800.2020.00166

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119. https://doi.org/10.3102/0013189X09332374

Pineda-Báez, C., Hennig Manzuoli, C., & Vargas Sánchez, A. (2019). Supporting student cognitive and agentic engagement: Students' voices. *International Journal of Educational Research, 96*(January), 81–90. https://doi.org/10.1016/j.ijer.2019.06.005

Pintrich, P. R., & Schrauben, B. (1992). Students' motivational beliefs and their cognitive engagement in classroom academic tasks. In D. H. Schunk, & J. L. Meece (Eds.), *Student perceptions in the classroom* (pp. 149–183). Erlbaum.

Plavšić, M., & Diković, M. (2016). Do teachers, students and parents agree about the top five good teacher's characteristics? *Education Provision to Every One: Comparing Perspectives from Around the World BCES Conference Books, 14*(1), 120–126.

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM - Mathematics Education, 50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & Mcfarland, D. A. (2009). Topic modeling for the social sciences. *NIPS 2009 workshop on applications for topic models: Text and Beyond, 5*, 1–4. Stanford University http://www.umiacs.umd.edu/~jbg/nips_tm_workshop/23.pdf.

Ramesh, A., Goldwasser, D., Huang, B., Daume, H., & Getoor, L. (2014). Understanding MOOC discussion forums using seeded LDA. In *Proceeding of 9th workshop on innovative use of NLP for building educational applications* (pp. 28–33). https://doi.org/10.3115/v1/w14-1804

Reedy, G. B. (2008). PowerPoint, interactive whiteboards, and the visual culture of technology in schools. *Technology, Pedagogy and Education, 17*(2), 143–162. https://doi.org/10.1080/14759390802098623

Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., & Stewart, B. (2015). Computer-assisted reading and discovery for student generated text in massive open online courses. *Journal of Learning Analytics, 2*(1), 156–184. https://doi.org/10.18608/jla.2015.21.8

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Russom, P. (2011). Big data analytics. In *TDWI best practices report*. Fourth Quarter: The Data Warehousing Institute (TDWI).

Sanchez, B., Byra, M., & Wallhead, T. L. (2012). Students' perceptions of the command, practice, and inclusion styles of teaching. *Physical Education and Sport Pedagogy, 17*(3), 317–330. https://doi.org/10.1080/17408989.2012.690864

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness: A critical review of the knowledge base*. Springer.

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods, 10*(2), 143–152. https://doi.org/10.18148/srm/2016.v10i2.6213

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. https://doi.org/10.3102/0034654307310317

Sekiya, T., Matsuda, Y., & Yamaguchi, K. (2015). Curriculum analysis of CS departments based on CS2013 by simplified, supervised LDA. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge—LAK, '15*, 16–20. https://doi.org/10.1145/2723576.2723594. *-Marc*, 330–339.

Sethi, J., & Scales, P. C. (2020). Developmental relationships and school success: How teachers, parents, and friends affect educational outcomes and what actions students say matter most. *Contemporary Educational Psychology, 63*, Article 101904. https://doi.org/10.1016/j.cedpsych.2020.101904

Shuell, T. J. (1993). Toward an integrated theory of teaching and learning. *Educational Psychologist, 28*(4), 291–311. https://doi.org/10.1207/s15326985ep2804_1

Shuell, T. J. (1996a). Teaching and learning in a classroom context. In D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 726–764). London, England: Prentice Hall International. Macmillan Library Reference Usa.

Shuell, T. J. (1996b). The role of educational psychology in the preparation of teachers. *Educational Psychologist, 31*(1), 5–14. https://doi.org/10.1207/s15326985ep3101_1

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. https://doi.org/10.3115/v1/w14-3110

Sins, P. H. M., van Joolingen, W. R., Savelsbergh, E. R., & van Hout-Wolters, B. (2008). Motivation and performance within a collaborative computer-based modeling task: Relations between students' achievement goal orientation, self-efficacy, cognitive processing, and achievement. *Contemporary Educational Psychology, 33*(1), 58–77. https://doi.org/10.1016/j.cedpsych.2006.12.004

Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M., & Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence. *Frontiers in Psychology, 10*, 2366. https://doi.org/10.3389/fpsyg.2019.02366

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. *Proceedings of the Third International Conference on Learning Analytics and Knowledge—LAK, '13*, 38–47. https://doi.org/10.1145/2460296.2460307

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. *Proceedings of the 31st International Conference on Machine Learning*, 337–345.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., Conover, L. A., & Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted, 27*(2–3), 119–145. https://doi.org/10.1177/016235320302700203

Unankard, S., & Nadee, W. (2019). Topic detection for online course feedback using LDA. In E. Popescu, T. Hao, T.-C. Hsu, H. Xie, M. Temperini, & W. Chen (Eds.), *4th international symposium: Emerging technologies for education* (pp. 133–142). Springer. https://doi.org/10.1007/978-3-030-38778-5.

Vargas-Calderón, V., Flórez, J. S., Ardila, L. F., Parra-A., N., Camargo, J. E., & Vargas, N. (2020). Learning from students' perception on professors through opinion mining. In , *Communications in computer and information science: Vol. 1277. F. H. & M. S. (Eds.), applied informatics. ICAI 2020*. Cham: Springer. https://doi.org/10.1007/978-3-030-61702-8_23.

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems, 94*, Article 101582. https://doi.org/10.1016/j.is.2020.101582

Wachen, J. (2017). *Media coverage of educational testing: Understanding issue dimensions using topic modeling* (Doctoral dissertation, University of North Carolina).

Wang, Y. (2021). When artificial intelligence meets educational leaders' data-informed decision-making: A cautionary tale. *Studies In Educational Evaluation, 69*, Article 100872. https://doi.org/10.1016/j.stueduc.2020.100872

Wang, Y., Bowers, A. J., & Fikis, D. J. (2017). Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of EAQ articles from 1965 to 2014. *Educational Administration Quarterly, 53*(2), 289–323. https://doi.org/10.1177/0013161X16660585

Watkins, D. (2004). Teachers as scholars of their students' conceptions of learning: A Hong Kong investigation. *British Journal of Educational Psychology, 74*(3), 361–373. https://doi.org/10.1348/0007099041552332

Yu, M. V. B., Johnson, H. E., Deutsch, N. L., & Varga, S. M. (2018). She calls me by my last name": Exploring adolescent perceptions of positive teacher-student relationships. *Journal of Adolescent Research, 33*(3), 332–362. https://doi.org/10.1177/0743558416684958