Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Pedagogical discourse markers in online algebra learning: Unraveling instructor's communication using natural la

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Shin, Jinnie	University of Florida, USA	
Balyan, Renu	SUNY Old Westbury, USA	
Banawan, Michelle P.	Asian Institute of Management, Philippines	
Arner, Tracy	Arizona State University, USA	
Leite, Walter L.	University of Florida, USA	
McNamara, Danielle S.	Arizona State University, USA	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) August 11, 2023

Check type of content being submitted and complete one of the following in the box below:

- ☑ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Computers & Education, Volume 205

DOI or URL to published work (if available) https://doi.org/10.1016/j.compedu.2023.104897

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

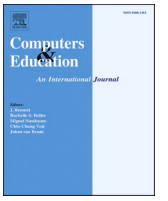
"This work was supported by U.S. Department of Education **[Office name]** National Center for Educat through **[Grant number]** R305C1600 to **Institution]** University of Fl. The opinions expressed are those of the authors and do not represent views of the **[Office name]** National Center for Educa or the U.S. Department of Education.

Computers & Education 205 (2023) 104897

Contents lists available at ScienceDirect

Computers & Education

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



Pedagogical discourse markers in online algebra learning: Unraveling instructor's communication using natural language processing

Jinnie Shin [a,*], Renu Balyan [b], Michelle P. Banawan [c], Tracy Arner [d], Walter L. Leite [a], Danielle S. McNamara [d]

[a] University of Florida, USA
[b] SUNY Old Westbury, USA
[c] Asian Institute of Management, Philippines
[d] Arizona State University, USA

ARTICLE INFO

Keywords:
Pedagogical communication
Online learning
Video lectures
Natural language processing

ABSTRACT

Despite the proliferation of video-based instruction and its benefits—such as promoting student autonomy and self-paced learning—the complexities of online teaching remain a challenge. To be effective, educators require extensive training in digital teaching methodologies. As such, there's a pressing need to examine and comprehend the intricacies of instructors' communication patterns within this context. This research addresses the pressing need to understand pedagogical discourse in online video lectures in Algebra classes by employing computational linguistic tools and natural language processing (NLP). Using transcripts from 125 Algebra 1 video lectures—comprising 4962 instances of pedagogical discourse—from five instructors at Math Nation, a virtual math learning environment, we analyzed the conveyance of linguistic, attitudinal, and emotional nuances. With the aid of 26 Coh-Metrix and SÉANCE features, we classified educators' language choices, achieving an accuracy of 86.7%. Furthermore, variations in language choices, as signified by discourse markers, were examined through a K-means clustering approach. The resulting 17 clusters were grouped into interpersonal, structural, and cognitive pedagogic functions. Through this exploration, we demonstrate the promising potential of NLP in efficiently deciphering pedagogical communication patterns in video lectures. These insights open a new avenue for research, aimed at assessing the efficacy of digital instruction by scrutinizing pedagogical discourse characteristics in computer-based learning environments.

1. Introduction

The broad adoption of video-based instruction in virtual learning environments has inspired a vast body of research examining their impact on students' learning outcomes. Video lectures have some potential to positively enhance students' experience and performance (Shanshan et al., 2022). Video lectures help students to access learning resources efficiently and effectively (Buckley & Smith, 2007; Choi & Johnson, 2005) and afford students greater independence in the learning process (MacKay, 2019; Scagnoli et al.,

* Corresponding author.
E-mail addresses: Jinnie.shin@coe.ufl.edu (J. Shin), balyanr@oldwestbury.edu (R. Balyan), mbanawan@aim.edu (M.P. Banawan), tarner@asu.edu (T. Arner), Walter.Leite@coe.ufl.edu (W.L. Leite), Dsmcnamara1@gmail.com (D.S. McNamara).

https://doi.org/10.1016/j.compedu.2023.104897
Received 15 December 2022; Received in revised form 7 August 2023; Accepted 9 August 2023
Available online 11 August 2023

2019). Video lectures are also used by teachers to provide individual remediation and support small groups of students.

While potentially effective and efficient, there are several considerations regarding video production such as voiceover, talking head, recorded live lecture (Beege et al., 2022; Pi et al., 2019), and, most importantly, linguistic choices that potentially impact the quality of video lectures (Chen & Wu, 2015). A vast body of previous work has investigated how video instruction should be formatted (i.e., how it looks) and how the instructional voice should be presented (i.e., how it sounds), using traditional research methods to compare learning outcomes in varied conditions (Clark & Mayer, 2011; Madariaga et al., 2021; Mayer, 2005). Moreover, researchers have focused on linguistic features and utilized linguistic analyses to inform the production and refinement of video lectures, aiming to optimize student learning outcomes.

Beyond the visual and auditory elements, linguistic features serve as a cornerstone in video-based instruction (Johnson & Picciuolo, 2022). Linguistic features refer to the observable and analyzable characteristics of language, encompassing various elements such as vocabulary, syntax, grammar, phonetics, and discourse spoken and written by instructors and/or students (Alghamdi & Velloso, 2022; Karasavvidis et al., 2022). These features are pivotal in communication during video-based lectures, enabling instructors to convey meaning, express ideas, and understand one another effectively (Shanshan et al., 2022; Li et al., 2023).

Scholars have employed linguistic analyses to improve the quality and effectiveness of learning materials. Strategies such as word embeddings and video fragmenting—which aligns the visual content of video lectures with the lecturer's discourse—have been explored as effective methods to enhance lecture quality using linguistic features. For example, word embeddings have been used to improve video fragmenting, a method of segmenting video lectures (Mayer, 2020), to align visual content of video lectures with discourse of the lecturer. Fragmenting video lectures to support student learning by improving indexing and enhancing searchability of video lectures has been explored (Galanopoulos & Mezaris, 2019). Further, linguistic analysis has been used to evaluate students' learning from video instruction using text produced in an integrated annotation tool. Joksimović and colleagues (2019) used Coh-metrix to analyze linguistic features of students' annotations from an integrated video annotation tool to evaluate students' self-reflections of their learning from video lectures.

While there has been a broad range of research investigating features of video lectures and integrated tool use, there remains a paucity of empirical research investigating the most effective linguistic features and structure of the words spoken by the lecturer (i.e., pedagogical communication) in video instruction (Henderson & Schroeder, 2021; Polat, 2022). Yet, pedagogical communication is a critical consideration in producing video lectures for more effective computer-based education (Johnson & Picciuolo, 2022; Beege et al., 2022, Shoufan, 2019).

Pedagogical communication, instructors' spoken words, has previously been studied in traditional classroom settings on multiple dimensions that contribute to student learning outcomes. It is also important to note that in addition to providing direct instruction that supports acquisition of knowledge, educators are frequently charged with affective support of their students (e.g., emotional support, belonging, motivation) which is also essential for positive learning outcomes (Aspelin, 2012).

Teachers' influence on students' success in the classroom stems from didactics as well as non-cognitive support such as motivation and engagement. Students' attitudes about learning are derived, in part, from their self-concept, self-efficacy, and feelings of relatedness within the classroom. These beliefs and attitudes about learning are key predictors of improved school outcomes, including achievement (Guay et al., 2019). The language choices that teachers make, informs students' development of beliefs and attitudes about learning, particularly in their ability to solve a problem, complete a task, or learn a concept (Al-Ghamdi, 2017; Burnett, 2010). Given the importan ce of students' attitudes and beliefs about learning and the influence of teachers' word choices on those beliefs, it is important to understand how the linguistic, attitudinal, and emotion in pedagogical communication in asynchronous video lectures may vary.

The objective of this study is to enhance our comprehension of pedagogical communication in the setting of online video lectures. We aim to achieve this by presenting and showcasing a methodical analytical approach that can identify and elucidate the linguistic attributes of pedagogical communication in video-based instruction utilized in virtual learning environments. Our approach addresses the limitations of prior research that necessitates a substantial amount of human-labeled data to interpret instructors' communication in a learning setting. In this study, we leverage multiple computational linguistic tools and natural language processing methods with an unsupervised learning algorithm. We mesh these affordances with pedagogical discourse theory Fung and Carter's (2007). Their framework has been primarily adopted in the research of traditional classroom settings, indicating positive associations between teachers' intentional communication strategies with student learning.

However, few to no studies, to our knowledge, have adopted the framework to understand pedagogical communication in virtual learning environments, such as online video lectures (Vickov and Jakupcevic, 2017; Tsai & Chu, 2017; Beege et al., 2022). Closing this gap in the research in pedagogical discourse, online learning, and natural language processing applications in education serves as the primary goal of this article. The following sections provide an overview of pedagogical communication in traditional and online classrooms as well as methods to evaluate pedagogical communication leveraging artificial intelligence and NLP. We conclude by providing evidence how the findings in this study extend the understanding of online communication in an instructional context.

2. Literature review

2.1. Pedagogical communication

Pedagogical communication includes instructors' intentional word use and language choices that have a positive impact on students' learning experiences, outcomes, and evaluation of teaching (Nussbaum, 1992). Studies investigating effective pedagogical communication have largely taken place in traditional, face-to-face classroom environments and evaluated on the dimensions of

instructional clarity (Chesebro & McCroskey, 1998), emotional support (She & Fisher, 2002), and cognitive support (Amadi & Paul, 2017). These dimensions are commonly highlighted as having significant positive impacts on student learning.

Communication naturally includes discourse markers (Clark, 2002). Discourse markers refer to elements and the units of talk that are sequentially dependent and primarily drawn from conjunctions, adverbs, and prepositional phrases from a spoken text (Schiffrin, 1987, p.31; Fraser, 1999). Carter and McCarthy (2006) defined discourse markers as the elements of discourse t hat reflect the speaker's intention and activity to "monitor, organize, and manage" the talk (p.208). Pedagogical discourse markers are presented in a pedagogic setting and shown close relations with teaching effectiveness, evidenced by positive associations between the use of various discourse markers and students' comprehension of lectures (Jung, 2006; Beege et al., 2022).

Among the few studies that integrated communication in the classroom with discourse markers (e.g., Othman, 2010; Yu, 2008), Fung and Carter (2007) demonstrated one of the first attempts to provide a comprehensive functional paradigm of discourse markers in a classroom setting. Instructors' intentional language choices are categorized to support the Interpersonal, Referential, Structural, and Cognitive functions (see Fig. 1). First, the interpersonal function concerns the language used to minimize the social distance between the students and instructor. Referential function concerns providing grammatical connectives to the discourse or sentences. Structural functions signal topic shifts and cognitive function denotes thinking processes. Their functional paradigm of discourse markers was adopted to understand non-native EFL teacher talk in primary and secondary school classrooms (Ding & Wang, 2015) and teacher communications in language learning classrooms (Jones, 2011; Yin, 2018). The discourse markers f requently used by non-native EFL teachers were often associated with structural and interpersonal functions (e.g., ok, so, and), which contribute to the pragmatic flow and facilitate student understanding (Flowerdew & Tauroza, 1995).

Unlike the other classroom communication frameworks and guidelines that emphasize the importance of talk moves to promote discussion (e.g., accountable talk or academically productive talk; Michaels et al., 2008), the functional framework introduced by Fung and Carter focuses solely on the teacher's discourse strategies. For instance, the accountable talk framework (Michaels et al., 2008) focuses on analyzing the interaction between the teacher and the student based on their talk moves in a situation where teachers attempt to organize and promote a rigorous discussion environment for collaborative lea rning (Suresh et al., 2019; 2022). Fung and Carter's (2007) discourse framework, on the other hand, provides a guideline that is universally associated with the varying aspects and components of learning, beyond the discussion environment.

2.2. *Evaluating pedagogical communication*

The importance of pedagogical communication has motivated education researchers to examine discourse in classrooms, small-group tutoring, and one-on-one tutoring to provide instructors with pragmatic strategies for effective instructional delivery (Hrastinski et al., 2018; McFarlane, 2016). Continuous, albeit sparse, efforts have been made in the computer-assisted learning and educational technology research community to evaluate what constitutes 'effective' learning. This is measured in terms of instructors' pedagogical communication strategies in computer-mediated online learning environments, such as video lectures. See Table 1 for a more detailed list of references and their key findings.

However, collecting these data often requires countless hours of observation and human coding of discourse, either live or from video recordings of instructional sessions (Suresh et al., 2019). More recently, AI-powered analytic approaches, such as natural language processing (NLP) and machine learning, have been used to systematically analyze the instructional communication strategies in traditional classroom settings (Jensen et al., 2020; Suresh et al., 2019, 2021, 2022).

Supervised Learning. AI-powered methods have been leveraged to detect basic discourse structure (e.g., question, homework),

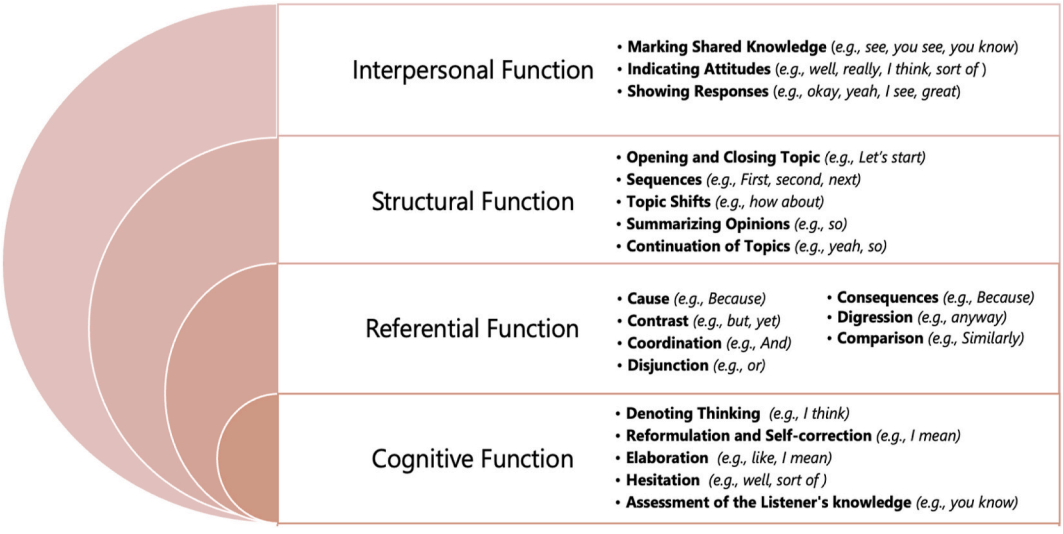


Fig. 1. Core functional categories of pedagogical discourse markers (Fung & Carter, 2007).

Table 1
Summary of the Literature and its Key Findings on Effective and Ineffective Pedagogical Communication Strategies in Online Learning.

Reference	Key Findings	Effective/Ineffective Pedagogical Communication
Johnson and Picciuolo (2022)	The study introduces the EmiBO corpus, and the findings from the corpus, which investigated and compared the verbal- and non-verbal communication in online and face-to-face learning. The study identifies a list of effective communication skills that are different in the two modalities of lectures.	• Use of lexico-grammatical importance markers to highlight key lecture points • Use of personal pronouns for inclusion such as "you" • Variational use of spatial deictics, such as "here" and "this" to help student understand the context of discussion • Use of lexical bundles for language learning when presenting new terminology
Lange and Costley (2020)	The study analyzed student feedback on the challenges of online video courses. The challenges are attributed to both media delivery issues and pedagogical discourse concerns.	• Speaking rate should neither be too fast nor too slow; the suggested speaking rate is 219 (Pastore, 2012) or 254 words per minute (Guo et al., 2014) • Speaking should be intelligible, well-balanced with background noise and/or music • Provide various forms of media that incorporate both audio and visual elements
Dinmore (2019)	The use of multimedia principles is essential for engaging video lectures. Student satisfaction was associated with high quality preproduction and pedagogical practices.	• Only cover a single concept or topic in each video ; Mayer's (2017) "segmenting principle" • Shorter videos promote user-paced segments in following the topics • Articulate each word • Providing various types of material in the video (e.g. desktop capture of interviews, case studies, etc.) with the presenter's recorded voice over aids in addressing student misconceptions of the material • Speak straight towards the camera to develop on-screen persona with supporting slides in the background
Choe et al. (2019)	Identified 8 online lecture video styles that improved student engagement and satisfaction, while maintaining high learning outcomes in online education. The researchers presented different lecture video styles with standardized material to evaluate its effect on student learning outcomes and satisfaction. Video styles that were impersonal and unfamiliar were rated poorly.	• Real-world challenges, case studies, illustrations; consider using the Learning Glass to take students through the process of problem solving; consider using the Demo on select topics. • Mayer's multimedia principles: redundancy, segmenting, modality principles. • Focus on important points, shorter videos are received more positively; ensure good pacing; include illustrations/diagrams to maximize learning. • Reduce awkwardness (or negative affective responses); design engaging lectures; talk directly into the camera and establish a connection with the viewer; use accessible language; create an entertaining product. Ineffective Strategies: • The interview style discourse received the lowest average scores. Students generally felt that the style was uncomfortable and awkward. However, students did see potential learning value from this style, and suggested that it be included as a supplemental questions segment.
Glover (2018)	The study demonstrates in-depth analysis of teacher talk and describes the effectiveness of evaluating teacher talk using the following categories of Elicitation, Responses, and Description from Mercer (1995)	• Use of words like "we" statement, which indicates the shared knowledge and reconstruction of recaps • Use of response strategies to students to Confirm, reject, Repeat, Reformulate, and Elaborate. • Frequent questioning; Raise questions that are cognitively challenging to students; Raise open-ended questions. • Guiding students through cues (e.g., topic shifts) and Clarify uncertainties
Atapattu & Falkner (2017)	Effective lecturer's discourse significantly improves student engagement and learning outcomes in MOOCs video interactions (identified by the pausing and seeking behaviors), as revealed through detailed video learning analytics	• Avoid long sentences unless they include important connectives indicating cause, contrast, or intent. • Decrease the number of sentences in a video, shortening its duration. • Adopt a slower pace of speech. • Favor the use of first-person and second-person pronouns instead of third-person pronouns. • Opt for tangible, specific words rather than abstract ones. • Incorporate content-specific words regularly. • Choose a narrative, casual style of language instead of a formal tone.
Brame (2016)	Effective use of video as an educational tool is enhanced when instructors consider 3 elements: how to manage cognitive load of the video; how to maximize student engagement with the video; and how to promote active learning from the video.	• Use signaling to highlight important ideas or concepts. • Use a conversational, enthusiastic style to enhance engagement. • Embed videos in a context of active learning by using guiding questions and interactive elements. • Keep the videos brief and targeted on learning goals. • Use audio and visual elements to convey appropriate parts of an explanation; consider how to make these elements complementary rather than redundant.

pedagogical functions, as well as specific discourse elements such as instructional questions (Patrick et al., 2016; Jensen et al., 2020; Stone et al., 2019) by leveraging human-labeled classification of discourse components. These approaches evaluated teachers' instructional communication from the transcripts based on the human-labeled data following several theoretical frameworks, such as accountability talk framework.

Specifically, studies have used machine learning techniques to analyze video recordings of teacher-student interactions, small group discussions, and teacher feedback on student writing, and have identified patterns in teacher speech that predict positive student outcomes, such as engagement, achievement, and critical thinking. Recent studies focused on leveraging deep-neural models and large-language models (e.g., transformers) to classify the pedagogical discourse elements without any explicit feature engineering. For instance, Suresh et al. (2019, 2021) investigated the presence of accountable talk, or academically productive talk (Michaels et al., 2010), within teachers' instructional communications in the classroom from a total 501 written transcripts of classroom conversations. Their best-performing model was Roberta-Lar ge (Liu et al., 2019) which resulted in an F1-score of 0.79 in their test set. Balyan et al. (2022) conducted similar analyses with online one-on-one tutoring student-teacher discourse and trained the RoBERTa model using several datasets to understand the pedagogical communication in the mathematics lessons. The system classified communication elements into the three foci of the Accountable Talk Framework (i.e., Learning Community, Content Knowledge, Rigorous Thinking) and achieved an accuracy in the range of 0.89–0.94 for the teacher talk moves and the student talk move model accuracy ranged between 0.80 and 0.86.

Several other approaches focused on engineering linguistic, contextual, and acoustic features to model important elements from pedagogical discourse. Jensen et al. (2020) evaluated a total of 142 lectures and conversations in classrooms using the audiotape and leveraged the voice recognition system to transcribe the pedagogical conversation. They engineered a total of 16 acoustic (e.g., time signal, crossing rate), 13 contextual features (e.g., speech rate), and 37 binary linguistic features (e.g., part-of-speech, named entities) to segment the conversation element into seven discourse types (i.e., instructional talk, questions, authentic questions, elaborated evaluation, high cognitive level, uptake, goal specificity, and ELA terms). Their best-performing random-forest prediction model achieved an average accuracy 0.71 (AUROC = 0.77).

Similarly, other studies have focused on detecting specific pedagogical elements, such as instructional questions (Donnelly et al., 2016; Kelly et al., 2018), and content-specific statements (Stone et al., 2019), with relatively strong accuracy (close to F1 $\sim = 0.70$) in correctly detecting communication elements within traditional classroom settings (Kelly et al., 2018; Stone et al., 2019). For example, Stone et al. (2019) collected audio recordings of 28 students from 162 classroom sessions. Their system extracted a comprehensive list of n-gram-based linguistic and acoustic features to train Random Forest models. The models distinguished non-instructional vs. instructional statements and questions within the teacher discourse with the F1-score ranging from 0.71 to 0.77. Further, Lampos et al. (2021) used a machine learning classifier to predict which type of teacher communication was more likely to generate a positive response from diverse students, such as the students with autism spectrum disorder (ASD). The teachers' communication strategies were coarsely categorized into verbal, gesture, physical, picture- and object-based prompts to evaluate its effectiveness using the Logistic Regression, Random Forest, Gaussian Process models. The results indicated that the classifiers could achieve close to 0.70 accuracy and F1-score in predicting the level of elaboration presented in the students' responses following teachers' communication strategy.

To summarize, previous research has utilized AI and NLP to provide teachers with feedback on their classroom communication, but there is still room for improvement, particularly in identifying detailed instructional and pedagogical elements in teacher communication. To classify and evaluate specific teacher discourse using transformers and supervised learning methods, it is necessary to have a substantial amount of human labeling of discourse markers, which may not always be available and presents a challenge for implementing these techniques at scale. These limitations highlight the necessity for new approaches, which we demonstrated in our study that focuses on disentangling teachers' discourse patterns using statistical n-gram modeling, unsupervised learning, and a theoretical framework on the use of pedagogical discourse markers. Furthermore, previous research has primarily focused on analyzing instructors' linguistic discourse patterns, with little consideration given to teachers' attitudinal and emotional differences. Therefore, a more explicit evaluation is necessary to fully comprehend their impact on teacher-student interactions.

2.2.1. Pedagogical communication in virtual learning environments

Pedagogical communication in a virtual learning environment is key to promoting effective learning (Schneider et al., 2022; Muir et al., 2022). It is al so exceptionally complex due to the multifaceted requirements of communicating effectively in a virtual learning environment. For example, in an online learning environment, instructors are expected to provide social, interpersonal, cognitive, and structural support (Lee, 2010; Conklin & Dikkers, 2021). Effective pedagogical communication is a solution to address each of these. Social and interpersonal interactions in communication have an important role in reducing negative effects resulting from the physical gap in computer-based online learning environments (Atapattu & Falkner, 2017; Rockey et al., 2020). Thus, it is important to understand how the linguistic choices in video-based instruction affect students' feelings of connection, motivation, and engagement.

Guo et al. (2014) analyzed close to 6.9 million video lectures collected from the edX MOOC platform. They focused on understanding various factors that affect student engagement and learning outcomes. They adopted two proxies to evaluate students' engagement, including the total length of time that a student spends on a video and the total number of attempts needed to resolve follow-up questions after the lecture videos. The findings showed that both engagement proxies were significantly higher for more personal lecture videos, such as talking-head videos. Students tended to show increased engagement with lecture videos when the instructors used conversational language with first-person pronouns (e.g., "I", "you", "your"). Similar and consistent findings were provided by recent studies, where the verbal and non-verbal signs that provide interpersonal traits (e.g., eye-contact, Beege et al., 2017).

Richardson (2001) explored relations among students' learning perceptions, course satisfaction, and instructors' social presence as perceived by the students in a virtual learning environment. Results from this study of 99 online students indicated a strong association between students' perceptions of social presence (e.g., students and instructors) and their overall satisfaction with the instructors and the perceived learning outcomes. A recent finding from Beege et al. (2022) indicated the importance of professional communication. Linguistic cues that lead to professional communication involve the accuracy, completeness, and the understandability of the instructors' talk (Liew et al., 2021). Teacher's verbal cues that reflect their professional communication skills were identified to positively affect students' school performance in traditional classroom settings (Spooren & Mortelmans, 2006). Beege et al. (2022) additionally indicated that the professional communication skills with the instructors' professional appearance could positively affect students' intrinsic motivation, social processes, and reducing cognitive load in virtual video lectures. *Yet,* there have been few studies that have systematically analyzed the linguistic features of pedagogical communication in video-based instruction used in virtual learning environments with a broader categories of communication strategies in learning settings.

2.3. The current study and research questions

The widespread access to virtual learning environments has resulted in accumulation of large numbers of transcripts from video-based instruction that can be systematically analyzed to better understand the features of pedagogical communication. However, few studies have demonstrated systematic analytic methods to understand the linguistic features of pedagogical communication in video-based instructions used in virtual learning environments. Hence, the goal of this study is to better understand pedagogical communication in video lectures automatically using computational linguistic tools and natural language processing approaches. We leveraged psycholinguistic tools and NLP to conduct theory-based discourse analysis to effectively analyze the complex and noisy video lecture data. Specifically, we employed Coh-Metrix (Graesser et al., 2004) and SÉANCE (Crossley, Kyle, & McNamara, 2017), to identify the linguistic indices representing each instructor's language choices.

We then examined the individual contribution of specific linguistic indices to identify linguistic dimensions of the pedagogical communication that varied significantly among the instructors. To do so, we constructed a machine-learning classifier which classifies the video based on the tutor using the linguistic indices identified in their communication component. After extracting the variations in pedagogical communication, they were mapped onto three functional categories in Fung and Carter's (2007) pedagogical discourse framework. This framework was used to identify common communication strategies with specific utterance-level examples of uses and adoptions. The goal of mapping utterances to this framework is to provide more specific examples of how the instructors used distinct patterns of pedagogical communications to generate instructional video lectures.

The Interpersonal, Structural, and Cognitive functions from the Fung and Carter's (2007) framework was adopted in our study to understand and analyze the instructors' pedagogical communication in Math Nation video-based instruction. We removed the referential category from the analysis. This is because this category mostly concerned the grammatical aspects of the linguistic functions of linking and transition words (e.g., cause, contrast) rather than communicatory functions. The three categories of pedagogical discourse markers selected for this study also provided a coherent categorization of indicators of intentional language choices previously highlighted as critical aspects of effective pedagogical discourse (instructional clarity; Chesebro & McCroskey, 1998; emotional support; She & Fisher, 2002, and cognitive support; Amadi & Paul, 2017). Hence, adopting the three categories from the guidelines by Fung and Carter (2007) provides a theoretical guide for our analytic process. Two specific research questions are answered in the study.

1) To what extent can natural language processing and machine learning approaches evaluate pedagogical communication patterns in online video lectures?
2) What underlying linguistic, attitudinal, and emotional differences in the instructors' pedagogical communication strategies are depicted in their video lectures?

3. Methods

We implemented a three-stage analytic process to investigate the pedagogical communication within a virtual learning environment (Math Nation) using multiple NLP approaches. This mixed-methods analytic process was designed to investigate the communication content in video lectures by understanding the linguistic and utterance-level instructional discourse within virtual mathematics learning environments.

3.1. Research context and sample

Algebra 1 video lectures (n = 125) that are created by 5 instructors are collected from an online mathematics learning platform, Math Nation. Math Nation provides varying types of learning modules including online practice tools (i.e., online workbook), interactive wall posts, and dynamic video lectures. The video lectures include a foregrounded human instructor (i.e., not a picture-in-picture frame) who uses more friendly language in describing the backgrounded graphics supporting instruction (see Fig. 2). The presentation of graphics and narration align with the principles of multimedia learning based on social cues: Personalization (e.g., friendly vocabulary), Voice (e.g., human), Image (e.g., a human face is present; Mayer, 2005). The graphics in the videos are worked examples, problem text, and animations that are simultaneously displayed with the accompanying narration. The graphics and narration in the videos follow the Redundancy and Modality principles such that learners' cognitive resources should not be overtaxed

by extraneous text, images, or narration. The videos covered a total of 25 topics from three Algebra 1 chapters.

The video content was aligned with the Algebra 1 workbook (Table 2, reference content) which introduced important concepts in Algebra 1 with practice exercises to support student learning. The video lectures constituted three components included by the instructor a) introducing and summarizing the workbook content (i.e., workbook component); b) introducing problem-solving strategies and corresponding algebra concepts (i.e., lecture component); and c) communicating with the audience (i.e., communication component). In Table 1, we provide an example discourse categorized into these three components extracted from one lecture video. Additionally, Math Nation provided algebra glossary resources which contained a list of algebra terms (e.g., quartile, range, rate of change, residual, rational number) with their detailed definitions (e.g., quartile: divides a sorted data set into four equal parts so that each represents ¼ of the data set; 25% of the data within a data set).

To prepare for analysis, 125 video lectures from five instructors were transcribed by human experts. These videos were randomly selected for manual transcription from a total of 465 videos which were generated from 93 topics. The lecture videos were transcribed manually by professionals from an external transcription company and proofread by researchers to ensure the quality of the transcription. Annotation was conducted and provided in the SubRip Subtitle (SRT) format, which is used for captioning videos and presentations. This format is characterized by the inclusion of time codes to indicate when each line of text should be displayed. Nonverbal actions were not signaled in the transcript, and standard punctuation systems were used. The 125 video lectures analyzed in this study stemmed from 3 units which comprised a total of 25 topics. The three units included "Using expression to represent real-world situations" and "Solving quadratic function" and "Exponential function". Each unit included up to 10 topics as presented in Table 3.

The lengths of the videos varied noticeably between instructors across the different topics with the longest video in Topic 1 [*Using expressions to represent real-world situations*] for instructors 2, 3, and 4 and the longest video in Topic 12 [*Solving quadratic equation by factoring*] for instructors 1 and 5 (Fig. 3).

Despite the fact that they are covering the same workbook content, the video lectures varied quite noticeably (e.g., length) mainly due to differences in the instructors and their discourse styles. The instructors' varying discourse styles refer to their problem-solving styles and international language choices to better connect with the students. In our study, we analyzed such pedagogical *communication components* which are geared towards motivating, guiding, inspiring, encouraging.

3.2. Research procedure

We introduced a three-stage analytic process. In the first stage, we extracted the communication components (or content) from the lecture transcripts. The communication content extracted from stage 1 served as an important focal point for analysis in stages 2 and 3. In stage 2, we investigated the overarching linguistic differences in pedagogical communications using two advanced psychometric tools, Coh-Metrix (Graesser et al., 2004) and SÉANCE (Crossley et al., 2017). We evaluated and compared the linguistic and sentimental differences between the instructors' communication components. In stage 3, we evaluated the utterance-level differences between the instructor's communication content with empirical evidence and their corresponding pedagogical functional categories. The three processes complemented each other by reducing the noise (e.g., variation of words, part of the transcript that are not related to the communication components) in the data to identify the important content for analysis (stage1), evaluating the overall differences in their linguistic differences (stage 2), and extracting and comparing the utterance-level patterns in the communication content and its theoretical functional category in pedagogical discourse settings (stage 3). Fig. 4 provides a conceptual representation of our analytic process.

3.3. Data analysis

3.3.1. Stage 1: communication component detection

We first categorize the sentences collected and transcribed from the video lecture, namely into the following three components: workbook, lecture, and communication (refer Table 1). First, we implemented an edit distance-based approach (Ristad & Yianilos, 1998) to filter the workbook component (see Table 1) from the video transcripts. The edit distance-based approach computes the minimum number of edits required to merge two sentences to evaluate their similarities. Two sentences (i.e., one from the workbook and the other from the video transcript) are compared based on the minimum number of characters inserted, deleted, and changed to convert one sentence to another. Hence, the specific parts of the video transcripts indicating strong evidence of overlap with the Algebra 1 workbook contents (i.e., conventional cut-off = 0.60) were categorized as workbook contents filtered out in this stage. For instance, in Table 1, the reference sentence from the workbook "The Florida Turnpike is a 312-mile stretch of highway running from Central Florida to South Florida" was compared to one of the sentences from the video transcript "It is a 312-mile stretch of highway running from central Florida to South Florida.". The edit distance between the two sentences results in 0.96, which is significantly above the conventional cut-off of 0.60. Hence, the sentence from the transcript is categorized as a workbook component.

The next filtering process was to identify the lecture component. Statistical language models learn to assign probabilities to a new or an unseen text based on the text that was used to train the models (Bengio et al., 2000). In this approach, the probability of the entire word sequence of the workbook is decomposed using the chain rule of probability, where X_i represents the i-th word in the workbook document (see Equation (1)). The n-gram approximation is applied using a bigram (i.e., a consecutive sequence of two words; "I and", "and You", "You Here") model to predict the consecutive words' conditional probabilities, $P(X_{1:k-1}) \approx P(X_{k-1})$. The bigram probability was approximated using the maximum likelihood estimation by counting the occurrence of words in the document, $P(X_{k-1}) = \frac{C(X_{k-1}, X_k)}{C(X_{k-1})}$. Laplace (or add-one) smoothing (Zhai & Lafferty, 2017) was used to avoid problems with words that have never been

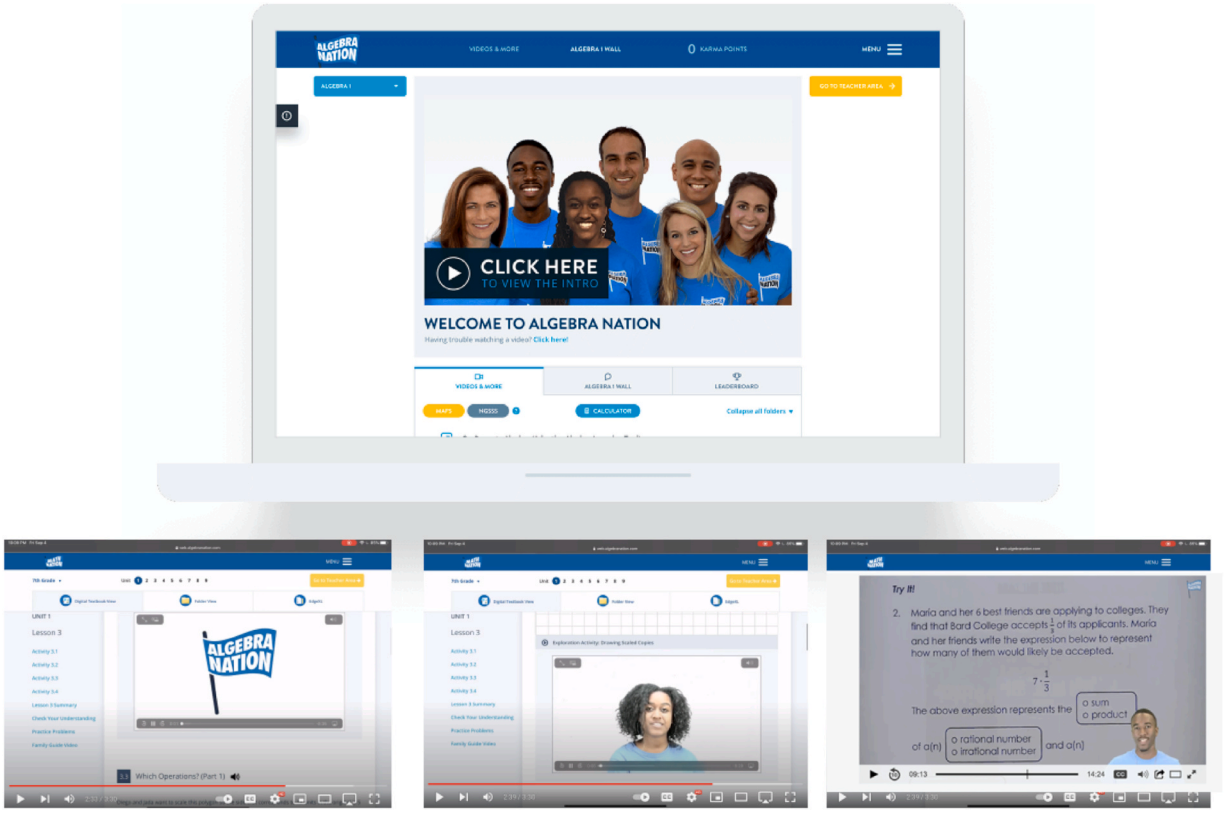


Fig. 2. Screenshot of the Math Nation lecture videos. *Retrieved from* https://www.youtube.com/watch?v=vZinESrnqak, https://www.mathnation.com.

introduced to the model such that each count with the Laplace smoothing is identified as, $P_{Laplace}(X_k) = \frac{C_k+1}{N+V}$, where V represents the total number of unique words, and the N represents the length of the document defined as the total number of words.

$$P(X_1 \ldots X_n) = P(X_1)P(X_1)P(X_{1:2}) \ldots P(X_{1:n-1}) = \prod_{k=1}^{n} P(X_{1:k-1}) \tag{1}$$

A lower perplexity score indicates that the training and the test documents (or corpus) contain similar word-sequences. The text with a low perplexity score was determined as the lecture component and the remaining text was considered to constitute the *communication* component – which is the focus of our study. The final perplexity scores were compared to the overall average of the text perplexity scores to identify the lecture components. The perplexity is calculated for a language model on a test set, in our case each sentence of video lecture transcript: $Perplexity(X_{test}) = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i|w_{i-1})}}$.

In this study, we trained unigram (i.e., a word; "I", "You", "Here") and bigram (i.e., a consecutive sequence of two words; "I and", "and You", "You Here") language models with the algebra glossary resource provided by Math Nation (see Methods). In other words, the language model is trained to learn the specific probabilities of how unique sequences of words (i.e., unigrams, bigrams) appear in the given text, in our case the glossary resource. The language model trained using this resource would be sensitive towards capturing the lecture component.[1]

3.3.2. Stage 2: linguistic differences in the communication component

Two psycholinguistic tools, Coh-Metrix (Graesser et al., 2004) and SÉANCE (Crossley et al., 2017), were used to identify linguistic differences between the instructors' communication strategies. We used the tools to extract an extensive list of psycholinguistic indices to quantify such differences. The indices were used to represent overarching linguistic dimensions present in the communication component of video lectures. *Coh-Metrix is a computer tool that can analyze English text and discourse on various linguistic and psycho-linguistic measures of cohesion, readability, and language. It is widely used by researchers in the areas of applied linguistics, English education,*

[1] The complete code for Stage 1 is available on GitHub (link omitted for blind review).

Table 2
Example transcript categorization used to extract the communication component.

Reference Content (Workbook)

Section 1: Expressions

Section 1 – Topic 1:
Using Expressions to Represent Real-World Situations

The Florida Turnpike is a 312-mile stretch of highway running from Central Florida to South Florida. Drivers who use a SunPass pay discounted toll rates. For 2-axle passenger vehicles with a SunPass, the toll at the Orlando (I-4) Mile Post is $0.53, and the toll at the Okeechobee Plaza exit in West Palm Beach is $1.04.

a. A 2-axle passenger vehicle with a SunPass drives through the Orlando I-4 Mile Post five times. Determine the total amount of money that will be deducted from this vehicle's SunPass account.

b. Create an algebraic expression to describe the total amount of money that will be deducted from this vehicle's SunPass account for any given number of drives through the Orlando (I-4) Mile Post.

c. The same 2-axle passenger vehicle drives through the Okeechobee Plaza exit in West Palm Beach seven times. Determine the total amount of money that will be deducted from its SunPass account.

d. Create an algebraic expression to describe the total amount of money that will be deducted from this vehicle's SunPass account for any given number of drives through the Okeechobee Plaza exit.

e. Write an algebraic expression to describe the combined total amount of money that will be deducted from this vehicle's SunPass account after it passes through the Orlando (I-4) Mile Post and the Okeechobee Plaza exit any given number of times.

Video Lecture Transcript	Workbook Component	● "It is a 312-mile stretch of highway running from central Florida to South Florida."
	Lecture Component	● "2 and 3- the only thing they would have in common is a 1.", ● "These are different variables." ● "So, it is a normal conversion."
	Communication Component	● "You guys can do this.", ● "This might be difficult to understand.", ● "What do you think?"

*and language psychology to analyze various English texts (*e.g.*, Jeon, 2014). Joksimović and colleagues (2019) used Coh-metrix to analyze linguistic features of students' annotations from an integrated video annotation tool to evaluate students' self-reflections of their learning from video lectures. SÉANCE, on the other hand, is a different tool developed to analyze student-teacher interactions in educational settings. It uses machine learning algorithms to identify patterns and topics in audio recordings of classroom interactions. SÉANCE reports on indices specifically related to sentiment, social cognition, and social order. Measuring sentiment features in a text can help better understand how emotions, opinions, affect, and feelings influence cognition and learner engagement. Therefore, we have used this tool for analyzing student-teacher interactions in educational settings.*

Specifically, Coh-Metrix extracts a comprehensive list of linguistic features at word and sentence levels, providing rich data regarding word information, syntactic patterns, syntactic complexity, situation model, lexical diversity, and other descriptive indices (Graesser et al., 2004). Similarly, SÉANCE (Crossley et al., 2017) provides semantic information about words in a text using a set of pre-existing sentiment, social positioning, and cognitive dictionaries. In this study, we specifically focused on component scores from SÉANCE. The component scores measure word polarity scores that are associated with a total of 20 sentiment categories.

A Random Forest (RF) classifier algorithm was then constructed to assess and compare the instructors based on the extracted indices. Specifically, we investigated the extent to which the psycholinguistic features (Coh-Metrix and SÉANCE) differentiated between the instructors. A random forest classifier is a supervised learning method that makes classification by combining multiple decision trees to increase the overall performance (Liaw & Wiener, 2002). The RF classifier also attempts to identify the best features among the random subset of variables using the feature importance indices, which help identify the relative contribution (out of 100%) of each feature to derive the classification output. Preprocessing and feature selection entailed addressing assumptions of normality, multicollinearity, and validity of the features. We removed features with zero variance and highly correlated features ($r > 0.90$) resulting in a final set of 25 linguistic features. These features included 11 Coh-Metrix linguistic measures of word information, incidence of noun phrases and prepositional phrases, situation model, syntactic structure, and text easability (narrativity and

Table 3
Algebra Topics introduced in the Algebra Nation Video Lectures.

Unit	Topic
Algebraic Expression	[1] Using expressions to represent real-world situations.
	[2] Understanding polynomial expressions.
	[3] Algebraic expressions using the distributive property.
	[4] Algebraic expressions using the commutative and associative properties
	[5] Properties of exponents.
	[6] Radical expressions and expressions with rational exponents.
	[7] Adding expressions with radicals and rational exponents.
	[8] More operations with radicals and rational exponents.
	[9] Operations with rational and irrational numbers.
Quadratic Equation	[10] Real-world examples of quadratic functions
	[11] Factoring quadratic expressions
	[12] Solving quadratic equations by factoring
	[13] Solving other quadratic equations by factoring
	[14] Solving quadratic equations by factoring – special cases
	[15] Solving quadratic equations by taking square roots
	[16] Solving quadratic equations by completing the square
	[17] Deriving the quadratic formula
	[18] Solving quadratic equations using quadratic formula
	[19] Quadratic functions in action
Exponential Functions	[20] Geometric Sequences
	[21] Exponential Functions
	[22] Graphs of exponential functions- part 1
	[23] Graphs of exponential functions- part 2
	[24] Growth and decay rates of exponential functions
	[25] Transformations of exponential functions

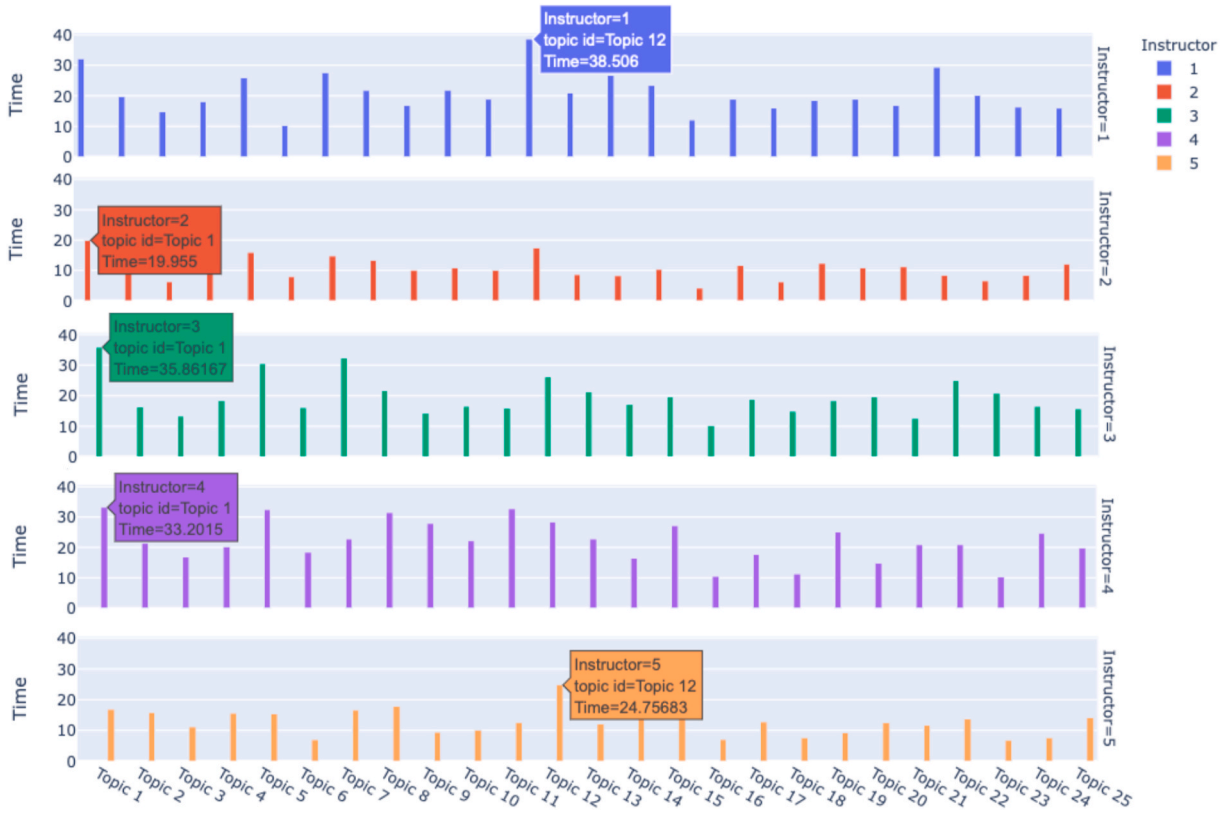


Fig. 3. Total and the average length of the video lectures across 25 Algebra 1 topics.

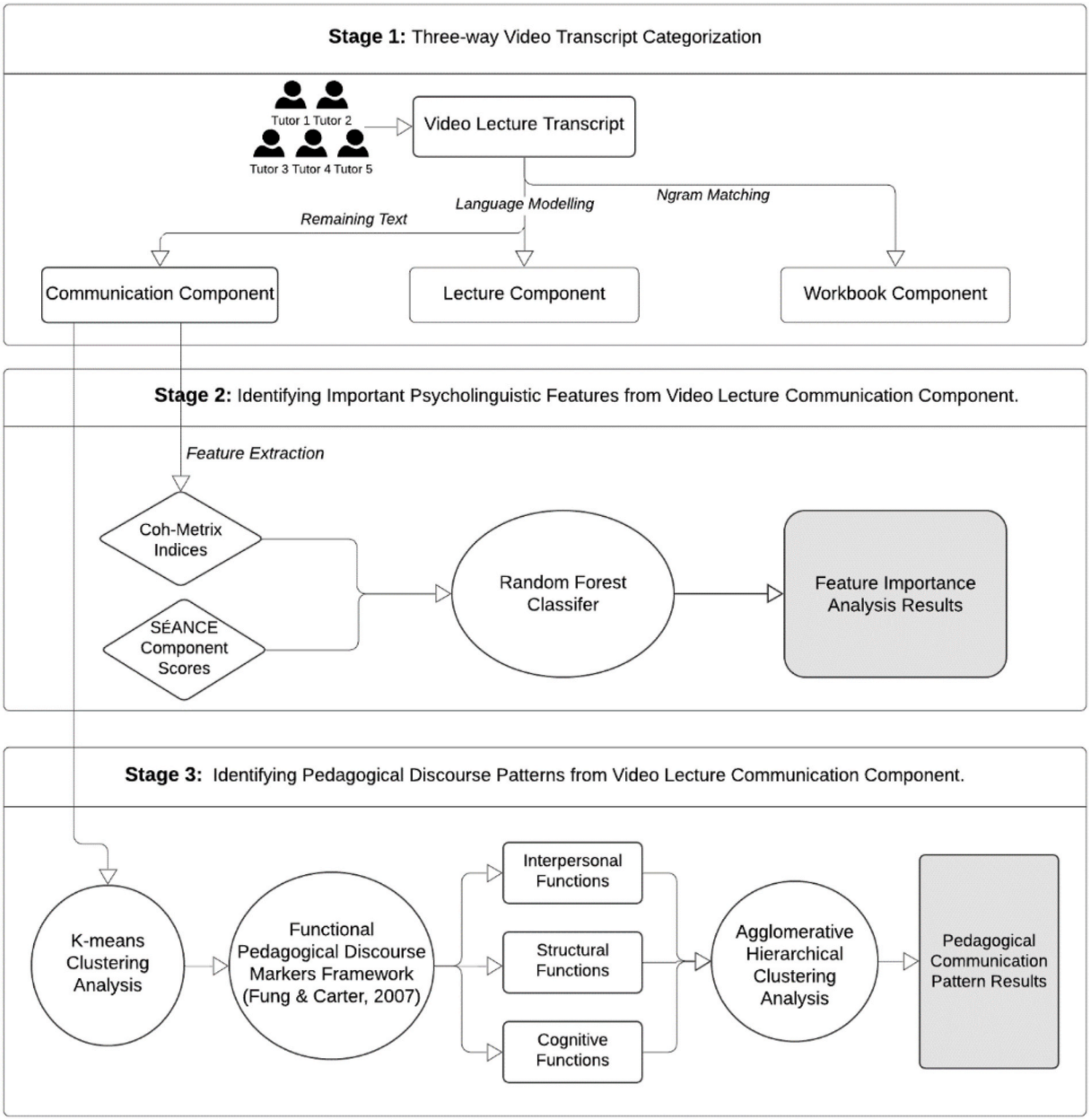


Fig. 4. A conceptual representation of the current study's three-stage analytic process.

connectivity components); and 14 SÉANCE sentiment component indices. Additionally, we computed the permutation-based feature importance analysis to iteratively remove the features that did not contribute to the performance accuracy (see Altmann et al., 2010 for more information about feature importance approaches). We provided a detailed description of the final input variables in Appendices A and B.

Our final random forest model was implemented using Python 3.6 and had optimal parameter settings of the number of estimators of 1,800, minimum sample split of 2, minimum sample leaf of 2, maximum depth of 80. The final average performance results and the importance of features for each model were computed using a 5-fold cross-validation (CV) approach wherein the dataset with a total of 25 features was randomly parsed into five folds (or blocks). Each fold was iteratively used to test the accuracy of the classification model while the remaining four folds were used to train the classification model.[2]

[2] The complete code for Stage 2 is available on GitHub (link omitted for blind review).

3.3.3. Stage 3: pedagogical function category classification

In the last stage, we focused on extracting the pedagogical discourse strategies at the utterance-level from the communication component. This stage allows us to disambiguate the specific language or word choices that are made by the instructors for clear pedagogical communication.

First, we applied a data-driven approach to extract statistical clusters of common discourse patterns from the communication component. We then recategorized the common discourse clusters based on their core pedagogical functions following the guidelines by Fung and Carter (2007). This step was conducted manually by comparing the examples provided in each pedagogical functional category by Fung and Carter (2007) with the cluster outcomes. For example, the list provided in Fung and Carter (2007) shown in Fig. 1 included Marking Shared Knowledge – "see", "you see", and "you know"; Opening and closing – "let's start"; Denoting Thinking – "I think". We manually compared the discourse clusters that we extracted with these example discourse markers to identify the best categorization of the clustering results to their functional discourse marker categorization guidelines.

K-means clustering algorithm with two text-vectorization approaches, namely word-count and term frequency-inverse document frequency (TF-IDF; Aizawa, 2003) with different n-gram ranges (1–5; i.e., sequence of *n* number of consecutive words appearing in the text), were used to generate meaningful pedagogical discourse marker clusters. The TF-IDF method transforms a sentence to a numeric vector based on the word count. The method assigns lower weights to the word that appears frequently across all sentences while providing high weights to the word that appears in particular sentences to highlight the unique semantic contribution of each word in the document.

The optimal number of clusters was identified using cluster quality measures (elbow method with cluster inertia; Bholowalia & Kumar, 2014) and the interpretability of each cluster. Lastly, the pedagogical discourse marker results are structured based on its function following the framework by Fung and Carter (2007; Fig. 1). This augmented theoretical approach integrating human experts' judgements with empirical findings from the NLP analysis benefited the study in two ways: (1) reducing the number of individual clusters using overarching thematic categorization and (2) grounding the empirical findings with a theoretical support for more meaningful interpretation.[3]

The evaluation of clustering results through qualitative methods has been shown to enhance the understanding, interpretability, and explainability of the extracted clusters or topics, as noted by Qiang et al. (2020) and Alves et al. (2021). In particular, previous learning analytics research has utilized clustering analysis and mapped the results based on the theoretical framework of students' self-regulated learning strategies (Maldonado-Mahauad et al., 2018) and engagement, utilizing a convergent mixed-method approach (Wilson et al., 2021). By employing such a process, the clustering results are interpreted based on the pedagogical framework and the theoretical guidelines, allowing for appropriate conclusions to be drawn from the findings. Moreover, this process ensures rigor in the unsupervised method's findings, which unlike the supervised method, does not rely on the ground truth (labels of communication discourse markers) for categorization.

Last, the two-way ANOVA analysis was conducted to identify and compare the patterns of pedagogical discourse and their functions adopted by different instructors in lecture videos.

4. Results

4.1. Stage 1 results: communication component detection

On average, approximately 18%, 59%, and 22% of the video lecture transcripts were identified to be associated with the workbook, lecture, and communication components respectively (see Table 4). To further evaluate and validate this categorization approach, we visualized the distribution of the three components across the 25 algebra topics presented by each Algebra 1 instructor. The cosine similarity results showed a strong indication of systematic patterns in the use of workbook (average similarity = 0.845) and lecture (average similarity = 0.788) with less systematic pattern in communication components (average similarity = 0.225). The strong similarities of the observed patterns can be attributed to the similar content and key terms that are embedded in the workbook and glossaries embedded in Math Nation's algebra curriculum as used by the instructors. Moreover, we observed that as a topic matures within the three algebra topics the instructors introduced fewer workbook components, and instead introduced more lecture components (Fig. 5)

These tendencies emerge, in part, due to the design of the workbook, wherein each algebraic concept starts with topics introducing real-world problems. Such topics inherently carried longer texts to introduce problem situations. Towards the end of each concept, however, problem statements tended to be simpler with more abstract algebraic expressions, leading to significant decrease in workbook features, and an increase in lecture components. Despite such changes in patterns, we did not identify systematic patterns in the use of communication components across topics.

Post-hoc qualitative evaluation of the sentence categorization was conducted on randomly selected subsets of the video transcripts to verify face validity of the categories. This post-hoc evaluation pointed to the accurate categorization, i.e., sentences categorized as communication, lecture, and workbook were confirmed as such.

[3] The complete code for Stage 3 is available on GitHub (link omitted for blind review).

Table 4
Percentage of sentences from the video transcript categorization by instructors.

Instructor	Workbook	Lecture	Communication
1	16.1%	62.2%	21.7%
2	19.7%	58.7%	21.6%
3	15.5%	59.8%	24.7%
4	19.7%	58.7%	21.6%
5	21.4%	56.4%	22.0%
Avg.	**18.5%**	**59.1%**	**22.3%**

4.2. Stage 2 results: linguistic differences in the communication component

The linguistic differences between the instructor's communication components were characterized by a total of 25 features (see Tables 5 and 6). . These features could differentiate the Algebra 1 instructors based on their communication pattern (Random Forest classifier accuracy = 86.7%, 5-fold CV). We presented relative importance (out of 100%) of each feature in the classification models. The indices from the Coh-Metrix categories: Word information (i.e., WRDPOLs, WRDVERB, WRDPOLc), Text easability (i.e., PCCONNp, PCSYNz, PCVERBz, PCNARz), and Syntactic structure (i.e., DRPP, CNCTemptx, DRNP), and Situational model (SMINTEp) varied among the instructors.

The word information index specifically showed the highest contribution to classifying the instructors based on their communication components in their video lectures. The frequency of using first-person plural pronouns (e.g., "we" and "us"), the polysemy for content words, and the verb incidence score were identified as important indicators. The ambiguity of language associated with possible varying interpretations are often related to the polysemy for content words. In other words, instructors adopted communication patterns that vary in terms of its levels of ambiguity in their pedagogical language use.

The easability of the communication (i.e., narrativity, verb cohesion, syntactic simplicity, and connectivity) was an important factor to distinguish instructors' communication patterns. Also, the concentration of syntactic patterns in the communication components varied significantly among the instructors. Lastly, the intentional content feature from the situational model category was identified to vary significantly among the instructors. This feature represents the "intentional actions, events, and particles" in the text (Crossley et al., 2017).

Table 6 presents the final SÉANCE indices identified as important features that are varying among the instructors. These indices communicate how the instructors differentially communicate with the students which may convey specific sentimental effect to guide students e.g., encouraging students with high certainty words. Especially, the frequency of adopting particular adverbs indicated noticeable contributions to the instructor communication pattern classification performance outcomes. These adverbs were often associated with attributes such as virtue, certainty, and failure.

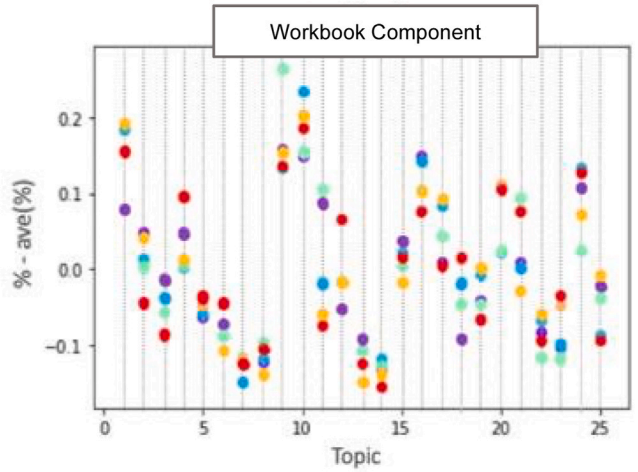
4.3. Stage 3 results: pedagogical function category

We evaluated a total of 4962 sentences (or discourse) retrieved from the lecture video transcripts. K-means clustering results identified 17 pedagogical discourse marker clusters from the communication components (K = 17, inertia (or sum of squared distances) = 153.17, n-gram: 1–5). TF-IDF vectorization was adopted in the final clustering analysis model. More specific information regarding the discourse markers is described in Table 7. The results indicated that instructors' discourse patterns tended to vary suggesting they made intentional word choices aimed at achieving effective pedagogical communication. Analysis revealed dominant categories such as greeting (Cluster 6: "Hi everyone", "Hey Algebra Nation"), encouraging participation (Cluster 3: "Let's see", "Let's jump in", "Let's take a look"), short correct/incorrect responses (Cluster 1: "That is correct", "Yes", "No").

4.3.1. Theoretical mapping with the core functional category

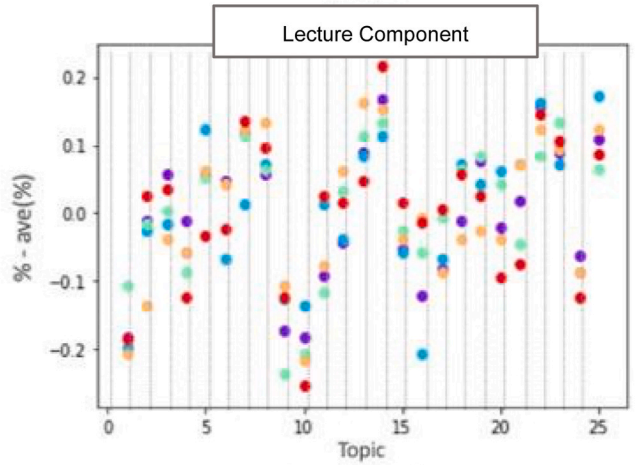
The theoretical mapping results using the functional framework (Fung & Carter, 2007) showed that the interpersonal functions appear as the primary function of pedagogical communication in video lectures. Clusters 12, 14, and 16 supported the interpersonal functions that are specifically associated with sharing knowledge components in lectures. Cluster 13 served the interpersonal function by providing instructors' attitudes about the topic explicitly. Lastly, Clusters 1, 4, 5, 8, and 9 supported interpersonal function by providing immediate feedback and responses to the students. Similarly, the cognitive function was supported by the Clusters 10 and 15 (i.e., Denote thinking processes), and Cluster 2 (i.e., Assess students' knowledge). The structural function was supported by a total of four clusters. These concerned the structural functions of Opening and Closing Topics (Cluster 6 and 7) and Shifting Topics (Cluster 3 and 11). Table 8 provides a distribution of the use of pedagogical discourse functions for each instructor.

Additionally, the results of a two-way ANOVA indicated a statistically significant difference in how frequently the different pedagogical functions are adopted and present in the communication components ($F = 2.507$, $p\text{-value} = .036$, $\eta^2 = 0.259$; Table 9). More specifically, discourse markers in the Structural Function 1 (Opening and Closing Topics) and the Interpersonal Function 3 were most frequently adopted in the communication components. On the other hand, the frequency of adopting pedagogical functions showed no significant interaction effect with the instructors. This indicates that the patterns (or the proportions) of adopting varying pedagogical functions showed no differences among the five instructors. In other words, instructors exhibited consistent patterns of adopting pedagogical functions, with the dominant categories of discourse markers within the Structural and Interpersonal functions



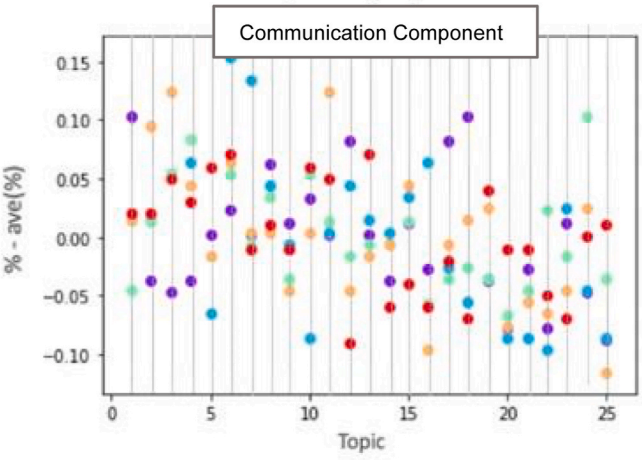
Instructor	1	2	3	4	5
1	1				
2	0.87	1			
3	0.88	0.85	1		
4	0.82	0.93	0.81	1	
5	0.76	0.87	0.78	0.86	1

Note. Cosine similarity between the '**workbook**' component distribution.



Instructor	1	2	3	4	5
1	1				
2	0.83	1			
3	0.87	0.72	1		
4	0.84	0.73	0.79	1	
5	0.79	0.66	0.81	0.77	1

Note. Cosine similarity between the '**lecture**' component distribution.



Instructor	1	2	3	4	5
1	1				
2	0.24	1			
3	-0.13	0.26	1		
4	0.15	0.40	0.57	1	
5	-0.14	0.06	0.35	0.48	1

Note. Cosine similarity between the '**communication**' component distribution.

Fig. 5. Cosine similarities among the instructors in terms of their time used in the three instructional components.

with relatively less adoption of the Cognitive functions. In sum, algebra instructors showed similar patterns and frequency of adopting pedagogical discourse for effective communication, while there are systematic differences in adopting certain categories of pedagogical functions (Table 9 and Fig. 6).

5. Discussion

Intentional language choice in video lectures is essential because it helps encourage more effective and efficient training for online

Table 5
Final Coh-Metrix Feature Indices and Categories and the Feature Importance in the RF model.

Category	Index	Description & *Importance		Implications
Situation Model	SMINTEp	Intentional content	3.7%	Instructors' use of intentional language significantly identifies the varying teacher-talk strategies among the algebra instructors.
Syntactic Pattern Density	DRPP	Preposition phrase density	3.5%	Instructors' use of proportions of particular word classes (e.g., prepositions, temporal connectives, and nouns) in the communication components.
	CNCTempx	Expanded temporal connectives incidence	3.7%	
	DRNP	Noun phrase density	4.6%	
Text Easability	PCCONNp	Connectivity	2.6%	Instructors' use syntactic and vocabulary choices that affect the difficulty, or easability to comprehend the communication components
	PCSYNz	Syntactic simplicity	3.2%	
	PCVERBz	Verb cohesion z score	3.4%	
	PCNARz	Narrativity	3.7%	
Word Information	WRDPOLc	Polysemy for content words	2.6%	Instructors' tendency to use first person plural, such as "we" and "us" encourages collaborative participation
	WRDVERB	Verb incidence	4.3%	
	WRDPRP1p	First person plural pronoun incidence	8.9%	Instructors use content words and verbs with different levels of ambiguity

Note. *Importance score refers to the relative contribution (effect) of each feature in predicting the outcome in the model. Higher feature importance indicates that the variable showed significant contribution to association with the dependent variable (see Saarela & Jauhiainen, 2021 for more information).

Table 6
SEANCE Component score and the Feature Importance in the RF model.

Component	Description and Importance	
Virtue adverbs	Nouns related to worship and forgiveness (e.g., *blame, conviction)*	7.0%
Certainty	Words indicating "a feeling of sureness, certainty, and firmness" (e.g., *absolute*)	5.8%
Failure	Verbs indicating power decreasing (e.g., *attack, banish, beat*)	3.9%
Economy	Words related to economy (e.g., *account, affluence, agricultural*)	3.6%
Politeness	Words concerned with a tools or forms of invoking formal power	3.5%
Negative adjectives	Negative adjectives (e.g., *abject, abrasive, abominable*)	3.5%
Positive nouns	Positive nouns (e.g., *acceptance, abundance, accomplishment*)	3.3%
Trust verbs	Words describing the state of trust (e.g., *admire, adore, anticipate*)	3.1%
Social order	Words concerning social order (e.g., *adhere, appall, corrupt*)	2.7%
Positive adjectives	Positive adjectives (e.g., *absorbent, able, accessible*)	2.5%
Joy	Words related to joy (e.g., *superb, affection, complement*)	2.4%
Affect Friends, Family	Nouns related to friends and family (e.g., *company, friend)*	2.4%

Table 7
K-means clustering results for the pedagogical discourse markers.

Cluster	Key Phrases (Discourse Markers)	Total Discourse
1	"That is true", "That is incorrect", "Yes", "No"	386
2	"Think about this", "What about this", "You know what?"	136
3	"Let's see", "Let's jump in", "Let's take a look"	855
4	"That is good", "That looks good", "Awesome job"	312
5	"Ok", "That is cool", "Cool", "All right", "All right", "Fantastic"	1459
6	"Hi everyone", "Hey Algebra Nation"	128
7	"Thank you", "I will see you next time", "See you in the next video"	140
8	"There you go", "Here you go", "Just like that", "There it is"	190
9	"You got this", "You guys are smart", "Keep studying hard"	178
10	"We are going to", "You are going to", "You can", "We can", "We will"	325
11	"Pause the video", "It is up to you", "You can do either way"	214
12	"We are close", "We are almost there"	45
13	"This is very important." "You should be careful." "This is a bit tricky."	98
14	"We know that", "We have", "We know what to do"	145
15	"Same thing here", "Same idea here"	131
16	"That is what is up", "That is how we do", "That is what is happening",	188
17	"Sorry", "Sorry about this"	32

instructors. The physical distance between the instructors and the students has been identified as a key element that can negatively impact students' motivation and engagement (Lee & Rha, 2009; Martin et al., 2022). Students often feel isolated and left out without immediate guidance and active interactions with the instructors. Currently limited studies in computer-based instructions are focused on understanding instructors' communication patterns in video lectures. Hence, the goal of this study is to better understand pedagogical communication in video lectures automatically using computational linguistic tools and natural language processing approaches.

Table 8
Distribution of the use of pedagogical discourse functions by instructors.

	Instructor				
	1	2	3	4	5
Interpersonal Function 1 [Marking shared knowledge]	3.82%	4.45%	8.65%	7.64%	5.26%
Interpersonal Function 2 [Indicating Attitudes]	1.04%	0.45%	3.26%	0.80%	1.82%
Interpersonal Function 3 [Showing Responses]	35.33%	49.33%	45.52%	38.54%	39.06%
Structural Function 1 [Opening and Closing Topics]	6.19%	4.23%	4.01%	4.30%	2.47%
Structural Function 2 [Topic Shifts]	23.16%	19.27%	9.78%	14.01%	24.79%
Cognitive Function 1 [Denoting Thinking Processes]	7.65%	6.24%	8.40%	9.16%	4.40%
Cognitive Function 2 [Assessing Knowledge]	2.99%	2.67%	2.38%	1.35%	1.50%
Unknown	19.82%	13.36%	17.99%	24.21%	20.71%
Total (%)	100%	100%	100%	100%	100%
Total N Discourse Markers	1438	898	1595	1256	923

Note. The unknowns represent the discourse clusters which could not be directly mapped to the functional framework and were omitted from the further analysis.

Table 9
Two-way ANOVA results.

	df	Mean Squares	*F* (*p*-value)	η^2
Instructor	4	2155.862	0.300 ($p = .876$)	0.027
Pedagogical Function	6	18,022.875	2.507 ($p = .036$*)	0.259
Instructor × Pedagogical Function	23	869.731	0.121 ($p = 1.000$)	
Error	43	7189.256		

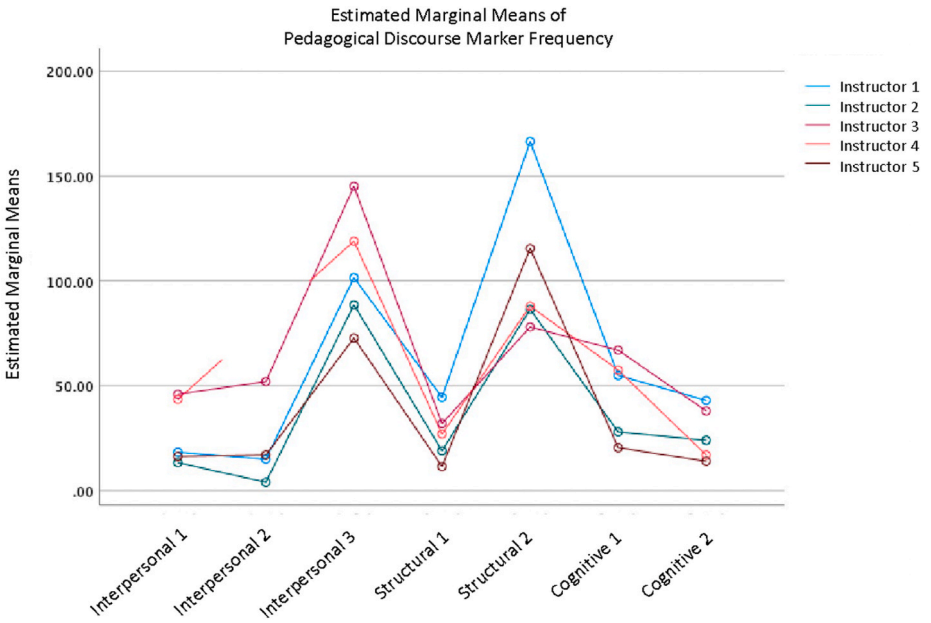


Fig. 6. Estimated marginal means of pedagogical discourse marker frequency.

Our study illuminates the essential nuances of instructor's communication strategies in online video lectures and reveals how they could vary among the instructors and may influence students' learning experience. It uncovers how a blend of intentional language choice, the use of varied discourse markers, and the incorporation of different linguistic cues in video lectures can contribute to effective online teaching. Two research questions were addressed to guide the study and the important summary of our findings are presented below.

RQ 1. To what extent can natural language processing and machine learning approaches evaluate pedagogical communication patterns in online video lectures?

Overall, the findings indicated that the instructors represented in the Math Nation videos used various linguistic cues and

pedagogical discourse functions to build rapport and establish social presence, convey sentimental connections, and challenge students with cognitively activating questions – which is consistent with other studies (see Belt & Lowenthal, 2023; Kim & Thayne, 2015; Wong et al., 2022). The findings from Stage 1 demonstrated that natural language processing techniques could analyze the video lectures and detect the communication component. As expected, instructors tended to show similar patterns of adopting the workbook and lecture components in their lecture videos. The lecture content was consistently dominant across all instructors in the video lectures (56.4%–62.2%). Workbook content showed relatively small proportions of 15.5%–16.1% of the video transcript across the five instructors. On average, close to 23% of the video transcripts were categorized as the communication component. The findings from Stage 1 allowed the following two stages to focus solely on the communication component from the noisy video lecture transcript data.

RQ2. What underlying linguistic, attitudinal and emotional differences in the instructors' pedagogical communication strategies are depicted in their video lectures?

The findings from Stage 2 investigated how the communication content extracted in Stage 1 varied among instructors in terms of various linguistic dimensions. Our findings reinforced suggestions communicated in the literature about pedagogical communication patterns. Additionally, our findings illustrated how the attitudinal and emotional components of instructors are incorporated within their communication attributes, such as virtue, certainty, and failure. For instance, the use of virtue adverbs included key indices from the Lasswell rectitude gain, sureness adverbs, and the GI concerns for hostility adverbs (e.g., "above," "absolute," "account," "actual," etc.). This is consistent with the conclusions of the previous literature. For example, Mazer and Hunt (2008), as well as Wiebe and Kabata (2010), demonstrated that instructors' use of positive language, such as positive slang during lectures, fosters an engaging and participatory learning environment. Consistent findings and support have been provided in the recent literature, where the positive discourse in teacher talk was positively associated with students' learning interest (Anggoro & Ningsih, 2023), academic performance (Soriano & Soriano, 2022), and improved motivation for learning (Mirivel et al., 2022).

Zhu (2023), based on the findings from 10 lecture transcripts analyzed in-depth, indicated that more than half (55%) of the teacher talk during lectures encompasses teachers' attitudes. This finding highlighted the importance of providing appropriate instructor attitudes during the lecture component. However, it was limited and did not focus on which attitude might be the most effective for student learning and engagement. Overall, previous literature has advocated for a discourse style characterized by supportive and positive language to create a more welcoming learning environment and boost student satisfaction, motivation, achievement, and self-regulated learning strategies (Perry, 2023). Although more focused research may be necessary, these studies collectively underscore the influence of linguistic choices in shaping the effectiveness and inclusivity of online education.

Algebra Nation instructors, in our dataset, who showed the highest component scores in virtue adverbs showed frequent use of words that are related to sureness (i.e., degree of assurance, certainty, and confidence in the message), such as "So that is actually very important", or "so you have got to be very, very certain of what you are looking for", "but make sure you look very carefully." Now, once again, as I said, you want to keep making sure you keep practicing on these." These examples reflect the characteristics of the video lectures in a virtual learning setting, where the instructors attempt to convey the instructional message by repeating and emphasizing ideas to provide improved clarity (Civikly, 1992; Horn, 2022).

A total of 17 different clusters of pedagogical discourse markers were identified from a total of 4962 discourse incidences collected from the instructors' communication in algebra video lectures. Interpersonal function was the prominent pedagogical discourse function serving the discourse markers. This is consistent with the important psycholinguistic features that were identified in our second stage of the analysis (i.e., from Coh-Metrix and SÉANCE). For instance, the instructors' use of first-person plural pronouns (WRDPRP1p), such as "we" and "us", could be associated with the instructors' pedagogical discourse tendency to encourage collaborative participation and reduce the gap between the students and instructor (Pennebaker, Booth et al., 2007). This finding aligns with Atapattu and Falkner (2017), who found that instructors who used more interpersonal discourse markers and focused on establishing a collaborative relationship with students (for example, by using "we" and "us") created more engaging and effective video lectures. They reported a significant reduction in video interactions (e.g., pause or seek events) when the instructor used a high frequency of first-person plural pronouns (i.e., "we" and "us"). The importance of pronoun choice in video lectures has been consistently supported as an effective means of communication in recent literature (Glover, 2018; Johnson & Picciuolo, 2022).

Additionally, the results showed Structural Function 2 (Topic Shifts) to be one of the most frequently appearing discourse functions in video lectures. Consistently, previous literature has emphasized the pivotal role of well-structured video lectures, utilizing pedagogical discourse that serves a structural function, in enhancing student engagement and learning outcomes. Lange and Costley (2020) underscored that clear organization of content, explicit statements of learning objectives, and concise summaries of key points significantly improve students' comprehension and engagement. This perspective aligns with previous literature that explored and emphasized the importance and effect of segmenting and providing pauses in instructional videos to improve student learning outcomes (Biard et al., 2018; Merkt et al., 2018; Bétrancourt & Benetos, 2018).

They highlighted the importance of clear conceptual explanations, the utilization of real-world examples, and the incorporation of formative assessments to measure student understanding. Finally, Fiorella (2022) offered practical strategies for structuring video lectures, including segmenting the lecture into manageable parts, signaling crucial points (e.g., Topic Shifts), and weaving new information into existing knowledge. This led to an enhancement in students' understanding and retention of information.

Interpersonal Function 3 (Showing Response) also frequently emerged in our video lecture as one of the important pedagogical discourse patterns. This function, which involves the instructor responding to potential queries or misconceptions, is crucial for creating a responsive, dynamic learning environment. Our study's findings were consistent with previous literature that used Fung and Carter's (2007) framework to evaluate teacher discourse in traditional classroom settings. For example, Vickov and Jakupcevic (2017) found that showing responses, such as "Ok," "All right," "Right," and "Yeah," were the most frequently used discourse markers in

classroom pedagogy. Our study also showed that showing responses accounted for a significant proportion of the total discourse marker use (35%–49%) in our dataset of five instructors (Table 7). Consistently, previous literature has highlighted the significance of instructor responsiveness, particularly Interpersonal Function 3 (Showing Response), as a factor that promotes engagement in video lectures and enhances the effectiveness of online learning (Choe et al., 2019; Johnson & Picciuolo, 2022). Vareberg and Westerman (2022, p. 81) assert that a dynamic learning experience can be achieved through responsive engagement from instructors. This perspective aligns with Lin (2022) and Bossman and Agyei (2022), who pinpointed instructors' timely feedback and interactive responses as key predictors of student satisfaction in online learning. Furthermore, studies have consistently emphasized the positive impact of immediate instructor responses and engagement on student satisfaction and learning outcomes in online classes (Anh, 2022; Mohammed et al., 2022). These findings are consistent with Bardach et al. (2022)'s stu dy, which concluded that responsive communication and feedback from instructors, reflecting on learning materials, significantly contribute to creating supportive and engaging learning environments, leading to better student self-efficacy and learning outcomes. Taken together, these studies underscore the critical role of Interpersonal Function 3 (Showing Response) in video lectures, demonstrating that active and responsive teaching substantially enhances student satisfaction, engagement, and learning outcomes in online contexts. Interestingly, a recent study by Ghafarpour (2022) demonstrated significant differences in discourse marker use patterns between traditional and online classroom settings. They reported that markers related to "marking shared knowledge" in the interpersonal function were used more frequently in traditional classroom settings, while our study found that such markers accounted for only 3%–8% of the total discourse marker incidences. Additionally, our analysis (Table 3) revealed that online instructors frequently used third-person pronouns like "We" and "Us" to build rapport with students, which was contrary to the findings of Yeo and Ting (2014) and O'Boyle (2014) who reported less frequent use of these pronouns in traditional classroom settings.

The findings of this study offer vital insights for advancing the future of video-based education, particularly in the context of Algebra classes. First, they provide educators with a more nuanced understanding of how to engage and connect with students in virtual math learning environments such as Math Nation. Through the meticulous analysis of 125 Algebra 1 video lectures—using computational linguistic tools and natural language processing (NLP)—this study uncovers the significance of conscious and deliberate choices of words, phrases, and discourse markers. These elements have been consistently identified in previous literature as crucial in enhancing student engagement and motivation. Second, our findings underscore the necessity for specialized training for instructors delivering video lectures in subjects such as Algebra. With the unique complexities of online teaching, as exemplified by variations in language choices and the threefold grouping into interpersonal, structural, and cognitive pedagogic functions, educators must be well-equipped with the right skills and knowledge to effectively communicate their instructional messages. Third, this study highlights the potential of machine learning and natural language processing as powerful tools for understanding and enhancing video-based education. By achieving an accuracy of 86.7% in classifying educators' language choices and utilizing a K-means clustering approach to examine variations, our research illuminates a promising avenue for using these technologies. As the field of online education progresses, these tools are expected to play an increasingly pivotal role in optimizing learning experiences and tailoring instruction to learners' needs.

In conclusion, this study represents a significant contribution to the growing body of literature on video-based education, teacher communication, and teacher discourse in the realm of digital Algebra instruction. Through the use of 26 Coh-Metrix and SÉANCE features, we have elucidated the intricate patterns of pedagogical communication, setting the stage for further research aimed at assessing the efficacy of digital instruction in computer-based learning environments.

6. Conclusion

To achieve our objectives, we combined the theoretical guidelines of Fung and Carter's (2007) pedagogical discourse marker framework with a data-driven approach using natural language processing to explore evidence of effective communication in video-based instruction.

Using a domain-independent text categorization approach combined with a variety of NLP techniques, our approach can be readily applied in lecture videos regardless of domain. Hence, the current results and three-stage analytic approach help establish a robust domain general evaluation framework for online lecture videos. In turn, this framework and analytic process provide an important piece of the puzzle for future investigations of virtual instruction. Our approach was novel because we bridged the gap of empirical findings using psycholinguistic and n-gram features supported by a theoretical pedagogical discourse marker framework to provide meaningful interpretations regarding instructor communication choices. Most importantly, our approach did not require any human labeling but still provided empirical evidence that is consistent and reinforces the findings from the previous literature that emphasize the need for better measures for teacher communication.

Our results reinforced the findings from the previous studies in automated pedagogical discourse categorization (Suresh et al., 2019, 2021). Consistent with prior literature, our findings identified interpersonal functions as one of the most frequently appearing pedagogical discourse functions in lecture videos (Suresh et al., 2019, 2021). In addition, our study provides further empirical evidence on the importance of understanding the communication component in the pedagogical discourse. Consistent with the findings of Stone et al. (2019), we identified that the content that is not specific to pedagogical communication component accounts for a significant amount (21%–24%) of proportions in the total pedagogical discourse. This highlights the importance of future research to focus on both the elements of content-specific and non-content-specific (i.e., communication) discourse to automatically detect and analyze teacher discourse (Sert, 2019, pp. 216–238). In addition, the attitudinal and emotional differences that the instructor's communication patterns conveyed significantly differed among the instructors. Little to no previous literature evaluated the sentimental differences that the instructors conveyed in the video lectures. Hence, this finding highlights the need for future research to investigate the

communication pattern and the sentimental (e.g., attitudinal and emotional) differences that the instructors could provide in their communication strategies.

6.1. Educational implications

This study contributes to scholarship about pedagogical communication patterns and language use in video-based instructional environments. The results of our study point out the potentialities of better computer-based instructions, and professional development of online instructors. First, our work revealed that there are distinctive patterns (e.g., overall linguistic and utterance-level evidence) that the instructors intentionally use to communicate in video lectures. These findings identify a new direction for future research in computer-based instructions and video-based learning to examine the efficacy of video-based instruction through the lens of evaluating characteristics of pedagogical communication. The novelty of our findings and methods is in the use of automated methods using natural language processing to evaluate such patterns from the lecture videos with a large number of discourse incidences (n = 4962). For example, future research is needed to investigate the overall linguistic dimensions, sentiment, frequency of utterance-level pedagogical markers, and their associated functional categories and how these features impact students' learning outcomes from video-based instruction.

Second, our findings and the methodology demonstrated the importance and the potential of incorporating natural language processing approaches to trace and investigate instructors' communication automatically. Such extensions would help instructors review and monitor their intentional verbal cues in virtual lecture videos, evaluate their pattern of speech, and potentially understand the effect of their communication skills. More personalized and individualized online instructions with instructor's effective instructional efforts and changes could positively affect students' learning (Xu et al., 2019, Beege et al., 2020). What is often m issing is the consideration for communication patterns that may directly affect students' interactions with the learning system. What is often missing is the consideration for communication patterns that may directly affect students' interactions with the learning system. Our findings provide a novel methodological demonstration to extend such capacity for future e-learning and online video lecture creation.

Third, our findings have direct implications in the research and the practices of instructor training and strategies. In Particular, our findings contribute to identifying the domains that are imperative in defining the effective pedagogical communication strategies in video lectures for new instructors in online learning platforms. Instructional designers and online instructors should be aware of the potential effect of language use (e.g., Cha et al., 2004; Beege et al., 2022) and monito r whether their language use covers different aspects of core functional pedagogy. The instructional designers will be able to access the common patterns of language use patterns of experienced instructors and evaluate the differences and the discrepancy with the less experienced instructors as professional development training resources. Additionally, instructional designers will have access to monitor which pedagogical communication patterns and its function contribute to the most for student learning based on their content, student knowledge, and the difficulty of the subject, thus catering effective instructional materials based on the learner's needs.

6.2. Limitations and future research

While the study was carefully designed and constructed to reduce the bias and limitations, we acknowledge the future studies may help contextualize and enrich the current findings. Our study focused on evaluating and demonstrating the capacity to extract communication patterns from algebra lecture videos, and evaluated close to 5000 discourses to provide generalizable results. Our findings provided empirical evidence regarding the instructor's varying communication patterns in online learning environments; however, the findings need to be generalized by incorporating our methodological framework to video lectures from other disciplines. Hence, we encourage future studies to evaluate the generalizability of our findings and the methods using lecture videos from varying mathematics domains and disciplines.

Furthermore, while our findings successfully uncovered the pedagogical language patterns and their functions from the instructor's language use, a limited investigation was conducted to associate the findings with student outcomes. Further explorations on whether our current findings may result in more effective, or less effective student learning would help us define the effectiveness of pedagogical communications in video lectures. Hence, we highly encourage future research to connect our current findings with student performance results to evaluate the effectiveness of pedagogical communication strategies identified in our study.

In addition, the primary dataset explored in our study was manually transcribed by human experts in order to make sure that the instructor's pedagogical communications are well extracted for investigation in various software. This may pose challenges in expanding the results as transcribing video lectures is a laborious process. Recent studies suggested that automated transcription systems may be incorporated to understand video lectures in learning platforms (Hukker & Gouder, 2022). Leveraging appropriate technological innovations to help navigate the potential hurdles in video lecture transcripts will be recommended in future studies.

One potential limitation of this study is that we only analyzed 125 videos from three algebra units, which covered a total of 25 topics, out of a total of 465 videos generated from 93 topics. While we randomly selected these videos, it is possible that our sample size was not large enough to capture the full range of topics and content covered in the video lectures. To address this limitation, we plan to expand our findings in future research by including a larger sample size of video lectures. This will allow for a more comprehensive analysis of the topics covered in video lectures and may reveal additional insights into the effectiveness of these lectures.

Credit author statement

Jinnie Shin: Methodology, Software, Conceptualization, Validation, **Renu Balyan**: Methodology, Software, **Michelle Banawan**:

Data curation, Conceptualization, **Tracy Arner**: Writing, Reviewing, and Editing, **Walter L. Leite**: Writing, Reviewing, and Editing, Data curation, **Danielle S. McNamara**: Writing, Reviewing, and Editing, Data curation, Conceptualization, Supervision

Data availability

The authors do not have permission to share data.

Acknowledgement

The research reported here was supported by the Institute of Education Sciences (IES), U.S. Department of Education through Grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Appendix A

Descriptive Statistics of the Final 11 Coh-Metrix Features in the RF model

Coh-Metrix		Instructor				
		1	2	3	4	5
WRDPRP1p	First person plural pronoun incidence	19.74 (7.51)	42.88 (12.17)	29.14 (9.00)	53.07 (12.12)	14.18 (4.14)
DRNP	Noun phrase density	341.84 (18.62)	351.96 (25.95)	313.45 (19.50)	327.60 (21.09)	344.89 (21.38)
WRDVERB	Verb incidence	137.62 (15.25)	129.18 (18.35)	132.32 (13.90)	133.63 (14.67)	133.59 (21.85)
SMINTEp	Intentional content	35.85 (6.24)	39.73 (8.68)	29.86 (5.90)	33.87 (5.25)	32.42 (5.81)
PCNARz	Narrativity	1.09 (0.30)	1.03 (0.29)	1.22 (0.24)	1.19 (0.20)	0.64 (0.26)
CNCTempx	Expanded temporal connectives incidence	15.00 (4.75)	10.64 (5.81)	14.12 (5.20)	11.29 (5.29)	11.32 (6.25)
DRPP	Preposition phrase density	75.30 (14.46)	75.01 (21.46)	68.54 (13.46)	64.51 (13.27)	78.50 (16.32)
PCVERBz	Verb cohesion z score	0.27 (0.44)	−0.03 (0.58)	0.56 (0.37)	0.02 (0.40)	0.03 (0.45)
PCSYNz	Syntactic simplicity	1.67 (0.32)	1.69 (0.36)	1.69 (0.30)	1.80 (0.31)	1.81 (0.39)
WRDPOLc	Polysemy for content words	4.33 (0.27)	4.33 (0.30)	4.57 (0.23)	4.41 (0.25)	4.13 (0.29)
PCCONNp	Connectivity	24.64 (13.39)	42.83 (22.79)	25.78 (19.07)	15.19 (14.49)	12.48 (11.53)

Note. Mean (Standard deviation).

Appendix B

Descriptive Statistics of the Final 14 SÉANCE Component Scores in the RF model

Instructor	Negative adjectives	Social order	Positive adjectives	Joy	Affect friends and family	Polarity nouns	Polarity verbs
1	−0.13	0.36	0.14	0.20	0.12	0.24	0.24
2	−0.29	0.44	0.29	0.30	0.19	0.32	0.19
3	−0.30	0.29	0.47	0.60	0.19	0.32	0.12
4	−0.13	0.39	0.35	0.31	0.13	0.30	0.15
5	−0.12	0.27	0.02	0.25	0.10	0.18	0.17
Instructor	Virtue adverbs	Positive nouns	Trust verbs	Failure	Economy	Certainty	Politeness
1	0.07	0.00	0.07	0.15	0.06	0.28	0.04
2	0.18	0.03	0.05	0.11	0.11	0.46	0.07
3	0.10	0.02	0.09	0.06	0.08	0.34	0.07
4	0.09	−0.01	0.12	0.08	0.07	0.31	0.06
5	0.01	0.00	0.07	0.11	0.03	0.17	0.05

Note. Mean (Standard deviation).

References

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*(1), 45–65.
Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics, 26*(10), 1340–1347.
Alves, F., Bettini, A., Ferreira, P. M., & Bessani, A. (2021). Processing tweets for cybersecurity threat awareness. *Information Systems, 95*, Article 101586.
Amadi, G., & Paul, A. K. (2017). Influence of student-teacher communication on students' academic achievement for effective teaching and learning. *American Journal of Educational Research, 5*(10), 1102–1107.

Anggoro, D., & Ningsih, T. W. R. (2023). *The effect of teacher interpersonal communication and communication climate in the class on students' learning interest in economic subjects*.
Anh, D. H. M. (2022). Factors affecting satisfaction on online education on students digital teaching page in Ho Chi Minh City, Vietnam. *Indonesian Journal of Multidiciplinary Research, 2*(1), 179–186.
Atapattu, T., & Falkner, K. (2017). March). Discourse analysis to improve the effective engagement of MOOC videos. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 580–581).
Balyan, R., Arner, T., Taylor, K., Shin, J., Banawan, M., Leite, W. L., & McNamara, D. S. (2022). Modeling One-on-One Online Tutoring Discourse Using an Accountable Talk Framework. *International Educational Data Mining Society*.
Beege, M., Krieglstein, F., & Arnold, C. (2022). How instructors influence learning with instructional videos-The importance of professional appearance and communication. *Computers & Education, 185*, Article 104531.
Beege, M., Schneider, S., Nebel, S., & Rey, G. D. (2017). Look into my eyes! Exploring the effect of addressing in educational videos. *Learning and Instruction, 49*, 113–120.
Belt, E. S., & Lowenthal, P. R. (2023). Synchronous video-based communication and online learning: An exploration of instructors' perceptions and experiences. *Education and Information Technologies, 28*(5), 4941–4964.
Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in Neural Information Processing Systems, 13*.
Bétrancourt, M., & Benetos, K. (2018). Why and when does instructional video facilitate learning? A commentary to the special issue "developments and trends in learning with instructional video". *Computers in Human Behavior, 89*, 471–475.
Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Application, 105*(9).
Biard, N., Cojean, S., & Jamet, E. (2018). Effects of segmentation and pacing on procedural learning by video. *Computers in Human Behavior, 89*, 411–417.
Bossman, A., & Agyei, S. K. (2022). Technology and instructor dimensions, e-learning satisfaction, and academic performance of distance students in Ghana. *Heliyon, 8*(4).
Brame, C. J. (2016). Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE-Life Sciences Education, 15*(4), es6.
Buckley, W., & Smith, A. (2007). Application of multimedia technologies to enhance distance learning. *RE: View, 39*(2), 57.
Cha, A., Hecht, B. R., Nelson, K., & Hopkins, M. P. (2004). Resident physician attire: Does it make a difference to our patients? *American Journal of Obstetrics and Gynecology, 190*(5), 1484–1488.
Chen, C. M., & Wu, C. H. (2015). Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Computers & Education, 80*, 108–121.
Chesebro, J. L., & McCroskey, J. C. (1998). The relationship of teacher clarity and teacher immediacy with students' experiences of state receiver apprehension. *Communication Quarterly, 46*(4), 446–456.
Choe, R. C., Scuric, Z., Eshkol, E., Cruser, S., Arndt, A., Cox, R., … Crosbie, R. H. (2019). Student satisfaction and learning outcomes in asynchronous online lecture videos. *CBE-Life Sciences Education, 18*(4), ar55.
Choi, H. J., & Johnson, S. D. (2005). The effect of context-based video instruction on learning and motivation in online courses. *American Journal of Distance Education, 19*(4), 215–227.
Civikly, J. M. (1992). Clarity: Teachers and students making sense of instruction. *Communication Education, 41*(2), 138–152.
Clark, H. H. (2002). Speaking in time. *Speech Communication, 36*(1–2), 5–13.
Clark, R. C., & Mayer, R. E. (2011). *E-learning and the science of Instruction Proven Guidelines for consumers and designers of multimedia learning* (3rd ed.) (Pfeiffer).
Conklin, S., & Dikkers, A. G. (2021). Instructor social presence and connectedness in a quick shift from face-to-face to online instruction. *Online Learning, 25*(1), 135–150.
Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods, 49*, 803–821.
Dinmore, S. (2019). Beyond lecture capture: Creating digital video content for online learning–a case study. *Journal of University Teaching and Learning Practice, 16*(1), 7.
Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., … D'Mello, S. K. (2016). Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 conference on user modeling adaptation and personalization* (pp. 45–53).
Fiorella, L., & Mayer. R.E. (2022). The Generative Activity Principle in Multimedia Learning. R.E. Mayer, L. Fiorella (Eds.), The cambridge handbook of multimedia learning (3rd ed.), Cambridge University Press (2021), pp. 339-350.
Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics, 28*(3), 410–439.
Ghafarpour, H. (2022). Interpersonal discourse markers in online vs. face-to-face EFL classes. *Teaching English Language, 16*(2), 63–84. https://doi.org/10.22132/TEL.2022.155086
Glover, P. (2018). How to evaluate the effectiveness of teacher talk. *International Online Journal of Education and Teaching (IOJET), 5*(3), 497–512. http://iojet.org/index.php/IOJET/article/view/473/249.
Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments. & computers, 36*(2), 193–202.
Guo, P. J., Kim, J., & Rubin, R. (2014). March). How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@Scale conference* (pp. 41–50).
Henderson, M. L., & Schroeder, N. L. (2021). A systematic review of instructor presence in instructional videos: Effects on learning and affect. *Computers and Education Open, 2*(October), Article 100059. https://doi.org/10.1016/j.caeo.2021.100059
Horn, D. C. (2022). *Instructional clarity examined as a transactional communication process*. Doctoral dissertation, Illinois State University.
Hrastinski, S., Cleveland-Innes, M., & Stenbom, S. (2018). Tutoring online tutors: Using digital badges to encourage the development of online tutoring skills. *British Journal of Educational Technology, 49*(1), 127–136.
Jensen, E., Dale, M., Donnelly, P. J., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020). Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–13).
Jeon, M. (2014). Analyzing the cohesion of English text and discourse with automated computer tools. *Journal of Pan-Pacific Association of Applied Linguistics, 18*(2), 123–133.
Johnson, J. H., & Picciuolo, M. (2022). The EmiBO corpus. EMI lecturer discourse across disciplines and lecture modes. *Lingue e Linguaggio, 53*(2022), 253–272.
Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher, 47*(7), 451–464.
Kim, Y., & Thayne, J. (2015). Effects of learner–instructor relationship-building strategies in online video instruction. *Distance Education, 36*(1), 100–114.
Lampos, V., Mintz, J., & Qu, X. (2021). An artificial intelligence approach for selecting effective teacher communication strategies in autism education. *Npj Science of Learning, 6*(1), 25.
Lange, C., & Costley, J. (2020). Improving online video lectures: Learning challenges created by the media. *International Journal of Educational Technology in Higher Education, 17*(1), 1–18.
Lee, M. C. (2010). Explaining and predicting users' continuance intention toward e-learning: An extension of the expectation–confirmation model. *Computers & Education, 54*(2), 506–516.
Lee, H. J., & Rha, I. (2009). Influence of structure and interaction on student achievement and satisfaction in web-based distance learning. *Journal of Educational Technology & Society, 12*(4), 372–382. https://www.learntechlib.org/p/74974/.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22. https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression- by-RandomForest.pdf.
Liew, T. W., Tan, S. M., Tee, J., & Goh, G. G. G. (2021). July). The effects of designing conversational commerce chatbots with expertise cues. In *2021 14th international conference on human system interaction (HSI)* (pp. 1–6). IEEE.
Lin, T. C. (2022). Student learning performance and satisfaction with traditional face-to-face classroom versus online learning: Evidence from teaching Statistics for Business. *E-learning and Digital Media, 19*(3), 340–360.
MacKay, J. R. (2019). Show and 'tool': How lecture recording transforms staff and student perspectives on lectures in higher education. *Computers & Education, 140*, Article 103593.
Madariaga, L., Nussbaum, M., Gutiérrez, I., Barahona, C., & Meneses, A. (2021). Assessment of user experience in video-based learning environments: From design guidelines to final product. *Computers & Education, 167*, Article 104176.
Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018). Mining theory-based patterns from big data: Identifying self-regulated learning strategies in massive open online courses. *Computers in Human Behavior, 80*, 179–196.
Martin, F., Xie, K., & Bolliger, D. U. (2022). Engaging learners in the emergency transition to online learning during the COVID-19 pandemic. *Journal of Research on Technology in Education, 54*(sup1), S1–S13.
Mayer, R. (Ed.). (2005). *The cambridge handbook of multimedia learning*. Cambridge University Press.
Mayer, R. E. (2020). The multimedia principle. In R. E. Mayer, & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed.). Cambridge University Press.
Mazer, J. P., & Hunt, S. K. (2008). "Cool" communication in the classroom: A preliminary examination of student perceptions of instructor use of positive slang. *Qualitative Research Reports in Communication, 9*(1), 20–28.
McFarlane, K. J. (2016). Tutoring the tutors: Supporting effective personal tutoring. *Active Learning in Higher Education, 17*(1), 77–88.
Merkt, M., Ballmann, A., Felfeli, J., & Schwan, S. (2018). Pauses in educational videos: Testing the transience explanation against the structuring explanation. *Computers in Human Behavior, 89*, 399–410.
Michaels, S., O'Connor, M. C., Hall, M. W., & Resnick, L. B. (2010). *Accountable talk sourcebook: For classroom conversation that works*. Pittsburgh, PA: University of Pittsburgh Institute for Learning.
Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*(4), 283–297.
Mirivel, J. C., Fuller, R., Young, A., & Christman, K. (2022). *Integrating positive communication principles and practices in business communication courses*. The Western Association for Business Communication Bulletin.
Mohammed, L. A., Aljaberi, M. A., Amidi, A., Abdulsalam, R., Lin, C. Y., Hamat, R. A., & Abdallah, A. M. (2022). Exploring factors affecting graduate Students' satisfaction toward E-learning in the era of the COVID-19 crisis. *European Journal of Investigation in Health, Psychology and Education, 12*(8), 1121–1142.
Nussbaum, J. F. (1992). Effective teacher behaviors. *Communication Education, 41*(2), 167–180.
Othman, Z. (2010). The use of okay, right and yeah in academic lectures by native speaker lecturers: Their 'anticipated' and 'real' meanings. *Discourse Studies, 12*(5), 665–681.
Perry, N. E. (2023). Introduction: Using qualitative methods to enrich understandings of self-regulated learning. In *Using qualitative methods to enrich understandings of self-regulated learning* (pp. 1–3). Routledge.
Pi, Z., Zhang, Y., Zhu, F., Xu, K., Yang, J., & Hu, W. (2019). Instructors' pointing gestures improve learning regardless of their use of directed gaze in video lectures. *Computers & Education, 128*, 345–352.
Polat, H. (2022). Instructors' presence in instructional videos: A systematic review. *Education and Information Technologies*, 1–33.
Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering, 34*(3), 1427–1445.
Richardson, J. C. (2001). *Examining social presence in online courses in relation to students' perceived learning and satisfaction*. State University of New York at Albany.
Ristad, E. S., & Yianilos, P. N. (1998). Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(5), 522–532.
Rockey, A., Gonzalez, L., Eberhardt-Alstot, M., & Merrill, M. (2020). Socially present? The perception of humanness online through video. In *International perspectives on the role of technology in humanizing higher education* (Vol. 33, pp. 13–32). Emerald Publishing Limited.
Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences, 3*(2), 1–12.
Scagnoli, N. I., Choo, J., & Tian, J. (2019). Students' insights on the use of video lectures in online classes. *British Journal of Educational Technology, 50*(1), 399–414.
Schneider, S., Krieglstein, F., Beege, M., & Rey, G. D. (2022). The impact of video lecturers' nonverbal communication on learning–An experiment on gestures and facial expressions of pedagogical agents. *Computers & Education, 176*, Article 104350.
Sert, O. (2019). *Classroom interaction and language teacher education*. The Routledge handbook of English language teacher education.
Shanshan, S., Chenhui, D., & Lijuan, L. (2022). Metaphor and board writing matter: The mediating roles of psychological distance and immersion in video lectures. *Computers & Education, 191*, Article 104630.
She, H. C., & Fisher, D. (2002). Teacher communication behavior and its association with students' cognitive and attitudinal outcomes in science in Taiwan. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 39*(1), 63–78.
Soriano, A. L. T., & Soriano, A. C. (2022). Power of positivity: Exploring the experiences of senior high school teachers and its impact to learners. *ASEAN Multidisciplinary Research Journal, 10*, 47–62.
Spooren, P., & Mortelmans, D. (2006). Teacher professionalism and student evaluation of teaching: Will better teachers receive higher ratings and will better students give higher ratings? *Educational Studies, 32*(2), 201–214.
Stone, C., Donnelly, P. J., Dale, M., Capello, S., Kelly, S., Godley, A., & D'Mello, S. K. (2019). *Utterance-level modeling of indicators of engaging classroom discourse*. International Educational Data Mining Society.
Suresh, A., Jacobs, J., Harty, C., Perkoff, M., Martin, J. H., & Sumner, T. (2022). *The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. arXiv preprint arXiv:2204.09652*.
Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J. H., & Sumner, T. (2021). *Using transformers to provide teachers with personalized feedback on their classroom discourse: The TalkMoves application. arXiv preprint arXiv:2105.07949*.
Suresh, A., Sumner, T., Jacobs, J., Foland, B., & Ward, W. (2019). In *Automating analysis and feedback to improve mathematics teachers' classroom discourse* (Vol. 33, pp. 9721–9728). Proceedings of the AAAI conference on artificial intelligence. No. 01.
Vareberg, K. R., & Westerman, D. (2022). *Seeking connection while connecting. The COVID-19 impact on*. Higher Education Stakeholders and Institutional Services.
Vickov, G., & Jakupcevic, E. (2017). Discourse markers in non-native EFL teacher talk. *Studies in Second Language Learning and Teaching, 7*(4), 649–671.
Wiebe, G., & Kabata, K. (2010). Students' and instructors' attitudes toward the use of CALL in foreign language teaching and learning. *Computer Assisted Language Learning, 23*(3), 221–234.
Wilson, D., Wright, J., & Summers, L. (2021). Mapping patterns of student engagement using cluster analysis. *Journal for STEM Education Research, 4*, 217–239.
Wong, M., Marshall, L. M., Blank, H. C., & Hard, B. M. (2022). Up close and personal: Examining effects of instructor video presence on student's sense of connection. *Scholarship of Teaching and Learning in Psychology, 8*(4), 374.
Xu, Z., Wijekumar, K., Ramirez, G., Hu, X., & Irey, R. (2019). The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology, 50*(6), 3119–3137.
Yeo, J.-Y., & Ting, S.-H. (2014). Personal pronouns for student engagement in arts and science lecture introductions. *English for Specific Purposes, 34*, 26–37.
Zhai, C., & Lafferty, J. (2017). A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR forum* (Vol. 51, pp. 268–276). New York, NY, USA: ACM. No. 2.
Zhu, L. (2023). Attitudes in teacher talk in EFL classroom from the perspective of appraisal theory. *Creative Education, 14*(5), 1111–1134.