Research Handbook on Classroom Observation

Edited by

Sean Kelly

Professor, Department of Educational Foundations, Organizations, and Policy, University of Pittsburgh, USA

ELGAR HANDBOOKS IN EDUCATION

Edward Elgar PUBLISHING

Cheltenham, UK · Northampton, MA, USA

Published by
Edward Elgar Publishing Limited
The Lypiatts
15 Lansdown Road
Cheltenham
Glos GL50 2JA
UK

Edward Elgar Publishing, Inc.
William Pratt House
9 Dewey Court
Northampton
Massachusetts 01060
USA

Authorised representative in the EU for GPSR queries only: Easy Access System Europe – Mustamäe tee 50, 10621 Tallinn, Estonia, gpsr.requests@easproject.com

A catalogue record for this book
is available from the British Library

Library of Congress Control Number: 2024952811

This book is available electronically in the Elgaronline
Sociology, Social Policy and Education subject collection
https://doi.org/10.4337/9781035321544

ISBN 978 1 0353 2153 7 (cased)
ISBN 978 1 0353 2154 4 (eBook)

11. Instructional improvement: leveraging computer-assisted textual analysis to generate insights from educational artifacts

Zewei (Victor) Tian, Min Sun, Alex Liu, Shawon Sarkar and Jing Liu

INTRODUCTION

Instructional improvement is an iterative process where educators harness data-driven insights to refine teaching practices and improve student learning. In recent years, richer and more complex educational data have been generated as researchers conduct systematic classroom observations for instructional improvement. These data offer vast potential for analyses but at the same time pose challenges in processing and analyzing this complex data. Conventional methods (such as observation rubrics and human-coding processes) are limited by various constraints, which prevent educators from utilizing data to improve teaching and learning outcomes in a timely fashion. These constraints include often needing a long duration and trained researchers to analyze such complex data. The lack of timeliness limits the utility of results generated from conventional methods to inform the refinement of educators' practices. Hence, artificial intelligence and machine learning (AI/ML) approaches are able to effectively process such complex data with scalability and precision (Berland et al., 2014), presenting an unprecedented opportunity to advance research and instructional improvement in education. Recent forms of generative AI can also generate insights and interpretations from the data and provide actionable insights. A new field of research has emerged, in which researchers integrate cutting-edge AI/ML techniques with educational domain knowledge of curriculum, teaching, and learning to explore crucial questions for instructional improvement. This new field of study allows researchers in education to broaden their discourse with peers in other disciplines who use similar methods to mutually learn from one another, enabling the next generation of methodology and theory development.

In this chapter, we will review and discuss how AI/ML methods provide us with innovative solutions to analyze textual data in education (e.g., transcripts from classroom observations, student assignments, lesson plans, etc.), as well as summarize the promises and pitfalls of these new methodological advancements. To guide our review of this emerging new field, we use the Instructional Core Framework that depicts the dynamic interactions among students, educators, and educational content to create learning opportunities and classroom environments (City et al., 2009; Gillies, 2015; Hennessy et al., 2023). Textual artifacts pertaining to each component and their interactions offer valuable insights into instructional practices.

Specifically, we identify three main areas where AI/ML analysis of textual data can be adapted: teacher contributions, student contributions, and curricular content. Each of these three components both utilize and generate textual data. Teachers design curricula and lesson plans as textual data. At the same time, the instruction inside classrooms can also be

transcribed to serve as data for analysis of instruction quality. On the other hand, teachers rely on the feedback from student assignments to provide tailored course content. Students submit their assignments as potential textual data for analysis while taking teachers' instruction as inputs. Content itself serves as a major textual data input, but also incorporates new data like feedback from both the teachers and the students to adapt and better serve the educational purposes. For each component, we discuss example studies relevant to understanding data-driven research and improvement targeting that component, and conduct a nuanced examination of the complexities, potentials, and limitations inherent in AI/ML-powered textual analysis of instructional artifacts related to that component. Moreover, we will explore how AI/ML can be utilized as an observational tool to enhance classroom observations, providing more precise and scalable insights into instructional practices to gain a deeper understanding of the dynamics within the classroom. In the end, we remain cautiously optimistic about the value of this line of research in advancing education and identify several major directions for future research.

THE INSTRUCTIONAL CORE FRAMEWORK

The Instructional Core Framework is a conceptual tool used to understand the dynamic interactions among teachers, students, and content within the classroom. It serves to facilitate the categorization of textual analysis in educational research, analyze use cases, and anticipate future directions. We use this framework to contextualize the findings and convey the broader implications of the results from the existing body of knowledge. The Instructional Core Framework challenges traditionally prevailing paradigms, advocating for a departure from teacher-centric instructional approaches toward a more dynamic interplay between teachers and students (City et al., 2009). The Instructional Core Framework functions as a guide through the multifaceted terrain of instruction, offering an analytic lens into the classroom dynamics that are conducive to the enhancement of both teaching and learning.

The instructional core includes three interdependent components, with dynamic interactions between one another. These components include *teachers'* knowledge and skill, *students'* engagement in learning, and academically challenging *content*. The components in the instructional core collectively influence the quality of teaching and learning in a classroom. To improve classroom instruction, either the components themselves or the relationships among the teacher, student, and content need to be changed. For example, improving students' engagement in learning may require changes in the knowledge and skill of teachers, in turn affecting content, and altering the ways in which teachers engage students with content (Elmore, 2008).

Instructional improvement can focus on any of the three components in the instructional core, including the quality and dynamics of the interactions between them. For example, efforts for improvement in the three components of the instructional core might involve providing feedback on teachers' knowledge and practices (Bain & Swan, 2011), academic assessment based on group and individual student practice opportunities (Doabler et al., 2019), and generating high-quality and coherent curriculum and materials as instructional content (Harris et al., 2015). These efforts aim to enhance the effectiveness of teaching and learning processes, contributing to instructional improvement in educational settings.

Besides the individual components, the connections between instructional core components are also key characteristics integrated within the framework, linking teachers, students, and content in a holistic manner. These interdependent relationships provide additional ways to improve educational practices, leading to another important aspect of the Instructional Core Framework; improvements in one component may depend on other components (City et al., 2009). As indicated in Figure 11.1, all three components are correlated, meaning that disentangling one from the other two may cause imbalance and consequently fail to create improvements in instructional practices. If we aim to enhance the content, we should assess whether teachers are equipped to handle the more advanced curriculum and also consider whether students can effectively engage with it (Elmore, 2010). This demonstrates the interplay between content and teachers, as well as between content and students. Merely incorporating high-level content into the curriculum, without taking into account the readiness and acceptance of both students and teachers, would not be prudent.

To achieve effective instructional improvement, it is crucial to thoroughly gather, organize, and analyze relevant data pertaining to the instructional core. Today, vast amounts of observational data on instruction are being generated each year, but the volume of this educational data poses new challenges and opportunities. With the help of AI/ML, educators and administrators might be better able to utilize this vast data, gaining valuable insights and making informed decisions based on evidence. Textual artifacts of classroom instruction take a variety of forms, including: transcripts of instructional discourse, students' written homework, chat history between tutors and students, textbooks, and different types of Open Educational Resources. These first-hand educational materials and data capture nuanced details and patterns of classroom dynamics, which are often absent from traditional, structured, administrative data produced by conventional methods (Abrahamson & Sánchez-García, 2016; Scribner & Donaldson, 2001). Natural Language Processing (NLP; a method concerned with giving computers the ability to support and manipulate human language) and other computer-assisted manipulation of linguistic data can be used to discover patterns in either structured or unstructured data. To make these methods useful to educators, researchers must integrate

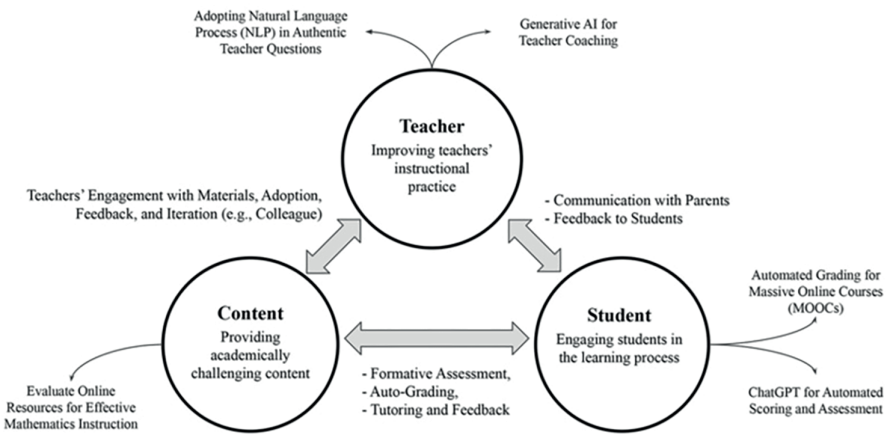

Figure 11.1 Opportunities for AI/ML to influence the instructional core

domain- and subject-specific knowledge into the inferential process that guides these methods' application.

To show the potential of AI/ML methods, when well informed by the Instructional Core Framework, to generate novel insights on teaching practices and enhance student learning experiences (Cardona et al., 2023), in this chapter we consider a set of use cases, along with implications for future research. For example, by providing automated feedback on teacher discourse, AI/ML becomes a powerful tool for professional development and instructional change. For students, it facilitates auto-grading and coursework evaluation, offering valuable feedback and access to intelligent tutoring systems. Additionally, AI/ML enables large-scale data analysis, providing unique insights. AI/ML methods can even assist in tasks of generating content and insights, such as lesson planning and individual student feedback. This integration not only streamlines administrative tasks but also reshapes the dynamics within the educational landscape. In terms of *teacher-content* dynamics, AI/ML offers tailored tasks suited to varying cognitive demands while providing teachers with manageable plans. It could even be used to bridge the gap between teachers and students by diminishing pre-existing perceptions of student achievement levels (Seo et al., 2021). The goal of such transformations is to eliminate the divide between teachers' assumptions, research-based expectations, and students' actual capabilities, fostering effective conversations and facilitating a more accurate understanding of students' potential. Ultimately, the integration of AI/ML into systems that provide information on various components and interactions, as understood through the Instructional Core Framework, has the power to propel education toward unprecedented possibilities and foster deeper connections and personalized learning experiences.

TEXTUAL ANALYSIS TO SUPPORT TEACHERS' LEARNING

In recent years, the landscape of professional development for educators has undergone a transformative shift by leveraging AI/ML capabilities to enhance instructional practices. Traditional feedback mechanisms, often sporadic and evaluative in nature, are evolving as technology becomes an integral part of this process (Demszky et al., 2023; Jensen et al., 2020). As part of practical demand, research has just started to demonstrate the efficacy of AI-driven tools in refining teacher discourse and fostering a more inclusive classroom environment. Within this context, NLP offers a powerful approach to measure teaching practices through the analysis of textual data like classroom transcripts by identifying linguistic nuances, communication styles, and the frequency and quality of teacher–student engagement (Tosey & Mathison, 2010). This provides educators with valuable insights into their strengths and areas for improvement. In this section, we summarize a few exemplary studies.

Advancing Classroom Discourse: Applying Natural Language Processing (NLP) to Identify Authentic Teacher Questions

Classroom discourse in educational settings manifests through diverse forms, ranging from teacher-directed lectures and procedural communication to open-ended discussions fostering collaborative exchanges between students and teachers (Alexander, 2008; Juzwik et al., 2015). By creating dialogic space and opening up classroom discourse, teachers are able to enhance students' active learning, particularly in language classrooms, as high-quality classroom

discourse is characterized by open-ended or authentic questions along with formative feedback, where student contributions are probed and elaborated on (Hardman, 2016; Soter et al., 2008).

An abundance of authentic questions can promote substantive conversation and is related to overall student engagement and achievement growth. Kelly et al. (2018) incorporated AI into a classroom observation system to address difficulties in measuring question authenticity at scale, an endeavor which previously relied on human observations and human coding. The study adopted automatic speech recognition, NLP, and ML to train models to detect authentic questions. The methodology for this study underwent iterative refinement, drawing insights from two primary data sources. First, a comprehensive archival database with text transcripts from 451 observations across 112 classrooms was utilized (Partnership data). Additionally, a new set of 132 high-quality audio recordings from 27 classrooms was collected (Class 5.0 data), adhering to technical constraints aimed at anticipating large-scale automated data collection and analysis. These diverse data sources laid the foundation for the investigation conducted by the research team.

Kelly et al. (2018) adopted different methods for archival and newly collected data. First, they applied a semi-automated measurement approach to the archival Partnership data, where human coders had previously carried out a wide range of different tasks, including segmenting the raw classroom audio into teacher utterances, identifying which utterances are questions, distinguishing instructional questions from non-instructional questions, and providing approximate transcriptions of the instructional questions. The next step included syntactic and discourse parsing, which was used to compute sentence (a teacher utterance) and multi-sentence discourse features (Manning et al., 2014; Surdeanu et al., 2015). Then, the study team identified various properties of language to arrive at a set of 244 features at each level of structure (word, sentence, discourse), including question stems (e.g., "what," "how," "why"), word order, and referential chains, and so forth. Using the newly collected Class 5.0 data, the team applied a fully automated approach with the ultimate goal of providing a measure of authenticity from a recording of classroom audio without any human involvement.

With different approaches and abundant datasets collected, the authors report a sufficiently high correlation between the computer- and human-coded authenticity, which suggests a promising avenue toward further automated measurement of classroom discourse. Kelly and colleagues' work highlights how NLP techniques can be used for such analysis purposes, not only offering a more efficient and scalable approach to measuring question authenticity but also holding the potential for improving teacher professional development and classroom instruction (2018). However, the authors also recognized the challenges that may occur when implementing in real-world cases, which include issues like noise, dialect diversity, data imbalance, and many more. It is undeniable that authenticity in itself is a complex construct, yet still, the inclusion of automated approaches empowered by NLP can offer some promising solutions to identify authentic questions and other classroom discourse features. By leveraging NLP, innovative approaches can provide high-quality feedback to teachers derived from transcripts of classroom interactions and detect instructional factors which are well aligned with commonly used observation protocols (Danielson, 2013). This unlocks the potential to enhance existing classroom observation systems through collecting far more data on teaching at a lower cost, higher speed, and with the detection of multifaceted classroom practices (Liu & Cohen, 2021).

Using NLP to Identify Accountable Talk and Address the Diversity of Lesson Study

Research teams have also applied automated methods to the study of "Accountable Talk," a framework for discourse in mathematics classes that facilitates student learning (Michaels et al., 2008). By utilizing advances in automatic speech recognition and NLP, automated approaches can help identify features of accountable talk to direct classroom discourse to be more engaging and efficient. A pilot study of feedback on Accountable Talk with the TalkMoves application by Jacobs et al. (2022) showcases a promising trend of increased use of accountable talk features, emphasizing the potential of platforms like the TalkMoves application to democratize access to high-quality feedback for educators. These advancements demonstrate the prowess of ML in swiftly and accurately classifying classroom discourse, offering automated insights that align remarkably well with human-coded evaluations (Wang et al., 2014). This convergence underscores the capacity of NLP to streamline feedback processes and provide invaluable support for instructional improvement by utilizing textual data generated through instructional practices.

Instructional processes in classrooms consist of a diverse set of activities. One approach to professional development that embraces this diversity is "lesson study," where teachers systematically examine their practice and improve their instructional practices (Makinae, 2019; Saito, 2012). "Lesson study" recognizes the complexity of cognitive demands inside classrooms and provides ways for teachers to understand such complexity. Appropriately trained and fine-tuned NLP models could, in theory, efficiently analyze huge amounts of data generated from inside the classroom and provide accurate feedback for teachers. With guidance from these analyses in the form of instructional feedback, teachers could then develop their own learning and understanding. This process can ultimately help both teachers and students achieve high-level tasks in classrooms. High-level tasks are critical because they require higher-order thinking skills, which are linked to higher performance and cognitive skills. By integrating high-level tasks with the right instructional practices, students can achieve very competitive outcomes (Newmann et al., 2001).

Generative AI for Teacher Coaching

Teachers' actions and inactions in the classroom are critical in determining the quality of the learning experience for students. Professional development serves as a dynamic and evolving mechanism through which educators engage in continuous learning and refinement of their skills. Internationally, classroom observation is widely recognized as a key tool for teachers' professional development and evaluation (Martinez et al., 2016). As part of formal, administrative evaluation, the feedback is conventionally provided by school administrators, peer teachers, and instructional coaches. Administrators conducting evaluations use pre-determined protocols and assessments that include various rubrics, facilitating and structuring the evaluation process. However, due to the variation in expertise, resources, and human subjectivity under different contexts, assessments of such kinds inherently have limitations in their validity and applicability (Kelly et al., 2020). The heterogeneity of evaluation processes and feedback induces variation (or disparities) in teacher development opportunities and practices, which in turn significantly impact student educational outcomes.

Prior efforts have shown the promise of adopting machine learning and natural language processing in assessing teachers' instructional practices, such as detecting authentic questions

(Kelly et al., 2018) and accountable talk moves (Jacobs et al., 2022; Suresh et al., 2022). It is important to note that the work by Kelly et al. (2018) and Jacobs et al. (2022) does not address teacher evaluation directly. Instead, it focuses on enhancing instructional practices through detailed feedback. Kelly (2023) introduces the concept of "agnostic" instructional observation systems that avoid making evaluative judgments at the point of coding. These systems offer detailed, fine-grained data that teachers can use to reflect on and improve their practices without the pressure of evaluative judgment. To provide actionable feedback similar to that provided by human evaluators and coaches during professional development sessions, generative AI, like GPT-3.5, has huge potential in providing relevant feedback that aligns with conventional coaching. These tools have already demonstrated capabilities in generating texts in education settings and are widely adopted by teachers and school professionals, playing multiple roles in pedagogical activities (Jeon & Lee, 2023). The use of agnostic systems, as advocated by Kelly (2023), can enhance these AI tools' effectiveness by ensuring that feedback remains descriptive and non-judgmental, thereby supporting professional growth rather than evaluation.

Leveraging Generative AI to Facilitate Teachers' Day-to-day Activities

In the selected case study for this section, Wang and Demszky (2023) investigated whether ChatGPT can help teachers and education professionals by applying observational rubrics to classroom observation data, and then providing effective feedback and generating helpful pedagogical suggestions. The authors assigned three separate tasks for ChatGPT to perform with observational data. The first task was to score a transcript segment for items derived from classroom observation instruments, according to criteria of the Classroom Assessment Scoring System (CLASS) and Mathematical Quality Instruction (MQI) rubrics. The second task was to identify highlights and missed opportunities for CLASS and MQI items, elaborating on these elements that occurred or did not occur during instruction. The third task was to provide open-ended, actionable suggestions to the teacher for eliciting more student mathematical reasoning in the classroom. These three stages of analysis are similar to what human evaluation and coaching would look like, but now applied more efficiently in an automated process.

The data used in this study was extracted from the National Center for Teacher Effectiveness (NCTE) Transcript dataset (Demszky & Hill, 2023) in this work, the largest publicly available dataset of U.S. classroom transcripts linked with classroom observation scores. The dataset consists of four years (2010–2013) of classroom transcripts from 4th and 5th grade elementary mathematics observations. At the time of data collection, experts assessed and rated the transcripts based on two instruments, both CLASS and MQI. CLASS instrument scoring segmented instruction into 15-minute sections, whereas MQI corresponds with 7.5-minute sections. All randomly selected segments were provided to ChatGPT in combination with zero-shot prompting (using a pre-trained language model to generate text in response to a task or query that it hasn't been specifically trained on). The authors measured zero-shot performance first because segments of the NCTE transcripts are long, causing annotated sections to exceed the input limit. Additionally, zero-shot prompting closely aligns with how an average teacher would interact with ChatGPT day-to-day, utilizing the generative AI as it is given without specific fine-tuning. Demszky and colleagues recruited human participants to examine the validity of ChatGPT's performance on all three tasks, evaluating the outputs

alongside human ratings. The results indicated that ChatGPT was able to provide relevant responses, yet failed to give novel and insightful feedback. The low quality of strategic feedback is likely attributable to the scarcity of instructional information at ChatGPT's disposal, leading to insufficient information around examples of teacher coaching. The incorporation of generative AI presents promising prospects for improving certain aspects of teacher coaching, such as automating the measurement of classroom interactions. However, whether entirely automated teacher coaching is possible or advisable remains uncertain. Adequate training data and illustrative examples might enhance the capabilities of AI, but we do not yet know if these will be sufficient. Moreover, while automating the measurement of classroom activities can provide valuable insights, automating the advice and feedback for teachers may not be as beneficial. It is essential to consider that teachers, who are professionals in their fields, should be empowered to interpret the data and reach their own conclusions about how to improve their instructional practices, based on the information provided by automated systems.

In sum, this study employed a zero-shot performance evaluation approach, mimicking how teachers might interact with generative AI in their day-to-day activities without specific fine-tuning. This approach aligns with practical implementations and offers insights into the model's adaptability. However, the effectiveness of generative AI, as demonstrated in the study, is contingent upon the availability of ample training data and illustrative examples, the lack of which points to challenges in replicating the novelty and insightfulness of human-generated feedback. Despite illustrating ChatGPT's performance with zero-shot prompting, the study still calls for development and fine-tuning required for AI models in the context of teacher coaching and professional development.

Textual Analysis for Student Support and Assessment

AI/ML technologies, including NLP, present diverse avenues for enhancing students' coursework. This transformative potential is evident in three major applications: (a) generating personalized and adaptive items through fine-tuned models tailored to specific subject areas; (b) analyzing and auto-grading coursework; and (c) providing students with feedback, including intelligent tutoring. Just as AI/ML contributes to the democratization of high-quality feedback for educators in professional development, its applications in supporting student learning hold the promise of more personalized, adaptive, and efficient learning experiences for students.

Integrating AI/ML in adaptive learning is an approach that can be used to customize instruction to learners' backgrounds, experiences, and prior knowledge (Alam, 2023; Peng et al., 2019; Walkington, 2013). AI/ML systems can select or recommend optimal content from a set of available resources. They can provide guidance on designing a well-structured long-term curriculum that sequences learning activities over a unit of instruction, and connect learners to appropriate content that matches their strengths and bypasses their weaknesses, ultimately aiding in precise performance evaluations and personalized learning (Maghsudi et al., 2021). AI/ML tools offer a range of customization spanning different subject areas, as well as various pedagogical approaches (Ruiz-Rojas et al., 2023).

Furthermore, teachers have used AI/ML-powered tools to grade assignments, provide efficient and timely feedback to students, and predict student learning trajectories and outcomes (Chen, 2018; Wilson et al., 2022). The automation of assessment processes offers benefits such as rapid evaluation, consistency in grading, and the ability to handle large volumes of student work. This type of technology takes as input student submissions through online platforms or

through transcribed responses, and then relies on computer-based algorithms like text mining in order to find the similarities between student responses and the assessment requirements (grading criteria) determined by teachers (Kakkonen & Sutinen, 2004). Traditionally, free-response questions on tests and assignments (e.g., where students compose paragraphs or essays) require a particularly heavy effort by teachers to grade. Here, AI/ML methods offer an efficient evaluation method by focusing on critical words and sentences present in the student composition, analyzing the logical semantic relationship of the content, and generating an overall grade (Wang et al., 2018). Even when the AI/ML system is not provided with a grading rubric, but only a training dataset of human-scored examples, currently available auto-grading technology can still handle the assessment and achieve good inter-rater agreement with expert grading (Yang et al., 2017).

The third application, intelligent tutoring, is a dynamic combination of the previous two. By leveraging student log data, which contains activities and previous performance, and utilizing sentence-level semantic representations of student responses to open-ended questions, AI/ML can provide a collaborative filtering-based approach to both predict student scores and recommend appropriate feedback messages for teachers to send to their students (Botelho et al., 2023). Intelligent Tutoring Systems (ITSs) have been designed to deliver adaptive guidance and instruction, evaluate learners, define and update the learner's model, and classify or cluster learners (Corbett et al., 1997; Mousavinasab et al., 2021; Nwana, 1990).

Automated Grading for Massive Open Online Courses (MOOCs)

From K-12 to higher education, a significant migration from physical materials to digital resources has occurred in recent years, especially following the COVID-19 pandemic. The fact that more and more students are relying on computers to complete and submit their schoolwork requires teachers to adapt, learning how to grade assignments and assess student performances in this new context. Zhao et al. (2017) provide a comprehensive exploration into the domain of automated grading tools, particularly focused on essay writing and open-ended assignments within the context of Massive Open Online Courses (MOOCs). The research addresses the burgeoning need for scalable grading solutions as more students rely on digital platforms for academic submissions. The proposed model, centered around memory networks, is particularly effective. This is because memory networks allow the model to utilize a large memory component to store and retrieve rich representations of previously graded responses. This capability enables the model to handle complex reasoning tasks and make more accurate predictions. The evaluation of this model was conducted using the Kaggle Automated Student Assessment Prize (ASAP) dataset, which is a publicly available dataset of student essays graded by human raters. The dataset consists of eight sets of essays, each corresponding to a different prompt and grading rubric. The model was able to achieve state-of-the-art performance, excelling in seven out of the eight essay sets. This was measured using the Quadratic Weighted Kappa (QWK) metric, which assesses the agreement between the predicted scores and the human-assigned scores. Achieving high QWK scores indicates that the model's predictions are closely aligned with human judgment, demonstrating its effectiveness in automated essay grading.

The authors relied exclusively on the Kaggle ASAP dataset, leveraging it for both training and evaluating the memory networks-powered automated grading model. The model of the study comprises four key layers. The first layer is input representation, responsible for

generating vector representations of student responses. The second layer is memory addressing, loading selected responses into memory with weighted assignments. The third layer is memory reading, retrieving content based on weighted summation. The fourth and last layer is output, making predictions from the resulting state. Stacking Memory Addressing and Reading layers enhances the model's ability to learn abstract representations through successive computational stages. The primary focus lies in establishing a robust and reliable representation of assignments through vectorization, coupled with the storage of pertinent samples in the memory component. However, the study recognizes the need for future endeavors to expand the model's applicability, urging the exploration of diverse datasets featuring varied assignment formats to enhance generalizability. Attention is also directed toward refining assignment representation and mechanisms for measuring relevance among assignments in forthcoming research.

The authors indicated that there are two key factors to the performance: reliable representation and memory component. By recognizing and accounting for these two factors, the model's potential generalizability to assignments from diverse subjects underscores its significance in the broader educational landscape. Acknowledging its merits, the study concurrently underscores its limitations, notably being tested solely on the Kaggle ASAP dataset. This calls for further exploration with diverse datasets containing various assignment formats. The other area marked for improvement is the representation of the assignment and the mechanism for measuring relevance among assignments, requiring enhancement in these aspects to improve the overall robustness and applicability of the model.

ChatGPT for Automated Scoring and Assessment in K-12 Settings

Generative AI like ChatGPT utilizes large language models (LLMs) to generate responses in accordance with prompts and inputs. LLMs, a type of language model that can both understand and generate texts, work better when they have been fine-tuned to the specific task at hand. For example, with sufficient training data, a 2-billion parameter model named MathGLM can provide multi-digit arithmetic operations with almost 100% accuracy, significantly surpassing GPT-4, the model on which ChatGPT Plus is based (Yang et al., 2023). Similarly, generative AI has been fine-tuned for performance of educational prompts, including automatically scoring student-written constructed responses using example assessment tasks in science education. The fundamental difference between this generative AI approach and the AI/ML approaches discussed in the previous MOOC section lies in their operational methodologies and applications. AI/ML approaches for assignment scoring, such as those utilizing memory networks, rely on comparing new responses with pre-stored graded samples to predict scores. These models are trained to recognize patterns and similarities based on specific features and historical data. In contrast, generative AI models like ChatGPT generate responses by predicting the next word or sequence of words based on the context provided by the input prompt. They are not just recognizing patterns from past data but are generating new content that aligns with the input requirements. When fine-tuned for specific tasks, such as rubrics-based scoring responses, they learn to generate evaluative feedback and scores in a manner that mimics human judgment. This allows them to provide more nuanced and contextually appropriate feedback to both teachers and students, although the models' effectiveness depends heavily on the quality and extent of the fine-tuning data.

Latif and Zhai (2024) have fine-tuned ChatGPT (GPT-3.5) for scoring assessments in science classrooms. While GPT-3.5 has demonstrated proficiency in natural language processing, its direct use for scoring student responses is limited by the contextual variation in language because it does not specifically focus on science instruction without fine-tuning. The study trained GPT-3.5 on specific assessment tasks using a dataset that included student responses and expert scores. A comparative analysis was conducted, comparing GPT-3.5 with BERT—which stands for Bidirectional Encoder Representations from Transformers, another pre-trained natural language processing model that has significantly advanced language understanding tasks by capturing bidirectional context representations (Devlin et al., 2019). The research aimed to gauge the effectiveness of GPT-3.5 in enhancing automatic scoring accuracy. Additionally, the study explored text data augmentation as a creative strategy to leverage GPT-3.5-turbo's capabilities for enhancing automated evaluation precision. It demonstrated the effectiveness of this approach in producing a more extensive and diverse training set for scoring mechanisms, leading to improved performance in evaluating student text responses (Cochran et al., 2023).

These findings highlight the potential of domain-specific fine-tuning in improving the performance of large language models for educational assessment tasks. This approach provides a valuable tool for educators and researchers by delivering more accurate and reliable scoring of student responses. The fine-tuned models have been made available for public use and community engagement, supporting their practical implications. However, potential limitations include the need for additional extensive fine-tuning and further studies on the generalizability of these models across different educational contexts, such as online versus in-person instruction, varying class sizes, and diverse subjects. This underscores the importance of continued research and development to maximize the utility and effectiveness of AI in educational settings.

Textual Analysis for Content Analysis and Development

High-quality instructional materials can have significantly positive effects on student learning and outcomes (Read, 2015). Large language models (LLMs) offer a way to both analyze existing curricular resources and provide new materials based on insights gained from the analyses. As the online landscape of educational resources grows, schools and educators are increasingly turning to digital materials alongside traditional textbooks. Open educational resources (OERs), with their undeniable promise of cost-effectiveness and accessibility, present a prospect for overcoming cost barriers by democratizing access to high-quality materials (Richter & McPherson, 2012). However, a crucial question looms: how can we vet the quality of these OER materials to ensure content rigor, engagingness of activities, and inclusivity to diverse students' backgrounds and needs? After the computer understands the quality of instructional materials, the generative AI models will. An interdisciplinary group of researchers, funded by the National Science Foundation, proposes a novel and rigorous approach to addressing these key questions by integrating domain knowledge in mathematics instruction with the latest technologies of AI/ML analysis of text data (National Science Foundation, 2023).

Open Education Resources as Data

Sun et al. began with an extensive data collection phase, amassing over 40,000 mathematics lesson plans from a variety of OER platforms, including widely known sites like BetterLesson, Illustrative Mathematics, and Achieve the Core. The research team paid special attention to the representativeness of the sampled lesson plans to guarantee that the collection exemplified the breadth of resources available for educators. The gathered lesson materials were then organized in data tables and stored using advanced data management solutions, enhancing efficiency in data handling and allowing for the integration of various types of data, such as information on how widely used the materials were (download numbers) and comments from the OER websites. The team also conducted regular exploratory data analyses to examine the distribution of key measures within the lesson materials. This process was crucial for making sense of the data and identifying any potential errors. This robust data collection and analysis formed a solid foundation for the subsequent phases of the study, where the quality of these lesson materials was scrutinized and enhanced through human-centered evaluations and AL/ML algorithms.

Classification Methods of High-Quality Lesson Plans: Human-Centered Data Science Approach

Central to the study is a human-centered method for assessing the quality of lesson plans, an approach instrumental in ensuring the reliability and validity of both the measures and the algorithms employed. This approach initiates with the development of a theory/conceptual framework for high-quality instruction through a comprehensive literature review and the application of the Delphi method, a mixed-method technique combining surveys and an expert panel to achieve consensus on key aspects of instructional quality and their measurement. Utilizing domain expertise from educators and researchers, the study formulates rigorous quality measures and coding rubrics, laying the groundwork for human coding. A team of experienced teachers and instructional coaches then manually codes 1,000 selected lesson plans, evaluating them against the established criteria for curriculum alignment and pedagogical efficacy. This step provides a nuanced understanding of what constitutes quality in K-12 mathematical instruction, ensuring the machine learning (ML) algorithms, trained on these evaluations, are based on a well-defined framework and an expert-endorsed training dataset.

The research team combines machine learning algorithms with human expertise to assess lesson plans in a comprehensive study to enhance the quality of educational resources. These algorithms are specifically developed to identify essential quality features in lesson plans, such as highly cognitively demanding tasks and meaningful student discourse. The multi-task classification models are based on two powerful pre-trained language models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and are trained on human-coded data. These ML algorithms identify patterns in lesson plans and offer a scalable solution for assessing the broad array of available lesson materials.

Researchers typically use several evaluation methods to validate the quality ratings generated by the classification. These include cross-validation, where a subset of the data is used to test the algorithm. Researchers can also compare algorithm ratings with human experts' ratings. For example, Sun and her team trained a new group of 20 math master teachers to rate 2,000 machine-coded lesson plans using the same rubrics developed for this study, to both

evaluate the algorithm's performance and offer human feedback to refine machine learning algorithms. Through this blend of AI and human judgment, the study aims to offer a robust tool for educators, guiding them toward high-quality, pedagogically sound resources in the evolving digital education landscape. Therefore, by fine-tuning AI's assessment capabilities through teachers' practical understanding of quality measures drawn from classroom experience, the study ensures that algorithms are not only theoretically sound but also practically relevant and effective. Integrating human expertise and automation presents a powerful tool for educators, enabling them to navigate and select high-quality OERs, thereby addressing the critical need for pedagogical efficacy and curricular alignment in the digital age of education.

Fine-Tune LLM to Generate Lesson Materials that are Consistent with these Quality Measures

The future direction of the research project involves fine-tuning LLMs to generate lesson materials that adhere to established quality measures. This initiative marks a pivotal shift from evaluating existing educational resources to actively creating new ones using cutting-edge AI technology. The primary goal of fine-tuning LLMs is to align the generated lesson materials with the high-quality benchmarks set in the study's earlier phases. Furthermore, insights from teacher evaluations of existing lesson plans will inform the iterative fine-tuning process. The continuous refinement will also involve pilot testing of the AI-generated materials in real classroom settings to gather empirical data on their effectiveness and to make context-specific adjustments. A key challenge will be bridging the gap between AI capabilities and educational expertise. Addressing this, the research emphasizes interdisciplinary collaboration, uniting AI researchers, data scientists, and educational experts to ensure that the LLMs are trained and fine-tuned in a manner that accurately reflects the intended pedagogical principles and practices. The team will also continuously explore customization and adaptability of LLMs, aiming to create tailored lesson materials that cater to diverse learning styles, student needs, and curriculum requirements. Ethical considerations, particularly regarding the use of data and the potential impacts on educational equity, will remain a priority.

FUTURE DIRECTIONS

This chapter summarizes the emerging field of adopting AI/ML-powered textual analysis for instructional improvement. We utilize the Instructional Core Framework to guide the review and discussion of existing studies pertaining to the three components of teacher, student, and content, providing one to two case studies to illustrate the scenarios under which AI/ML can be implemented. AI, particularly large language models (LLM) and generative AI, holds the potential to expedite the translation of research into EdTech products and redefine research methodologies. We highlight the transformative capacity of these technologies in enhancing instructional efficiency and effectiveness. By weaving together learning and instructional theories with the latest technologies, we envision that AI-powered textual analysis and generation will make a significant contribution to K-12 research and practices. This chapter is written to inspire dialogue, foster innovation, and chart a path toward an era of unparalleled educational excellence. Specifically, we envision the following trends will be extended.

Trend 1: The human-centered AI partnership model involves the mutual reinforcement and integration of AI/ML methodologies, educational domain knowledge, and human expertise. The advancements in AI/ML bring about revolutionary changes in various aspects of education, such as reshaping teaching and learning experiences and enhancing educational tools. On the other hand, it is crucial to incorporate educational domain knowledge to ensure that AI/ML models are relevant, useful, and effective in educational settings. Additionally, as the integration of AI in education progresses, addressing ethical considerations becomes increasingly important. By incorporating human elements and domain knowledge into AI/ML algorithms, fairness, accountability, transparency, and explainability can be ensured. Conversely, effective analysis of instructional artifacts by AI/ML may uncover new foundational knowledge about instruction and human learning or create more personalized instructional materials and pedagogical practices to support student learning. The synergy between AI's ability to recognize patterns and automate tasks with human insight into context-specific understanding has the potential to foster an adaptive, inclusive, and equitable educational ecosystem that empowers all learners and educators.

Trend 2: AI technology and research aim to amplify human creativity. This perspective acknowledges the essential value of human experience in both educational processes and AI tool development, positioning AI as a catalyst for fostering creativity and developing skills. This trend emphasizes technology's role in enabling human potential by encouraging critical thinking, nuanced problem-solving, and collaboration. It involves examining AI's role in facilitating human cognition and behavioral development through adaptive algorithms and flexible accessibility. Drawing on nudging and other socio-technical learning theories, where humans and technology iterate and bounce off ideas, will be critical to inform technology design features and become an important perspective for researchers studying instructional improvement.

Trend 3: Developing high-quality and inclusive annotated data will be key to supporting AI-powered analysis for further improvement. As exemplified in several summarized studies in this chapter, researchers have invested significant resources in annotating important features that align with instructional domain knowledge. These annotated data help train or fine-tune AI algorithms to understand the nuances and intricacies of educational content, enabling them to provide more accurate analyses and insights. Furthermore, it is crucial to ensure that the annotated data used to train algorithms represent diverse student and educator backgrounds and preferences. This helps combat algorithmic bias by ensuring that the AI models are trained on a wide range of perspectives, experiences, and cultural contexts. Sharing these high-quality annotated datasets is also essential for supporting the broader research community. By making these datasets available to other researchers, we can foster collaboration, encourage innovation, and accelerate progress in developing effective AI-powered analysis tools for instructional improvement.

In conclusion, the human-centered AI partnership model in education involves integrating AI/ML methodologies, educational domain knowledge, and human expertise. Future research should continue to explore how AI/ML can enhance classroom observations, providing more detailed and scalable insights into instructional practices. This will help in developing more personalized and effective educational interventions, bringing about revolutionary changes in teaching and learning experiences by enhancing educational tools and reshaping instructional practices. To fully leverage AI in the continuous improvement of instruction and learning, the development of common, shared data infrastructure is key. This infrastructure should support

the seamless integration of analytics and generation, enhancing AI-powered tools' effectiveness, combating algorithm bias, and amplifying human learning. Future research should focus on the integration of these analytics tools to provide real-time, actionable insights for educators.

REFERENCES

Abrahamson, D., & Sánchez-García, R. (2016). Learning is Moving in New Ways: The Ecological Dynamics of Mathematics Education. *Journal of the Learning Sciences*, *25*(2), 203–239. https://doi.org/10.1080/10508406.2016.1143370

Alam, A. (2023). Harnessing the Power of AI to Create Intelligent Tutoring Systems for Enhanced Classroom Experience and Improved Learning Outcomes. In G. Rajakumar, K.-L. Du, & Á. Rocha (Eds.), *Intelligent Communication Technologies and Virtual Mobile Networks* (pp. 571–591). Springer Nature. https://doi.org/10.1007/978-981-99-1767-9_42

Alexander, R. (2008). Culture, Dialogue and Learning: Notes on an Emerging Pedagogy. *Exploring Talk in School*, *2008*, 91–114.

Bain, A., & Swan, G. (2011). Technology Enhanced Feedback Tools as a Knowledge Management Mechanism for Supporting Professional Growth and School Reform. *Educational Technology Research and Development*, *59*, 673–685. https://doi.org/10.1007/S11423-011-9201-X

Berland, M., Baker, R., & Blikstein, P. (2014). Educational Data Mining and Learning Analytics: Applications to Constructionist Research. *Technology, Knowledge and Learning*, *19*. https://doi.org/10.1007/s10758-014-9223-7

Botelho, F., Tshimula, J. M., & Poenaru, D. (2023). Leveraging ChatGPT to Democratize and Decolonize Global Surgery: Large Language Models for Small Healthcare Budgets. *World Journal of Surgery*, *47*(11), 2626–2627. https://doi.org/10.1007/s00268-023-07167-2

Cardona, M. A., Rodríguez, R. J., & Ishmael, K. (2023). *Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations*. https://policycommons.net/artifacts/3854312/ai-report/4660267/

Chen, H. (2018). *Predicting Student Performance Using Data from an Auto-Grading System* [Master Thesis, University of Waterloo]. https://uwspace.uwaterloo.ca/handle/10012/13435

City, E. A., Elmore, R. F., Fiarman, S. E., & Teitel, L. (2009). *Instructional Rounds in Education* (Vol. 30). Harvard Education Press. https://www.education.ne.gov/wp-content/uploads/2021/11/Instructional-Rounds-in-Education-Elmores-Instructional-Core.pdf

Cochran, K., Cohn, C., Rouet, J. F., & Hastings, P. (2023). Improving Automated Evaluation of Student Text Responses Using GPT-3.5 for Text Data Augmentation. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 217–228). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_18

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent Tutoring Systems. In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of Human–Computer Interaction* (2nd Ed.) (pp. 849–874). North-Holland. https://doi.org/10.1016/B978-044481862-1.50103-5

Danielson, C. (2013). *The Framework for Teaching Evaluation Instrument*. https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/17302/2013_FfTEvalInstrument_Web_v1.2_20140825_.pdf

Demszky, D., & Hill, H. (2023). *The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts* (arXiv:2211.11772). arXiv. https://doi.org/10.48550/arXiv.2211.11772

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023). Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence from a Randomized Controlled Trial in a Large-Scale Online Course. *Educational Evaluation and Policy Analysis*, 01623737231169270. https://doi.org/10.3102/01623737231169270

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Doabler, C. T., Stoolmiller, M., Kennedy, P. C., Nelson, N. J., Clarke, B., Gearin, B., Fien, H., Smolkowski, K., & Baker, S. K. (2019). Do Components of Explicit Instruction Explain the Differential Effectiveness of a Core Mathematics Program for Kindergarten Students With Mathematics Difficulties? A Mediated Moderation Analysis. *Assessment for Effective Intervention*, *44*(3), 197–211. https://doi.org/10.1177/1534508418758364

Elmore, R. (2008). Improving the Instructional Core. *Draft Manuscript*. https://achievethecore.org/content/upload/Improving%20The%20Instructional%20Core_Elmore%20Article.pdf

Elmore, R. (2010). Leading the Instructional Core. *Conversation*, *11*(3), 1–12.

Gillies, R. M. (2015). *Enhancing Classroom-based Talk: Blending Practice, Research and Theory.* Routledge.

Hardman, J. (2016). *Opening-up Classroom Discourse to Promote and Enhance Active, Collaborative and Cognitively-Engaging Student Learning Experiences.* https://doi.org/10.14705/rpnet.2016.000400

Harris, C. J., Penuel, W. R., D'Angelo, C. M., DeBarger, A. H., Gallagher, L. P., Kennedy, C. A., Cheng, B. H., & Krajcik, J. S. (2015). Impact of Project-based Curriculum Materials on Student Learning in Science: Results of a Randomized Controlled Trial. *Journal of Research in Science Teaching*, *52*(10), 1362–1385. https://doi.org/10.1002/tea.21263

Hennessy, S., Calcagni, E., Leung, A., & Mercer, N. (2023). An Analysis of the Forms of Teacher–Student Dialogue that are Most Productive for Learning. *Language and Education*, *37*(2), 186–211. https://doi.org/10.1080/09500782.2021.1956943

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting Rich Discussions in Mathematics Classrooms: Using Personalized, Automated Feedback to Support Reflection and Instructional Change. *Teaching and Teacher Education*, *112*, 103631.

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. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376418

Jeon, J., & Lee, S. (2023). Large Language Models in Education: A Focus on the Complementary Relationship between Human Teachers and ChatGPT. *Education and Information Technologies*, *28*(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Juzwik, M. M., Borsheim-Black, C., Caughlan, S., & Heintz, A. (2015). *Inspiring Dialogue: Talking to Learn in the English Classroom*. Teachers College Press. https://books.google.com/books?hl=en&lr=&id=yqdDAwAAQBAJ&oi=fnd&pg=PR7&dq=Juzwik+et+al.,+2013+classroom&ots=NBZk7y27MS&sig=0FvRibIh0Sf2oeeOEywS879rWb8

Kakkonen, T., & Sutinen, E. (2004). Automatic Assessment of the Content of Essays based on Course Materials. *ITRE 2004. 2nd International Conference Information Technology: Research and Education*, 126–130. https://doi.org/10.1109/ITRE.2004.1393660

Kelly, S. (2023). Agnosticism in Instructional Observation Systems. *Education Policy Analysis Archives*, *31*. https://doi.org/10.14507/epaa.31.7493

Kelly, S., Bringe, R., Aucejo, E., & Fruehwirth, J. C. (2020). Using Global Observation Protocols to Inform Research on Teaching Effectiveness and School Improvement: Strengths and Emerging Limitations. *Education Policy Analysis Archives*, *28*, 62. https://doi.org/10.14507/epaa.28.5012

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. https://doi.org/10.3102/0013189X18785613

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for Automatic Scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210.

Liu, J., & Cohen, J. (2021). Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *Educational Evaluation and Policy Analysis*, *43*(4), 587–614. https://doi.org/10.3102/01623737211009267

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Maghsudi, S., Lan, A., Xu, J., & van der Schaar, M. (2021). Personalized Education in the Artificial Intelligence Era: What to Expect Next. *IEEE Signal Processing Magazine*, *38*(3), 37–50. https://doi.org/10.1109/MSP.2021.3055032

Makinae, N. (2019). The Origin and Development of Lesson Study in Japan. In R. Huang, A. Takahashi, & J. P. da Ponte (Eds.), *Theory and Practice of Lesson Study in Mathematics: An International Perspective* (pp. 169–181). Springer International Publishing. https://doi.org/10.1007/978-3-030-04031-4_9

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In K. Bontcheva & J. Zhu (Eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-5010

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom Observation for Evaluating and Improving Teaching: An International Perspective. *Studies in Educational Evaluation*, *49*, 15–29.

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. https://doi.org/10.1007/s11217-007-9071-1

Mousavinasab, E., Zarifsanaiey, N. R., Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent Tutoring Systems: A Systematic Review of Characteristics, Applications, and Evaluation Methods. *Interactive Learning Environments*, *29*(1), 142–163. https://doi.org/10.1080/10494820.2018.1558257

National Science Foundation. (2023). Integration of Computer-Assisted Methods and Human Interactions to Understand Lesson Plan Quality and Teaching to Advance Middle-Grade Mathematics Instruction. Award number: 2300291. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2300291

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?* Consortium on Chicago School Research. https://consortium.uchicago.edu/publications/authentic-intellectual-work-and-standardized-tests-conflict-or-coexistence

Nwana, H. S. (1990). Intelligent Tutoring Systems: An Overview. *Artificial Intelligence Review*, *4*(4), 251–277. https://doi.org/10.1007/BF00168958

Peng, H., Ma, S., & Spector, J. M. (2019). Personalized Adaptive Learning: An Emerging Pedagogical Approach Enabled by a Smart Learning Environment. *Smart Learning Environments*, *6*(1), 9. https://doi.org/10.1186/s40561-019-0089-y

Read, T. (2015). *Where Have All the Textbooks Gone? Toward Sustainable Provision of Teaching and Learning Materials in Sub-Saharan Africa*. World Bank Publications.

Richter, T., & McPherson, M. (2012). Open Educational Resources: Education for the World? *Distance Education*, *33*(2), 201–219. https://doi.org/10.1080/01587919.2012.692068

Ruiz-Rojas, L. I., Acosta-Vargas, P., De-Moreta-Llovet, J., & Gonzalez-Rodriguez, M. (2023). Empowering Education with Generative Artificial Intelligence Tools: Approach with an Instructional Design Matrix. *Sustainability*, *15*(15), Article 15. https://doi.org/10.3390/su151511524

Saito, E. (2012). Key Issues of Lesson Study in Japan and the United States: A Literature Review. *Professional Development in Education*, *38*(5), 777–789. https://doi.org/10.1080/19415257.2012.668857

Scribner, J. P., & Donaldson, J. F. (2001). The Dynamics of Group Learning in a Cohort: From Nonlearning to Transformative Learning. *Educational Administration Quarterly*, *37*(5), 605–636. https://doi.org/10.1177/00131610121969442

Seo, K., Dodson, S., Harandi, N. M., Roberson, N., Fels, S., & Roll, I. (2021). Active Learning with Online Video: The Impact of Learning Context on Engagement. *Computers and Education*, *165*, 104132. https://doi.org/10.1016/j.compedu.2021.104132

Soter, A. O., Wilkinson, I. A., Murphy, P. K., Rudge, L., Reninger, K., & Edwards, M. (2008). What the Discourse Tells Us: Talk and Indicators of High-level Comprehension. *International Journal of Educational Research*, *47*(6), 372–391. https://doi.org/10.1016/j.ijer.2009.01.001

Surdeanu, M., Hicks, T., & Valenzuela-Escárcega, M. A. (2015). Two Practical Rhetorical Structure Theory Parsers. In M. Gerber, C. Havasi, & F. Lacatusu (Eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 1–5). Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-3001

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:2204.09652). arXiv. https://doi.org/10.48550/arXiv.2204.09652

Tosey, P., & Mathison, J. (2010). Neuro-linguistic Programming as an Innovation in Education and Teaching. *Innovations in Education and Teaching International*, *47*(3), 317–326. https://doi.org/10.1080/14703297.2010.498183

Walkington, C. A. (2013). Using Adaptive Learning Technologies to Personalize Instruction to Student Interests: The Impact of Relevant Contexts on Performance and Learning Outcomes. *Journal of Educational Psychology*, *105*(4), 932–945. https://doi.org/10.1037/a0031882

Wang, R. E., & Demszky, D. (2023). *Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction* (arXiv:2306.03090). arXiv. https://doi.org/10.48550/arXiv.2306.03090

Wang, Z., Liu, J., & Dong, R. (2018). Intelligent Auto-grading System. *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 430–435. https://doi.org/10.1109/CCIS.2018.8691244

Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014). Automatic Classification of Activities in Classroom Discourse. *Computers & Education*, *78*, 115–123. https://doi.org/10.1016/j.compedu.2014.05.010

Wilson, J., Pollard, B., Aiken, J. M., Caballero, M. D., & Lewandowski, H. J. (2022). Classification of Open-ended Responses to a Research-based Assessment Using Natural Language Processing. *Physical Review Physics Education Research*, *18*(1), 010141. https://doi.org/10.1103/PhysRevPhysEducRes.18.010141

Yang, X., Zhang, L., & Yu, S. (2017). Can Short Answers to Open Response Questions be Auto-Graded Without a Grading Rubric? In E. André, R. Baker, X. Hu, M. Ma, T. Rodrigo, & B. Du Boulay (Eds.), *Artificial Intelligence in Education* (Vol. 10331, pp. 594–597). Springer International Publishing. https://doi.org/10.1007/978-3-319-61425-0_72

Yang, Z., Ding, M., Lv, Q., Jiang, Z., He, Z., Guo, Y., Bai, J., & Tang, J. (2023). *GPT Can Solve Mathematical Problems Without a Calculator* (arXiv:2309.03241). arXiv. https://doi.org/10.48550/arXiv.2309.03241

Zhao, S., Zhang, Y., Xiong, X., Botelho, A., & Heffernan, N. (2017). A memory-augmented neural model for automated grading. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale* (pp. 189–192).