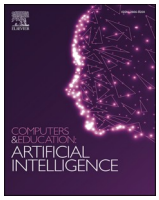
Computers and Education: Artificial Intelligence 6 (2024) 100207

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



Automatic classification of activities in classroom videos

Jonathan K. Foster [a,*], Matthew Korban [b], Peter Youngs [c], Ginger S. Watson [d], Scott T. Acton [b]

[a] Department of Educational Theory and Practice, University at Albany, United States
[b] C.L. Brown Department of Electrical and Computer Engineering, University of Virginia, United States
[c] Department of Curriculum, Instruction, and Special Education University of Virginia, United States
[d] Virginia Modeling, Analysis, and Simulation Center, Old Dominion University, United States

ARTICLE INFO

Keywords:
Elementary education
Classroom video
Classroom activity recognition
Neural networks
Computer vision

ABSTRACT

Classroom videos are a common source of data for educational researchers studying classroom interactions as well as a resource for teacher education and professional development. Over the last several decades emerging technologies have been applied to classroom videos to record, transcribe, and analyze classroom interactions. With the rise of machine learning, we report on the development and validation of neural networks to classify instructional activities using video signals, without analyzing speech or audio features, from a large corpus of nearly 250 h of classroom videos from elementary mathematics and English language arts instruction. Results indicated that the neural networks performed fairly-well in detecting instructional activities, at diverse levels of complexity, as compared to human raters. For instance, one neural network achieved over 80% accuracy in detecting four common activity types: whole class activity, small group activity, individual activity, and transition. An issue that was not addressed in this study was whether the fine-grained and agnostic instructional activities detected by the neural networks could scale up to supply information about features of instructional quality. Future applications of these neural networks may enable more efficient cataloguing and analysis of classroom videos at scale and the generation of fine-grained data about the classroom environment to inform potential implications for teaching and learning.

Many educational researchers rely on videos to study phenomena that occur in classrooms; videos provide several advantages to examine classroom interactions for research as well as teacher education and professional development (Gaudin & Chaliès, 2015; Janík & Seidel, 2009; Xu et al., 2018). One advantage is supporting multiple research purposes and opportunities for secondary analysis (Andersson & Sørvik, 2013; Derry et al., 2010; Jacobs et al., 1999; Klette, 2022). Because of the video record, one person (or group) can watch a classroom interaction multiple times or freeze the frame to attend to several features at once. Another added value is the ability to analyze at various timescales (Dalland et al., 2020; Derry et al., 2010).

With the use of video in education, there has been a concurrent development of technologies to assist with tasks such as the organization and storage of videos, creation of video transcripts, software for video annotations, and production of analytic schemes or reports (Derry et al., 2010; Goldman et al., 2014; Jacobs et al., 1999; Klette, 2022; Pea & Hoffert, 2007). Emerging technological advances for videos are a critical component of the classroom research agenda and researchers should carefully consider how these technologies contribute to the construction of data and analysis (Hall, 2000). In recent years, deep learning neural networks have emerged as one of the leading approaches for human activity recognition in videos due to their robustness for extracting video-based features and promising performance for highly complex and critical tasks (Beddiar et al., 2020; Gupta et al., 2022). In this paper, we contribute to the literature by considering the application of neural networks for constructing and analyzing classroom video data and consider their potential implications for teaching and learning.

Transforming video recordings into useful data is a time-consuming process (Derry et al., 2010). A growing number of researchers are investigating whether machine learning applications, such as neural networks, can be efficiently applied to video and audio data to study classrooms (e.g., Dale et al., 2022; Demszky & Hill, 2022; Jacobs et al., 2022; Pang et al., 2023; Sun et al., 2021; Wang et al., 2014). For instance, Kelly et al. (2018) developed automated methods to detect authentic teacher questions from audio recordings and transcripts in secondary English language arts (ELA) lessons. Other fields, such as medicine, have found that the application of neural networks to video can augment traditional approaches to make them more efficient and

* Corresponding author.
E-mail address: jkfoster@albany.edu (J.K. Foster).

https://doi.org/10.1016/j.caeai.2024.100207
Received 1 September 2023; Received in revised form 12 January 2024; Accepted 13 January 2024
Available online 17 January 2024

cost effective (e.g., Saba et al., 2019).

Even though there is potential for using machine learning to construct and generate data from videos, substantial amounts of annotated classroom video data are needed to develop these machine learning algorithms. Unfortunately, large and high-quality annotations of classroom video datasets specifically for machine learning development are not widely available, although a few attempts have been made (Sharma et al., 2021; Sun et al., 2021). In this paper, we describe our development of a large and high-quality annotated dataset of classroom videos from elementary classrooms in the United States and then we examine whether neural networks can detect instructional activities. Our goals are to determine whether we can (a) discriminate between different classroom instructional activities using neural networks of video data (b) rapidly and with accuracy levels comparable to that of humans.

1. Challenges scaling classroom videos in research and practice

Previous video technologies have aided researchers' and teachers' analysis of videos; the development of some these technologies arose from methodological and practical needs from large-scale classroom studies and professional development (Borko et al., 2008; Goldman et al., 2014; Jacobs et al., 1999; Klette, 2022; Pea & Hoffert, 2007). However, there remain challenges when using videos at scale, whether using them for descriptive research, evaluating an intervention, or providing feedback to teachers. Next, we highlight a few of these challenges and describe ways in which automated systems, embedded with neural networks, may be able to augment existing practices.

First, collecting a large dataset (i.e., hundreds of hours) of classroom videos presents the practical challenges of storing and cataloguing the data. Automated systems could efficiently summarize video content for cataloguing and searching large video collections with little to no human intervention (Pea & Hoffert, 2007). Second, due to financial and time constraints, researchers must strategically choose their unit of analysis and consider the feasibility of coordinating analysis at multiple timescales (Derry et al., 2010; Stigler et al., 2000). Researchers select timescales based on their research questions, theoretical perspective, and practical constraints but even minor adjustments in timescale analysis could lead to different interpretations (e.g., Dalland et al., 2020). To help researchers account for these differences, automated systems could assist in the extraction of certain timescale features (e.g., teacher versus student talk time) while freeing up time and resources for researchers to systematically extract features at more complex timescale (e.g., quality of student engagement during small group interactions).

Lastly, analyzing large-scale classroom video datasets presents the financial and time burdens of training humans to generate useful data (Hiebert et al., 2003; Stigler et al., 2000). For instance, human raters may need to complete frequent calibration sessions to ensure they are not diverging from accepted coding procedures (e.g., Walkowiak et al., 2014). Despite all these efforts to train human raters to create useful data, some studies suggest that humans may be the greatest source of error (Casabianca et al., 2015; Hill et al., 2012; Kelly et al., 2020; Klette, 2022). Automated systems could assist with some of these burdens by supplementing human efforts. For instance, some observation protocols capture the frequency of instructional activities (i.e., quantity) and describe their qualities. Automated systems may prove capable of summarizing the frequency of instructional activities efficiently and accurately; this offloading would enable more concentrated focus on instructional quality.

2. Conceptual framing: agnostic and fine-grained classroom observation measures aligned with ambitious instruction

Classroom videos offer rich records for studying human activities such as gestures, eye gaze, speech, tone of voice, and use of physical artifacts (Barron, 2003). These human activities occur along a spectrum; they can range from simple and short in duration to interactively complex and longer in duration. From least to most complex, these action classes are *simple* (e.g., raising a hand), *interaction* (e.g., reading a book), *group* (e.g., teacher supporting multiple students), and *event* (e.g., whole-class discussion).

To focus our selection of instructional activities from the classroom videos, we took a "fine-grained" and "agnostic" approach (Kelly, 2023) while also acknowledging the neural network models would not be able to use any audio or speech features from the classroom videos. Fine-grained analysis typically lends itself to binary labeling and can be applied exhaustively to the data; for instance, labeling whether students are raising their hands in videoframes. At the point of selecting instructional activities, we were agnostic regarding whether the instructional activities aligned with effective instruction. For example, a teacher transitioning students to a new instructional format may be considered effective or ineffective. We were inclusive of instructional activities that, at the point of labeling in the video, one could suspend evaluation of the activity. However, this approach does not mean we would be unable to evaluate instructional activities in the future. As in the case of the teacher guiding the transition, noting whether the transition was under 2 min lends itself to some evaluation of the transition's efficiency. In addition, any fine-grained and agnostic instructional activities had to be identifiable by manual annotation without the use of audio signals or speech data from the classroom video.

Two reliable and validated classroom observation instruments, the Mathematics-Scan (M-Scan; Berry et al., 2013; Walkowiak et al., 2014) and the Protocol for Language Arts Observations (PLATO; Corr, 2011; Grossman et al., 2013) guided our conceptualization of instructional activities. M-Scan and PLATO have been informed by many years of classroom-based research (see Section 2.1). These classroom observation instruments can be operationalized in fine-grained and agnostic ways. While Kelly (2023) only argued PLATO could be applied in fine-grained and agnostic ways, his argument could similarly be applied to M-Scan as well. M-Scan and PLATO measure features of ambitious instruction (Grossman et al., 2014; Walkowiak et al., 2018). Ambitious instruction seeks to foster conceptually rich understanding of disciplinary content (Franke et al., 2007; Newmann & Associates, 1996; Thompson et al., 2013).

2.1. Instructional activity labels

Our fine-grained and agnostic instructional activity labels are organized under 6 parent-level labels: activity type, teacher location, student location, teacher supporting, discourse, and representing content (see Table 1). Next, we provide an overview and rationale for our video-based instructional activities and summarize research related to these instructional activities.

2.1.1. Activity type

Activity type labels are the instructional formats the teacher engages in with students. These labels included whole class activity, individual activity, small group activity, and transition. We included these labels for three reasons. First, there is evidence to suggest a positive relationship between instructional time and student achievement (Baker et al., 2004; Bodovski & Farkas, 2007; Borg, 1980; Brophy & Good, 1984; Carroll, 1989; Gettinger, 1984; Stallings, 1980; Wiley & Harnischfeger, 1974) and some argue that certain activity types may (or may not) have implications for student engagement (see Kelly & Turner, 2009). Second, activity types have been manually labeled in many video studies (e.g., Hiebert et al., 2003). Previous research investigating the capabilities of neural networks to detect features in classroom videos has not focused on activity types (e.g., Ahuja et al., 2019; Sharma et al., 2021; Sun et al., 2021). Therefore, developing a neural network capable of detecting activity type may provide an efficient means for labeling this fine-grained measure. Third, we are interested in whether these fine-grained and agnostic measures may scale up for global observation

Table 1
List of instructional activity labels.

Instructional Activity	Definition	Action Level
Activity Type		
Whole class activity	All students are involved in one activity, with the teacher leading the activity (e.g., lecture, presentation, carpet time).	Event
Individual activity	All students privately work (e.g., independent practice, reading) at a separate desk or in small groups with no interaction between students.	Event
Small group activity	Students working together with peers (e.g., think-pair-share, book club); this is prioritized when there are students interacting or somewhat interacting near one another.	Event
Transition	The students and teacher transition from one instructional activity to another (e.g., whole class to small group). The teacher and students move from one spot in the room to another (e.g., from the carpet to desks). Other than specific behavioral directions, no instruction or meaningful instructional activity is occurring during the transition.	Event
Discourse		
On task student talking with student	Students conversing together without direct teacher support, which may overlap with small group activity. This is specific to mouth-movements within the parent code time interval.	Group
Student raising hand	A student's hand is up for more than 1 s; clearly and purposefully raising hand.	Simple
Teacher Location		
Teacher sitting	Teacher sitting (chair, stool, floor, crouching, on desk, kneeling).	Simple
Teacher standing	Teacher standing in generally the same spot to keep the same orientation to students.	Simple
Teacher walking	Teacher walking with purpose to change orientation to students.	Simple
Student Location		
Student(s) sitting on carpet or floor	Students sitting on floor or carpet.	Simple
Student(s) sitting at group tables	Students sitting at tables.	Simple
Student(s) sitting at desks	Students at individual desks.	Simple
Student(s) standing or walking	Students standing up or walking around the room.	Simple
Teacher Supporting		
Teacher supporting one student	Teacher uses proximity to offer assistance to one student; support can be verbal or non-verbal.	Group
Teacher supporting multiple students with student interaction	Teacher uses proximity to offer assistance to multiple students; support can be verbal or non-verbal. Individual students are also interacting with one another.	Group
Teacher supporting multiple students without student interaction	Teacher uses proximity to offer assistance to multiple students who are engaged in an activity; support can be verbal or non-verbal. Students are sitting close to one another or in a small group, but they are not interacting with one another.	Group
Representing Content		
Using or holding book	A book is used or held by a teacher or student.	Interaction
Using or holding worksheet	A worksheet is used or held by a teacher or student.	Interaction
Presentation with technology	A interative whiteboard, document camera, or projector is used to show content.	Interaction
Using or holding instructional tool	A tangible object (e.g., ruler, math manipulative; anything in someone's hand other than what is already listed, but does not include pencil/pen) is used or held by teacher or student for instructional purposes.	Interaction
Using or holding notebook	A notebook is used or held by a teacher or student.	Interaction
Individual technology	Student or teacher using a laptop, tablet, etc.	Interaction
Teacher writing	Teacher inscribing on paper, whiteboard, or document camera; includes erasing.	Interaction
Student writing	Student inscribing on paper, whiteboard, or document camera; includes erasing.	Interaction

protocols like M-Scan and PLATO. For example, some researchers have suggested that the primary activity type in a lesson was linked to ratings for certain dimensions of PLATO (Luoto et al., 2023).

Previous research provided some perspective on the frequency and duration of the activity types we would likely observe in our dataset. The Beginning Teacher Evaluation Study was a relevant example of this research (Rosenshine, 1981). It was an observational study on how time was spent in elementary classrooms in the United States. The study revealed that in grade 2 students spent 35 min and 90 min in mathematics and ELA lessons each day, respectively, and in grade 5 students spent 45 min and 110 min in mathematics and ELA, respectively. A recent survey of elementary teachers' self-reported time spent in mathematics and ELA instruction reported comparable results (Banilower et al., 2018).

Burns (1984) provided a comprehensive review of research on time allocation in elementary classrooms; he found notable differences in instructional time and engagement in activity types. For instance, in one study he reviewed, small group activities accounted for 73% of the time a second grader spent in ELA but, by the fifth grade, only 55% of ELA instruction was spent in small group activities; in comparison, small group instruction in mathematics was 40% and 34%, respectively (Lambert & Hartsough, 1976). Rosenshine (1981) reported that elementary students spent most of their time working on independent seatwork, 66% of their ELA instruction time and 75% of their mathematics instruction time. More recent studies have noted great variability across classrooms in how students spend their time in mathematics and ELA instruction (Hiebert et al., 2003; Phelps et al., 2012; Pianta et al., 2007).

2.1.2. Discourse

The discourse labels focused on students' participation in classroom talk. These labels included on task student talking with other students and students raising their hand. These labels were included because of recent educational reform movements emphasizing providing students opportunities to learn disciplinary content through interaction in a learning community (National Council of Teachers of English & International Reading Association, 1996; National Council of Teachers of Mathematics, 2000). Furthermore, studies have shown students engaging with other's ideas leads to positive outcomes such as student achievement or performance on disciplinary practices (Applebee et al., 2003; Barron, 2003; Bishop, 2021; Cobb et al., 1992; Goodwin et al., 2021; Howe et al., 2019; Murphy et al., 2009; Nussbaum, 2008; Resnick et al., 2018; Sedova et al., 2019; Webb et al., 2014, 2021).

Prior research suggests that certain participation structures exist in classrooms. Typically, the teacher is the dominant participant in the

classroom talk and there is a recitation pattern of the teacher initiating a question, students responding, and then the teacher following up on the students' response or sometimes referred to as IRE/F (Cazden, 1988; Edwards & Mercer, 1987; Kawanaka & Stigler, 1999; Mehan, 1979; Nystrand & Gamoran, 1997; Sinclair & Coulthard, 1975). Even though many discussion-based pedagogies exist to support teachers, this pattern is pervasiveness and persistent (Alexander, 2008; Howe & Abedin, 2013; Spillane & Zeuli, 1999). As such, student talk is typically less frequent with some studies reporting students usually contributing between 4% and 25% of classroom talk in the United States (Burns, 1984; Kawanaka & Stigler, 1999; Silverman et al., 2014).

2.1.3. Teacher and student location

The teacher and student location labels captured details about the teachers' and students' positions in the classroom or their movements; for instance, whether the teacher was standing or sitting and if students were sitting at individual desks or at group tables. These labels were included given prior research on the organization of classrooms (e.g., Fernandes et al., 2011) and spatial pedagogy (Lim et al., 2012), and the recent advancements in multimodal learning analytics (e.g., Chan et al., 2020; D'Mello et al., 2015; Prieto et al., 2018). Patterns in a teacher's position in the classroom and proximity to students may be an influential factor for student participation and student motivation (Chan et al., 2020; Hur & Bosch, 2022; Yan et al., 2022).

Again, the review by Burns (1984) provided some insights into teacher and student locations. For instance, in one of the studies reviewed (Lambert & Hartsough, 1976), second-grade teachers in mathematics were observed on average spending about 12% of their time circulating around the room, but in reading, the teachers spent about 5% of their time circulating around the room. There were also grade level differences in some aspects of the teacher's location. Second-grade teachers spent on average between 0.40% and 0.45% of the lesson time sitting at their desk in reading and mathematics lessons, but fifth-grade teachers spent on average between 5.65% and 5.93% of the lesson time at their desk, respectively. In another study reviewed by Burns, students spent 9% and 5% of their class time sitting and walking, respectively (Good & Beckerman, 1978).

2.1.4. Teacher supporting

The teacher supporting labels were instances when the teacher aided students while engaged in an academic task. Sometimes this support could be verbal such as offering spoken feedback to students or non-verbal such as looking over students' shoulders to monitor their progress. Research suggests links between teacher support and student factors such as engagement and achievement (Dietrich et al., 2015; Hattie & Timperley, 2007; Klem & Connell, 2004; Marks, 2000; Roorda et al., 2011).

Large-scale observation studies of classrooms in the United States have found considerable variation in the nature, quality, and quantity of teachers' interactions with students (e.g., Burchinal et al., 2008; Pianta et al., 2007). Generally, elementary students have few opportunities to extensively interact with their teacher. Some process-product research studies reported the duration of interactions between a teacher and students ranging between 7% and 46% of the class time and some reporting more teacher-student interactions in reading instruction than mathematics (Burns, 1984).

2.1.5. Representing content

Many instructional resources exist in classrooms, ranging from teacher-created worksheets, to commercially produced texts, to web-based tools, software, and videos. The representing content labels included occasions when the teacher or students were holding or interacting with instructional resources (e.g., book, worksheet, or instructional tool), generating content such as through their writing, or displaying content such as on an interactive whiteboard or projector screen. Accessibility to high-quality instructional resources is essential to students' opportunities to learn (Chiu & Khoo, 2005; Oakes & Saunders, 2004). Standards documents explicitly recommend that students have access to instructional tools during instruction (e.g., National Council of Teachers of Mathematics, 2000).

Some classroom observation studies and teacher survey studies provided some insights into duration or frequency of representing content. In one study, it was found that elementary students spent most of their time writing or reading, on average about 22% of their time writing and 12% of their time reading (Good & Beckerman, 1978). Hiebert and Stigler (2000) reported that "reform" mathematics teachers in the United States were less likely to use a textbook during lessons in comparison to their non-reformed peers. A more recent teacher survey study suggests that most elementary teachers reported having adequate instructional resources (e.g., instructional technology, measurement tools, manipulatives, etc.; see Banilower et al., 2018). From that same survey, 35% of elementary teachers reported that their school provided a student with a laptop or tablet and 89% reported having access to a classroom set of laptops or tablets. When asked about their most recent mathematics lesson, 65% of elementary teachers reported students used manipulatives and 77% reported students completed textbook/worksheet problems.

2.2. Related automated efforts in classrooms

Efforts to automatically detect instructional activities in classroom videos has emerged as an area of research. Several researchers have been successful in detecting activities occurring in audio signals or transcripts from classroom videos such as teacher questioning and feedback (Dale et al., 2022; Kelly et al., 2018), revoicing or taking up student contributions (Dale et al., 2022; Demszky & Hill, 2022; Jacobs et al., 2022), and activity type (Wang et al., 2014). Even though these previous studies use slightly different approaches, they all have reported models with high accuracy rates. For instance, Wang et al. (2014) reported that their automated system correctly matched human performance to detect the activity type about 80% of the time.

Other scholars have explored detecting activities occurring in the video signals, rather than audio or transcripts, from classroom videos. These efforts have primarily focused on building annotated video datasets for computer vision techniques. The EduNet (Sharma et al., 2021) and Student Class Behavior Dataset (Sun et al., 2021) are two examples of efforts to build annotated video datasets from classroom videos. EduNet uses an annotation labeling scheme that is teacher- and student-centric. For example, writing on the board and holding a book are two teacher-focused labels; raising hand and sitting at desk are two student-focused labels. Training a two-stream I3D-ResNet-50 model on the EduNet dataset, Sharma and colleagues found an overall accuracy rating of 72.3%. Student Class Behavior Dataset features a student-centric annotation scheme (e.g., listening, using a computer, and raising hand). Sun and colleagues found an overall accuracy rating of 73.5% using the spatial stream of ResNet-101.

To date, EduSense is the only real-time video automation deployable system for the classroom (Ahuja et al., 2019). Using two wall-mounted cameras, one teacher-facing camera in the back of the room and one student-facing camera in the front of the room, the system can detect such instructional activities such as whether the student or teacher is standing or sitting, whether students are raising their hands, and the facial features and body pose of students and teachers. Testing EduSense in real university classrooms, the developers found high levels of accuracy. For instance, EduSense was able to detect with over 90% accuracy whether the university instructor was sitting or standing.

The findings from these studies using the EduNet, Student Class Behavior Dataset, and EduSense suggest the feasibility of video-based automation methods for detecting classroom-based activities. Furthermore, these studies acknowledged detecting complex activities and activities of lengthy duration is still a challenge. It is also important to consider these studies' limitations. First, the video datasets are small.

Most video datasets for developing automated systems are hundreds of hours (Beddiar et al., 2020). Second, the activities are mostly student-centric and are primarily simple in nature and short in duration (e.g., student raising hand). The datasets did not contain complex activities or activities that were longer in duration and thus presented a challenge for neural network model development.

Complex and long-duration activities could be more informative for teachers (e.g., providing insights into how much time in a lesson is taken up by transitions between activities). Furthermore, the emphasis on student-centric labels limits potential future applications of the automation that examines instructional quality. For instance, those researchers developing automated methods from transcripts have found that certain features in teachers' speech are correlated with certain classroom observation scores and student learning outcomes (e.g., Demszky & Hill, 2022). And in the case of EduSense system, the requirement of two, high-mounted video recordings may be limiting for some settings.

Given these related automated efforts in classrooms, we hypothesized the following outcomes for our study: (1) The more complex instructional activities in our annotation scheme would be harder to detect by the neural networks; (2) The more prevalent an instructional activity was across the dataset then the more likely the neural networks would perform accurately in detecting that instructional activity; and (3) The more frequent an instructional activity then the more likely the neural network would perform accurately in detecting that instructional activity.

3. Methods

3.1. Video dataset

The dataset came from a prior research project: the Developing Ambitious Instruction (DAI) project (Youngs et al., 2022). The DAI followed 83 graduates from five elementary teacher preparation programs into their first two to three of elementary (K-5) teaching. Up to 3 mathematics lessons and 3 ELA lessons were recorded for each teacher for every year of observation. The DAI dataset resulted in approximately 1000 h of video recordings.

From the DAI dataset, we selected a subset of approximately 244 h of video recordings from 80 graduates which yielded 279 lessons. Of those lessons, 140 were ELA lessons and 139 were mathematics lessons. On average, the videos were about 54 min in duration. The ELA videos had more variability in their duration than mathematics (see Fig. 1). Few lessons recordings were less than 25 min, and most lesson recordings were between 25 and 75 min. There were a few ELA videos greater than 90 min.

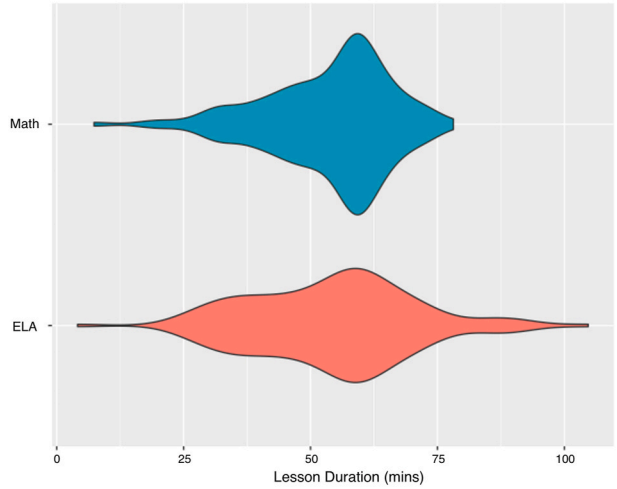


Fig. 1. Violin plot for lesson duration for ELA and mathematics lessons.

3.2. Human annotation of video dataset

To develop a large, annotated video dataset, we used a part-to-whole deductive approach (Erickson, 2006). Human annotators exhaustively identified and annotated every second in which an instructional activity occurred. Instructional activities less than a second in duration were excluded. Annotations were created using a free and open-source computer software called ELAN (ELAN, 2021), a software capable of tiered multi-label annotations of videos. ELAN has been used in video-based research studies across disciplines (Wittenburg et al., 2006), including education (de Freitas et al., 2017).

Fig. 2 is an example of the ELAN interface. A video player displays video in the top left with video playback controls to the right. Below the video player and playback controls there is a video timeline with a tiered multilabel system for an annotator to select start and stop time for each instructional activity. As the video plays, a red vertical bar runs along the timeline. For instance, in the videoframe in Fig. 2, we see the teacher sitting in a chair and students sitting on the carpet during a whole class lesson; in the annotation timeline, the red vertical line overlaps with the labels of whole class activity under activity types, sitting under teacher location, and sitting on the carpet/floor under student location.

3.3. Descriptive summary of DAI-244 dataset

The classroom videos were collected in elementary classrooms with instructional activities as they naturally occurred. Videos were annotated in their entirety without alternation or sampling of specific activities. Therefore, the classes of instructional activities in the annotated dataset are imbalanced, but typical of instructional activities in elementary classrooms in the United States. Fig. 3 lists the most frequent instructional activity, in terms of duration, from the top (i.e., teacher sitting at 120 h) to the bottom (i.e., on task student talking with student at 8 h).

Many of the instructional activities (13 of 24) were prevalent across the dataset; that is, appearing in at least 70% of the 279 lesson videos (see Table 2). A considerable minority of the instructional activities (9 of 24) were somewhat prevalent across lesson videos appearing in at least 40% but less than 70% of the lesson videos. Only two instructional activities (on task student talking with student and using or holding notebook) appeared in just under 40% of the lesson videos.

The duration of the instructional activities varied to some degree across videos. All instructional activity labels appeared in some lesson videos where they had a cumulative duration of less than 1 min and some videos with cumulative durations of up to 90 min. For instance, consider the individual activity label in the violin plot for instructional activity type in Fig. 4. Most individual activity durations ranged between 0 and 10 min and some continued for 20 min, but few lesson videos with individual activity persisted longer than 20 min.

3.4. Inherent challenges in the DAI-244 dataset

As our goal is to develop a dataset for training neural network models to detect instructional activities in elementary classrooms regardless of condition of the environment, we recognized the imbalance of the annotated dataset presented challenges as neural networks favor majority labels. One example of this imbalance is seen with the teacher location labels as shown in Fig. 3. Teacher sitting, on average, was almost seven times the size of teacher walking. Therefore, the neural networks may systematically under select those instructional activity labels with fewer hours (i.e., less than 50 h).

Another potential challenge for neural network models was how prevalent the instructional activities were across lesson videos. If a particular instructional activity only came from a limited selection of lessons, then the dataset may not be robust enough for training the

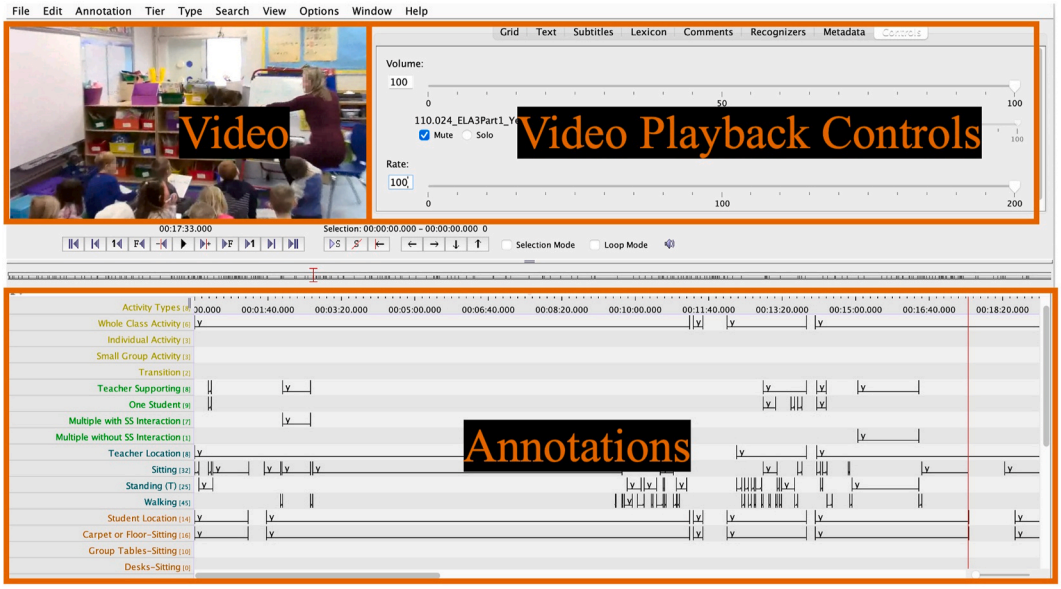


Fig. 2. ELAN annotation tool interface.

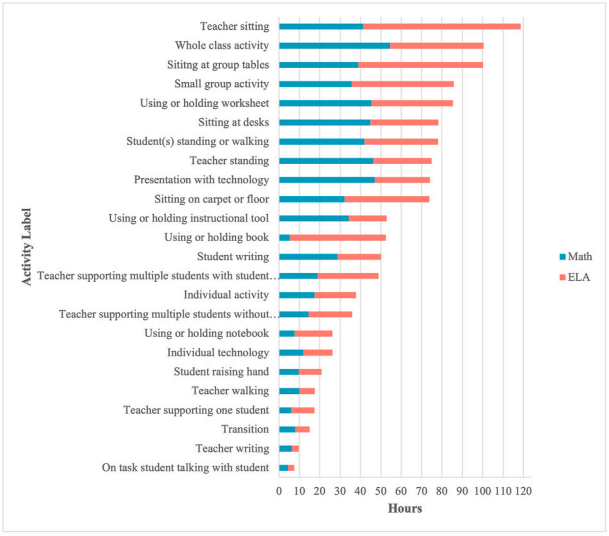


Fig. 3. Cumulative hours of instructional activities in the ETPP-244 dataset.

neural networks to detect instructional activities across a range of classroom videos. For instance, neural networks may infer instructional activities from similar environments or actors rather than the instructional activity of interest. This is a potential concern in the case of using or holding a notebook as the activity appeared in a little less than 40% of the classroom video lessons. Neural networks may systematically under select those instructional activity labels appearing less prevalent across lesson videos.

3.5. Neural network models

Next, we describe the neural networks and the approach used to evaluate their performance to detect the 24 instructional activities given the challenges of the imbalanced dataset and the complexity inherently found in classroom videos.

3.5.1. Background suppression network

Classroom scenes are often crowded with distracting details. In classroom videos, the foreground typically provides valuable information while the background is often irrelevant. As an example, Fig. 5 highlights regions in the foreground that include the teacher and student sitting at a group table examining a clock but ignores background scene information such as the backpacks and coats hanging on the wall. From an earlier pilot study, we found The Background Suppression Network (BaS-Net; Lee et al., 2020) performed better than other state-of-the-art neural network models (Korban et al., 2023). The BaS-Net emphasizes the foreground over the background for each videoframe. Given the initial positive outcomes of BaS-Net in detecting

Table 2
Frequency and typical duration of the instructional activities in the dataset.

Instructional Activity Label	# Lessons	Mean Duration	Max Duration
Activity Type			
Whole class activity	238	22 min	62 min
Individual activity	149	8 min	64 min
Small group activity	188	19 min	75 min
Transition	250	3 min	12 min
Discourse			
On task student talking with student	109	2 min	36 min
Student raising hand	263	5 min	24 min
Teacher Location			
Teacher sitting	261	26 min	89 min
Teacher standing	262	17 min	79 min
Teacher walking	258	4 min	27 min
Student Location			
Student(s) sitting on carpet or floor	209	16 min	83 min
Student(s) sitting at group tables	204	22 min	77 min
Student(s) sitting at desks	169	17 min	81 min
Student(s) standing or walking	270	17 min	58 min
Teacher Supporting			
One student	199	4 min	43 min
Multiple students with student interaction	156	11 min	76 min
Multiple students without student interaction	134	8 min	69 min
Representing Content			
Using or holding book	154	12 min	79 min
Using or holding worksheet	220	19 min	70 min
Presentation with technology	189	17 min	90 min
Using or holding instructional tool	177	12 min	59 min
Using or holding notebook	106	6 min	75 min
Individual technology	165	6 min	73 min
Teacher writing	241	2 min	12 min
Student writing	248	11 min	49 min

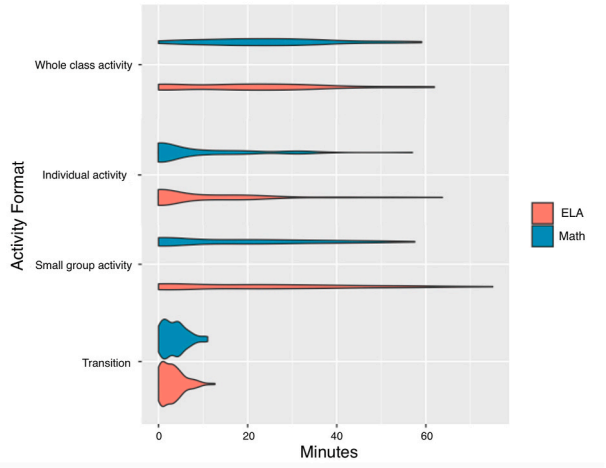


Fig. 4. Violin plot for activity format durations for ELA and mathematics lessons.

instructional activities in the pilot, we decided to further evaluate the performance of BaS-Net on a larger dataset.

3.5.2. Enhancing background suppression network for improved classroom detection

After evaluating the performance of BaS-Net on a larger dataset (see Section 4.1), we decided to enhance BaS-Net for classroom videos. These enhancements included three components (see Korban et al., 2023 for more details). First, we added a new loss function that covers both frame- and sequence-level predictions. Such a loss function is essential to model short and lengthy instructional activities. Second, an adaptive frame sampling based on important keyframes was included to remove irrelevant temporal dependencies between non-important frames. This adaptive frame sampling is particularly useful for processing long instructional activities in videos by making it more efficient. Third, a motion enhancement algorithm was included to boost the quality of motion features in the classroom videos due to camera movements, which occurred with some frequency across the dataset, that can reduce the quality of the video data.

3.6. Procedure and evaluation metrics

In our experimental setup, the proportions of training and testing sets were 80% and 20%, respectively. For the improved background suppression model, six convolutional layers were used. The learning rate was 0.00001 which was decayed by 0.1, for every 1500 iterations (from the total number of 7500 iterations). All the experiments were conducted using PyTorch 1.7 on a PC with dual Nvidia RTX 3090 GPUs (24 GB VRAM), AMD Ryzen Threadripper 3990X 64-Core Processor, and 256 GB of RAM.

To evaluate the two neural networks, we used accuracy and F1 measurements. Accuracy is the percentage of correct predictions relative to the total number of videoframes. The advantage of accuracy is that it is easily interpreted but does not provide a robust measurement for imbalanced datasets. F1 measures the precision and robustness of the classification; it is the harmonic mean of precision and recall. F1 as a performance measurement has some advantages over accuracy when the dataset is imbalanced. For accuracy and F1, the closer the measurement is to 1, the better the performance and the closer to 0, the worse the performance. To further explicate the performance of the two neural networks, we provide a comparison of the starting and ending frames for activity types from one classroom video.

4. Results

We present our results in three parts. First, we describe the performance of a baseline neural network, BaS-Net. Second, we describe the performance of our improvements to the baseline neural network, which we call BaS-Net+. Finally, we illustrate the boost in performance between BaS-Net and BaS-Net+ with a video example.

4.1. Performance of the background suppression network

The overall performance of the BaS-Net is shown in Fig. 6. The unweighted average F1-scores across the instructional activity labels was 0.47. The average F1-scores by parent-level instructional activity were as follows: activity type 0.42, discourse 0.24, representing content 0.41, student location 0.68, teacher location 0.57, and teacher supporting 0.40. Next, we further explicate the results for each instructional activity by parent-level activity.

4.1.1. Activity type

BaS-Net performed somewhat well in detecting the activity type in a classroom video with an average unweighted F1-score of 0.42. This was somewhat expected given the complexity of the activity (i.e., hypothesis 1). However, comparing the F1-scores for each activity type label reveals some differences. Small group activity was the only activity format label for which BaS-Net achieved an F1-score greater than 0.4. The F1-score for whole class activity is somewhat surprising considering it was the second most frequent activity with over 100 h; we expected BaS-Net to be somewhat biased towards it (i.e., hypothesis 3). BaS-Net performed similarly in detecting individual activity and transition, even though the duration of these activities was low. This suggests BaS-Net may struggle with detecting complex activities that can be relatively short and long in

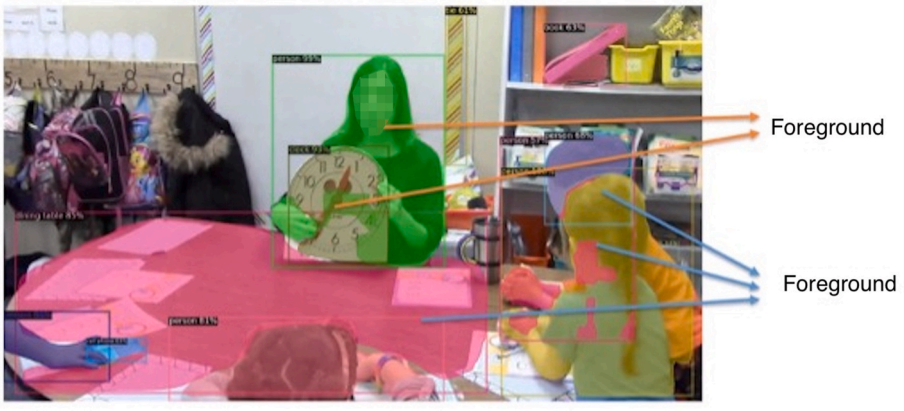


Fig. 5. Example foreground highlighted in a classroom scene.

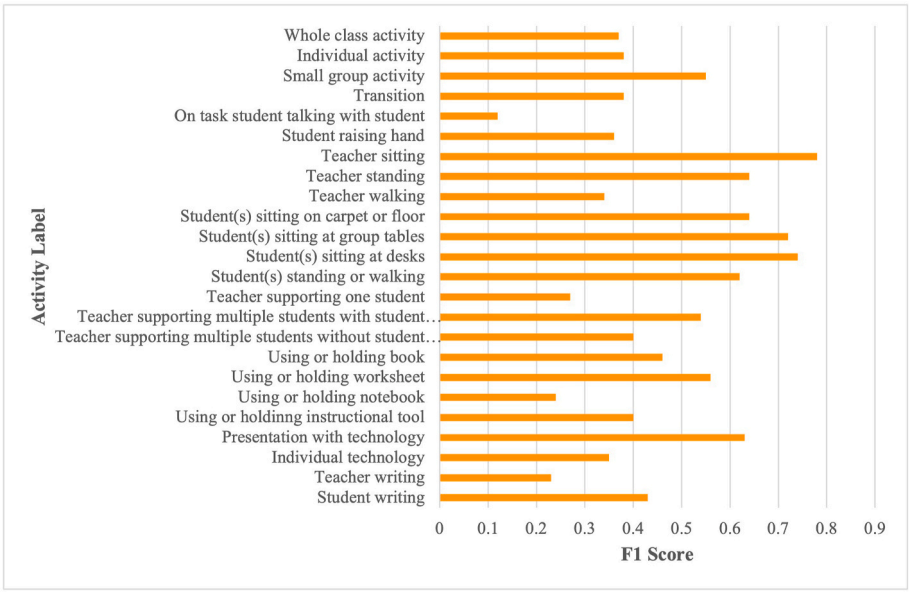


Fig. 6. F1-scores by instructional activity labels for BaS-Net.

duration (e.g., transition and whole class activity).

4.1.2. Discourse

BaS-Net performed poorly for the discourse labels. On task student talking with student and raising hand received F1-scores of 0.12 and 0.36, respectively. These results were somewhat expected given the occurrences of these labels were less than 25 h (i.e., hypothesis 3). Given student talking to another student was less prevalent (i.e., hypothesis 2) and the complexity of this action (i.e., hypothesis 1), the fact that BaS-Net performed the worst for that label matched expectations.

4.1.3. Representing content

With the representing content labels, the BaS-Net performed moderately well by reaching an average, unweighted F1-score of 0.41. BaS-Net performed particularly well in detecting presentation with technology with a F1-score of 0.63 and performed moderately well in detecting student writing, using or holding a book, using or holding an instructional tool, using or holding worksheet, and student writing. However, BaS-Net performed somewhat poorly in detecting individual technology, teacher writing, and using or holding notebook.

Overall, these results were as expected. Presentation with technology was only second to using or holding a worksheet in terms of prevalence and both appeared somewhat often. Thus, it was not too surprising BaS-Net performed moderately well. However, it was somewhat surprising that presentation with technology edged out using or holding a worksheet given the frequency of using or holding a worksheet (i.e., hypothesis 3). BaS-Net likely struggled to differentiate using or holding a worksheet from using or holding a book or notebook and was likely biased towards using or holding worksheet. Given that individual technology and using or holding notebook were less prevalent (i.e., hypothesis 2), it was expected that BaS-Net would likely perform poorly for those activities.

4.1.4. Teacher and student location

BaS-Net performed better in detecting student location activities versus teacher location. For all student location activities, BaS-Net attained F1-scores greater than 0.6. In contrast, BaS-Net attained F1-scores greater than 0.6 for all activities in teacher location except for teacher walking. This result is fairly expected given the lesser extent of teacher walking compared to the other teacher and student locations. However, given that almost all teacher and student locations labels were frequent and well represented, we would have expected similar performances. A plausible reason for the performances is BaS-Net may have experienced some difficulty differentiating between the teacher and the students. For instance, it may have detected when an actor was walking but not be able to differentiate the actor (i.e., teacher versus student).

4.1.5. Teacher supporting

BaS-Net's performance was mixed for detecting teacher supporting activities. It performed moderately well in detecting when the teacher was supporting multiple students with an average, unweighted F1-score of 0.47. When detecting if the teacher was supporting one student, however, the performance was low (0.27). Given that there were approximately 5 h of video with the teacher supporting multiple students for every hour the teacher was supporting one student, the mixed performance for teacher supporting activities was to be expected (i.e., hypothesis 3). Even though teacher supporting was prevalent in the classroom videos, the complexity of the teacher supporting activities and their low frequency likely contributed to the middling performance.

4.1.6. Rational for improving the background suppression network

With the high prevalence and frequency across classroom videos for some instructional activity type labels (e.g., whole class and small group), we postulated these labels would perform better than others. However, this outcome was not the case. Due to the potential impact of the complexity of these activity types, we decided to optimize BaS-Net for detecting these instructional activity types; we anticipated the performance could be improved by 30% or more. We decided to prioritize activity type labels because of the positive relationship between instructional time and student achievement (Baker et al., 2004; Brophy & Good, 1984), the popularity of manually identifying these labels in large-scale video studies (e.g., Hiebert et al., 2003), and detecting the primary activity type may have implications for certain classroom observation ratings (Luoto et al., 2023).

4.2. Performance of the background suppression network for improved classroom detection

Next, we describe the performance of our improvements to BaS-Net (i.e., BaS-Net+) for detecting activity types. Overall, BaS-Net+ performed well in detecting instructional activity types with an average, unweighted F1-score of 0.62. It performed exceptionally well at

detecting small group activity with a F1-score of 0.75. The F1-scores for whole class activity, individual activity, and transition were 0.57, 0.61, and 0.53, respectively (see Fig. 7). Using BaS-Net+, the accuracy for whole class activity, small group activity, individual activity, and transition was 0.88, 0.84, 0.89, and 0.93, respectively (see Fig. 8). These accuracy rates are similar to automated methods using classroom audio recordings (Wang et al., 2014).

4.3. Performance comparison

In comparison to BaS-Net, the improvements made with BaS-Net+ increased the performance, as measured by F1-scores, between 36% and 60% depending on the activity type (see Fig. 7). The most substantial boosts in performance, by over 50%, were for detecting whole class activity and individual activity. Thus, the improvements in BaS-Net+ were useful for relatively short activity types (i.e., transitions) and longer activity types (i.e., whole class activity). Next, we illustrate these improvements using videoframes from a classroom video of a first-grade ELA lesson.

Fig. 9 compares the performance of BaS-Net and BaS-Net+ on the starting and ending frames of three consecutive instructional activity types in a classroom video as identified by a human annotator. The first activity type, individual activity, occurs about 3 min into the video and lasts for about 3 min. Students are pasting a poem into their notebooks and then circling sight words as they appear in the poem. The teacher is walking around the classroom monitoring and assisting students. The second activity format is a transition. Students begin walking over to the large, carpeted rug and sitting down while the teacher monitors students and gathers lesson materials. The transition lasts for approximately 2 min. The third activity is a whole class activity that lasts about 15 min. The teacher begins by reviewing question words with students. Then, the teacher informs students that they will be reading a short story and then directs students to discuss with a partner what possible questions they might ask about the story (e.g., Who are the characters?). After reading the story, the teacher and students discuss several questions about the story with the teacher summarizing student responses on the board.

As shown in Fig. 9, BaS-Net+ outperformed BaS-Net in predicting the activity type at the starting and ending frames. BaS-Net+ improved the prediction score of these activity formats at the starting frames. However, for this video, there were only slight improvements in the prediction scores at the ending frames for individual activity and whole class activity. In contrast, BaS-Net+ noticeably improved the prediction score at the starting and ending frames for transitions.

5. Discussion

In this study, we examined whether a neural network (BaS-Net) was able to detect instructional activities during elementary classroom instruction captured by video recording with accuracy levels comparable to that of manual human annotations. We speculated BaS-Net would be well-situated for classroom videos as classroom scenes often contain cluttered backgrounds, but we were not optimistic that BaS-Net would perform well with more complex activities (e.g., teacher support multiple students and small group activity). Our experiment with BaS-Net confirmed our hypotheses. Taken from what we learned by applying BaS-Net, we enhanced it to optimize the performance for detecting activity formats (i.e., BaS-Net+), as this agnostic measure has several practical and methodological implications. We found BaS-Net+ to be an improvement in detecting all activity types over BaS-Net. The accuracy of BaS-Net+ to detect activity format is comparable to an automated method using audio recordings of classrooms (Wang et al., 2014).

Previous studies evaluating automated efforts for classroom video observations have primarily used audio recordings or transcripts from videos to detect instructional activities (e.g., Dale et al., 2022; Demszky & Hill, 2022; Jacobs et al., 2022; Kelly et al., 2018; Wang et al., 2014) and those studies that used video have primarily focused on simple actions of students such as students raising their hands (e.g., Ahuja et al., 2019; Sharma et al., 2021; Sun et al., 2021). A limitation for video automated efforts has been the absence of large, annotated video datasets specifically for machine learning development. As part of this study, we developed a large, high-quality video annotated dataset with 244 h of annotated classroom videos with instructional activity labels at various levels of complexity. This dataset is larger and more comprehensive than other similar datasets (Sharma et al., 2021; Sun et al., 2021).

5.1. Implications for teaching and learning

Our results provide further evidence of the validity of automated efforts to document instructional activities in classroom videos. In particular, BaS-Net+ performance was comparable with human annotations for detecting activity formats: whole class activity, small group activity, individual activity, and transition. As such, the application of BaS-Net+ to classroom videos could lead to several implications for teaching and learning. A teacher-facing application using BaS-Net+

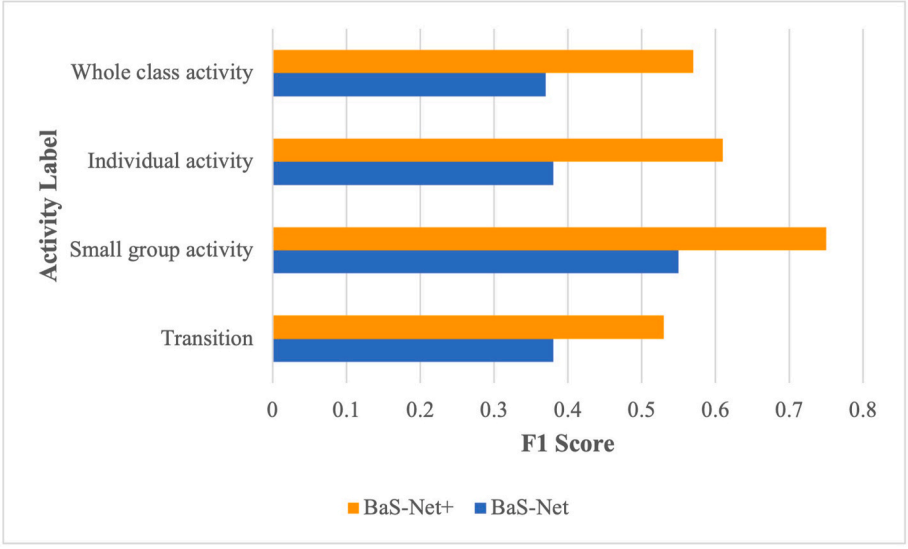


Fig. 7. Comparison of BaS-Net and BaS-Net+ performance.

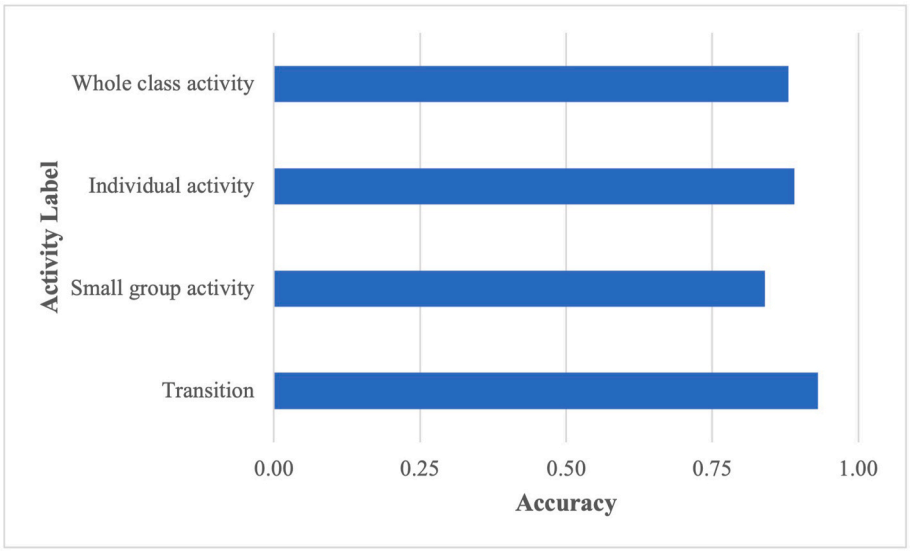


Fig. 8. BaS-Net+ performance.

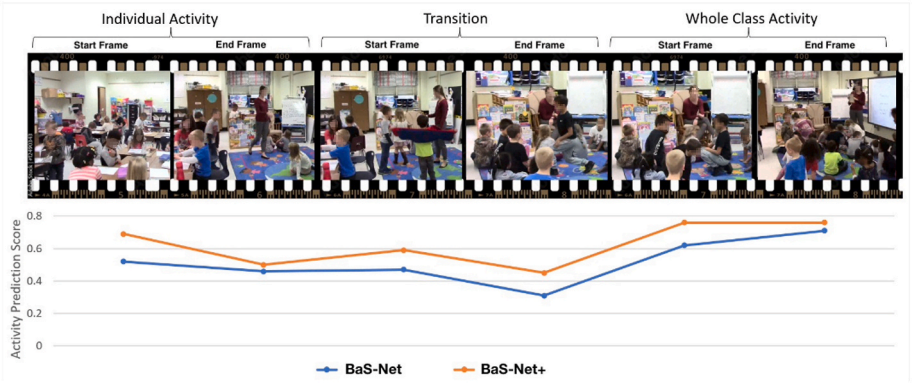


Fig. 9. Comparison of BaS-Net and BaS-Net+ on the starting and ending frames for three consecutive activity formats.

could process teachers' classroom videos the same day to provide analytics on the frequency and duration of activity types in their lessons without the need for specialized equipment. For instance, teachers adjusting a routine could receive data whether the adjustment led to reduced transition time. Second, deploying BaS-Net+ may make future large-scale video studies more efficient and provide insights into differences in instructional time and activity types across contexts such as in the case of international comparisons (e.g., Hiebert et al., 2003). The efficiency of BaS-Net+ compared to human annotations could bring reform and intervention efforts more readily to teachers and their students. Relatedly, BaS-Net+ has the potential to support future research that calls into question whether instructional time and activity types have implications for teaching and learning such as student engagement (e.g., Kelly & Turner, 2009) and measuring teaching quality (e.g., Luoto et al., 2023).

5.2. Contributions to using emerging video technologies in research

Previous investigations have not explored whether video-based automation endeavors could assist in facilitating data construction and analysis for educational research. This study suggests that emerging video technologies (i.e., computer vision and neural networks) can extract and construct some data near to the level of humans. These emerging technologies may support large-scale learning analytics research in classrooms that have been limited by excessive costs and time limitations. For instance, video recordings of classrooms across an entire school year could be completed and analyzed within that same year and thus potentially provide valuable insights for educational researchers and teachers. However, there are still some unknown consequences for education with these emerging video technologies.

5.3. Limitations

There are some limitations of the study. First, the neural networks were applied only to elementary classroom videos of mathematics and ELA instruction in the United States. It is unknown how these neural networks may perform on other videos. Second, these neural networks do not engage with the content of teachers' or students' speech. While the neural networks could accurately provide an estimate for quantity of instruction time spent in whole class, they could not provide any metrics about instructional quality. Therefore, the application of these neural networks should not be used for evaluative purposes. Furthermore, these instructional activity labels are agnostic, but it is still debatable how and to what extent these labels scale to qualitative recommendations; for instance, whether small group instruction is engaging for students (Kelly & Turner, 2009).

We are therefore cautious in recommending the immediate application of these technologies without fully understanding the implications. Nevertheless, we foresee these emerging video technologies as part of the history of other video technologies for education research and

like others (e.g., Goldman et al., 2014; Hennessy et al., 2020; Kelly, 2023) are optimistic about the potential for these tools to aid in our understanding of teaching and learning.

6. Conclusion

In conclusion, this study reported on the application of neural networks to classify instructional activities using video signals from a collection of classroom videos from elementary mathematics and English language arts instruction in the United States. The neural networks detected instructional activities in the classroom videos at a high rate of accuracy. This result suggests neural networks could become an important technological tool for contributing to the construction of data and data analysis for classroom research. Additionally, these data from neural networks could be useful to efficiently bring classroom analytics to key communities and support efforts to reform teaching and learning in classrooms.

Statement on open data and ethics

The DAI-244 Dataset is not available due to the nature of the consent form at the time of video data collection. The video data comes from a prior study: The Development of Ambitious Instruction. The overall research project had ethics approval from the Institutional Review Board for Social and Behavioral Sciences at the University of Virginia.

CRediT authorship contribution statement

Jonathan K. Foster: Writing – review & editing, Writing – original draft, Visualization, Resources, Investigation, Formal analysis, Data curation. **Matthew Korban:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Peter Youngs:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Ginger S. Watson:** Writing – original draft, Validation, Supervision, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization. **Scott T. Acton:** Validation, Supervision, Software, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Science Foundation [Grant No. 2000487] and the Robertson Foundation [Grant No. 9909875]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

References

Ahuja, K., Kim, D., Xhakaj, F., Varga, V., Xie, A., Zhang, S., Townsend, J. E., Harrison, C., Ogan, A., & Agarwal, Y. (2019). Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3*(3), 1–26. https://doi.org/10.1145/3351229

Alexander, R. J. (2008). In *Towards dialogic teaching: Rethinking classroom talk* (4th ed.). Dialogos.

Andersson, E., & Sørvik, G. O. (2013). Reality lost? Re-Use of qualitative data in classroom video studies. *Forum Qualitative Sozialforschung, 14*(3). https://www.duo.uio.no/handle/10852/50551.

Applebee, A. N., Langer, J. A., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal, 40*(3), 685–730.

Baker, D. P., Fabrega, R., Galindo, C., & Mishook, J. (2004). Instructional time and national achievement: Cross-national evidence. *Prospects, 34*(3), 311–334. https://doi.org/10.1007/s11125-004-5310-1

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M. L. (2018). *Report of the 2018 NSSME+*. Horizon Research, Inc.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*(3), 307–359. https://doi.org/10.1207/S15327809JLS1203_1

Beddiar, D. R., Nini, B., Sabokrou, M., & Hadid, A. (2020). Vision-based human activity recognition: A survey. *Multimedia Tools and Applications, 79*(41), 30509–30555. https://doi.org/10.1007/s11042-020-09004-3

Berry, R. Q., Rimm-Kaufman, S. E., Ottmar, E. M., Walkowiak, T. A., Merritt, E. G., & Pinter, H. H. (2013). *The mathematics scan (M-Scan): A measure of standards-based mathematics teaching practices*. University of Virginia.

Bishop, J. P. (2021). Responsiveness and intellectual work: Features of mathematics classroom discourse related to student achievement. *The Journal of the Learning Sciences, 30*(3), 466–508. https://doi.org/10.1080/10508406.2021.1922413

Bodovski, K., & Farkas, G. (2007). Do instructional practices contribute to inequality in achievement?: The case of mathematics instruction in kindergarten. *Journal of Early Childhood Research, 5*(3), 301–322. https://doi.org/10.1177/1476718X07080476

Borg, W. (1980). Time and school learning. In C. Denham, & A. Lieberman (Eds.), *Time to learn: A review of the beginning teacher evaluation study* (pp. 33–63). US Department of Education.

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education, 24*(2), 417–436. https://doi.org/10.1016/j.tate.2006.11.012

Brophy, J., & Good, T. L. (1984). *Teacher behavior and student achievement*. Institute for Research on Teaching, Michigan State University. https://eric.ed.gov/?id=ED251422.

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher-child interactions and instruction. *Applied Developmental Science, 12*, 140–153. https://doi.org/10.1080/10888690802199418

Burns, R. B. (1984). How time is used in elementary schools: The activity structure of classrooms. In *Time and school learning (1984)*. Routledge.

Carroll, J. B. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher, 18*(1), 26–31. https://doi.org/10.2307/1176007

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311–337. https://doi.org/10.1177/0013164414539163

Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning*.

Chan, M. C. E., Ochoa, X., & Clarke, D. (2020). Multimodal learning analytics in a laboratory classroom. In M. Virvou, E. Alepis, G. A. Tsihrintzis, & L. C. Jain (Eds.), *Machine learning paradigms: Advances in learning analytics* (pp. 131–156). Springer International Publishing. https://doi.org/10.1007/978-3-030-13743-4_8.

Chiu, M. M., & Khoo, L. (2005). Effects of resources, inequality, and privilege bias on achievement: Country, school, and student level analyses. *American Educational Research Journal, 42*(4), 575–603. https://doi.org/10.3102/00028312042004575

Cobb, P., Yackel, E., & Wood, T. (1992). Interaction and learning in mathematics classroom situations. *Educational Studies in Mathematics, 23*(1), 99–122.

Corr, M. K. (2011). *Investigating the reliability of classroom observation protocols: The case of PLATO*. Stanford University. http://platorubric.stanford.edu/Cor%20M%20K%20%20(2011).pdf.

Dale, M. E., Godley, A. J., Capello, S. A., Donnelly, P. J., D'Mello, S. K., & Kelly, S. P. (2022). Toward the automated analysis of teacher talk in secondary ELA classrooms. *Teaching and Teacher Education, 110*, Article 103584. https://doi.org/10.1016/j.tate.2021.103584

Dalland, C. P., Klette, K., & Svenkerud, S. (2020). Video studies and the challenge of selecting time scales. *International Journal of Research and Method in Education, 43*(1), 53–66. https://doi.org/10.1080/1743727X.2018.1563062

de Freitas, E., Lerman, S., & Parks, A. N. (2017). Qualitative methods. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 159–182). National Council of Teachers of Mathematics.

Demszky, D., & Hill, H. C. (2022). *The NCTE transcripts: A Dataset of elementary math classroom transcripts (EdWorkingPaper No. 22-682)*. Annenberg Institute at Brown University. https://doi.org/10.26300/npxh-kf69

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G., & Sherin, B. L. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *The Journal of the Learning Sciences, 19*(1), 3–53. https://doi.org/10.1080/10508400903452884

Dietrich, J., Dicke, A.-L., Kracke, B., & Noack, P. (2015). Teacher support and its influence on students' intrinsic value and effort: Dimensional comparison effects across subjects. *Learning and Instruction, 39*, 45–54. https://doi.org/10.1016/j.learninstruc.2015.05.007

D'Mello, S. K., Olney, A. M., Blanchard, N., Samei, B., Sun, X., Ward, B., & Kelly, S. (2015). Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 557–566). https://doi.org/10.1145/2818346.2830602

Edwards, D., & Mercer, N. (1987). *Common knowledge: The development of understanding in the classroom*. Methuen.

ELAN (Version 6.2) [Computer software]. (2021). *Nijmegen: Max planck institute for psycholinguistics, the language archive*. Retrieved from https://archive.mpi.nl/tla/elan.

Erickson, F. (2006). Definition and analysis of data from videotape: Some research procedures and their rationales. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.),

Handbook of complementary methods in education research (pp. 177–191). Lawrence Erlbaum Associates Publishers.
Fernandes, A. C., Huang, J., & Rinaldo, V. (2011). Does where a student sits really matter? - the impact of seating locations on student classroom learning. *International Journal of Applied Educational Studies, 10*(1), 66–77.
Franke, M. L., Kazemi, E., & Battey, D. (2007). Mathematics teaching and classroom practice. In F. K. Lester, & National Council of Teachers of Mathematics (Eds.), *Second handbook of research on mathematics teaching and learning: A project of the national Council of teachers of mathematics* (pp. 225–256). Information Age Publishing.
Gaudin, C., & Chaliès, S. (2015). Video viewing in teacher education and professional development: A literature review. *Educational Research Review, 16*, 41–67. https://doi.org/10.1016/j.edurev.2015.06.001
Gettinger, M. (1984). Individual differences in time needed for learning: A review of literature. *Educational Psychologist, 19*(1), 15–29. https://doi.org/10.1080/00461528409529278
Goldman, R., Zahn, C., & Derry, S. J. (2014). Frontiers of digital video research in the learning sciences: Mapping the terrain. In R. K. Sawyer (Ed.), *The cambridge handbook of the learning Sciences* (2nd ed., pp. 213–232). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.014.
Good, T. L., & Beckerman, T. M. (1978). Time on task: A naturalistic study in sixth-grade classrooms. *The Elementary School Journal, 78*(3), 193–201. https://doi.org/10.1086/461101
Goodwin, A. P., Cho, S.-J., Reynolds, D., Silverman, R., & Nunn, S. (2021). Explorations of classroom talk and links to reading achievement in upper elementary classrooms. *Journal of Educational Psychology, 113*(1), 27–48. https://doi.org/10.1037/edu0000462
Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293–303. https://doi.org/10.3102/0013189X14544542
Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445–470. https://doi.org/10.1086/669901
Gupta, N., Gupta, S. K., Pathak, R. K., Jain, V., Rashidi, P., & Suri, J. S. (2022). Human activity recognition in artificial intelligence framework: A narrative review. *Artificial Intelligence Review, 55*(6), 4755–4808. https://doi.org/10.1007/s10462-021-10116-x
Hall, R. (2000). Videorecording as theory. In A. E. Kelly, & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 647–664). Routledge. eBook Collection (EBSCOhost) https://proxy1.library.virginia.edu/login?url= https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=19353&site=ehost-live&scope=site.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487
Hennessy, S., Howe, C., Mercer, N., & Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture and Social Interaction, 25*, Article 100404. https://doi.org/10.1016/j.lcsi.2020.100404
Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., Chui, A. M.-Y., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E., Etterbeek, W., Manaster, C., Gonzales, P., & Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. National Center for Education Statistics.
Hiebert, J., & Stigler, J. W. (2000). A proposal for improving classroom teaching: Lessons from the TIMSS video study. *The Elementary School Journal, 101*(1), 3–20. https://doi.org/10.1086/499656
Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64. https://doi.org/10.3102/0013189X12437203
Howe, C., & Abedin, M. (2013). Classroom dialogue: A systematic review across four decades of research. *Cambridge Journal of Education, 43*(3), 325–356. https://doi.org/10.1080/0305764X.2013.786024
Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *The Journal of the Learning Sciences, 28*(4–5), 462–512. https://doi.org/10.1080/10508406.2019.1573730
Hur, P., & Bosch, N. (2022). Tracking individuals in classroom videos via post-processing OpenPose data. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 465–471). https://doi.org/10.1145/3506860.3506888
Jacobs, J. K., Kawanaka, T., & Stigler, J. W. (1999). Integrating qualitative and quantitative approaches to the analysis of video data on classroom teaching. *International Journal of Educational Research, 31*(8), 717–724. https://doi.org/10.1016/S0883-0355(99)00036-1
Jacobs, J. K., 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*, Article 103631. https://doi.org/10.1016/j.tate.2022.103631
Janík, T., & Seidel, T. (Eds.). (2009). *The power of video studies in investigating teaching and learning in the classroom*. Waxmann.
Kawanaka, T., & Stigler, J. W. (1999). Teachers' use of questions in eighth-grade mathematics classrooms in Germany, Japan, and the United States. *Mathematical Thinking and Learning, 1*(4), 255–278. https://doi.org/10.1207/s15327833mtl0104_1
Kelly, S. (2023). *Agnosticism in instructional observation systems, 31*. Education Policy Analysis Archives. https://doi.org/10.14507/epaa.31.7493
Kelly, S., Bringe, R., Aucejo, E., & Cooley Fruehwirth, J. (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
Kelly, S., & Turner, J. (2009). Rethinking the effects of classroom activity structure on the engagement of low- achieving students. *Teachers College Record, 111*(7), 1665–1692. https://doi.org/10.1177/016146810911100706
Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health, 74*(7), 262–273. https://doi.org/10.1111/j.1746-1561.2004.tb08283.x
Klette, K. (2022). The use of video capturing in international large-scale assessment studies: Methodological and theoretical considerations. In *International handbook of comparative large-scale studies in education* (pp. 470–510). Springer. https://www.duo.uio.no/handle/10852/101442.
Korban, M., Youngs, P., & Acton, S. T. (2023). Instructional activity detection using deep neural networks. In *2023 24th international conference on digital signal processing* (pp. 1–4). https://doi.org/10.1109/DSP58604.2023.10167935
Lee, P., Uh, Y., & Byun, H. (2020). Background Suppression Network for weakly-supervised temporal action localization, 07. In *Proceedings of the AAAI conference on artificial intelligence, 34*. https://doi.org/10.1609/aaai.v34i07.6793. Article 07.
Lim, F. V., O'Halloran, K. L., & Podlasov, A. (2012). Spatial pedagogy: Mapping meanings in the use of classroom space. *Cambridge Journal of Education, 42*(2), 235–251. https://doi.org/10.1080/0305764X.2012.676629
Luoto, J., Klette, K., & Blikstad-Balas, M. (2023). Possible biases in observation systems when applied across contexts: Conceptualizing, operationalizing, and sequencing instructional quality. *Educational Assessment, Evaluation and Accountability, 35*(1), 105–128. https://doi.org/10.1007/s11092-022-09394-y
Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal, 37*(1), 153–184. https://doi.org/10.3102/00028312037001153
Mehan, H. (1979). *Learning lessons: Social organization in the classroom*. Harvard University Press.
Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*, 740–764. https://doi.org/10.1037/a0015576
National Council of Teachers of English, & International Reading Association. (1996). In *Standards for the English language arts*. International Reading Association; National Council of Teachers of English.
National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. National Council of Teachers of Mathematics.
Newmann, F. M., & Associates. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. Jossey-Bass.
Nussbaum, E. M. (2008). Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology, 33*(3), 345–359. https://doi.org/10.1016/j.cedpsych.2008.06.001
Nystrand, M., & Gamoran, A. (1997). The big picture: Language and learning in hundreds of English lessons. In M. Nystrand (Ed.), *Opening dialogue: Understanding the dynamics of language and learning in the English classroom* (pp. 30–74). https://www.jstor.org/stable/417942?origin=crossref.
Oakes, J., & Saunders, M. (2004). Education's most basic tools: Access to textbooks and instructional materials in California's public schools. *Teachers College Record, 106*(10), 1967–1988. https://doi.org/10.1111/j.1467-9620.2004.00423.x
Pang, S., Lai, S., Zhang, A., Yang, Y., & Sun, D. (2023). Graph convolutional network for automatic detection of teachers' nonverbal behavior. *Computers and Education: Artificial Intelligence.* , Article 100174. https://doi.org/10.1016/j.caeai.2023.100174
Pea, R., & Hoffert, E. (2007). Video workflow in the learning sciences: Prospects of emerging technologies for augmenting work practices. In *Video research in the learning Sciences*. Routledge.
Phelps, G., Corey, D., DeMonte, J., Harrison, D., & Loewenberg Ball, D. (2012). How much English language arts and mathematics instruction do students receive? Investigating variation in instructional time. *Educational Policy, 26*(5), 631–662. https://doi.org/10.1177/0895904811417580
Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. (2007). Opportunities to learn in America's elementary classrooms. *Science, 315*(5820), 1795–1796. https://doi.org/10.1126/science.1139719
Prieto, L.p., Sharma, K., Kidzinski, Ł., Rodríguez-Triana, M.j., & Dillenbourg, P. (2018). Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data. *Journal of Computer Assisted Learning, 34*(2), 193–203. https://doi.org/10.1111/jcal.12232
Resnick, L. B., Asterhan, C. S. C., Clarke, S. N., & Schantz, F. (2018). Next generation research in dialogic learning. In *The Wiley handbook of teaching and learning* (pp. 323–338). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118955901.ch13.
Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81*(4), 493–529. https://doi.org/10.3102/0034654311421793
Rosenshine, B. V. (1981). How time is spent in elementary classrooms. *Journal of Classroom Interaction, 17*(1), 16–25.
Saba, L., Biswas, M., Kuppili, V., Cuadrado Godia, E., Suri, H. S., Edla, D. R., Omerzu, T., Laird, J. R., Khanna, N. N., Mavrogeni, S., Protogerou, A., Sfikakis, P. P.,

Viswanathan, V., Kitas, G. D., Nicolaides, A., Gupta, A., & Suri, J. S. (2019). The present and future of deep learning in radiology. *European Journal of Radiology, 114*, 14–24. https://doi.org/10.1016/j.ejrad.2019.02.038

Sedova, K., Sedlacek, M., Svaricek, R., Majcik, M., Navratilova, J., Drexlerova, A., Kychler, J., & Salamounova, Z. (2019). Do those who talk more learn more? The relationship between student classroom talk and student achievement. *Learning and Instruction, 63*, Article 101217. https://doi.org/10.1016/j.learninstruc.2019.101217

Sharma, V., Gupta, M., Kumar, A., & Mishra, D. (2021). EduNet: A new video dataset for understanding human activity in the classroom environment. *Sensors, 21*(17), 5699. https://doi.org/10.3390/s21175699

Silverman, R. D., Proctor, C. P., Harring, J. R., Doyle, B., Mitchell, M. A., & Meyer, A. G. (2014). Teachers' instruction and students' vocabulary and comprehension: An exploratory study with English monolingual and Spanish–English bilingual students in grades 3–5. *Reading Research Quarterly, 49*(1), 31–60. https://doi.org/10.1002/rrq.63

Sinclair, J. M., & Coulthard, M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford University Press.

Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis, 21*(1), 1–27. https://doi.org/10.3102/01623737021001001

Stallings, J. (1980). Allocated academic learning time revisited, or beyond time on task. *Educational Researcher, 9*(11), 11–16. https://doi.org/10.2307/1175185

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS Video Studies. *Educational Psychologist, 35*(2), 87–100. https://doi.org/10.1207/S15326985EP3502_3

Sun, B., Wu, Y., Zhao, K., He, J., Yu, L., Yan, H., & Luo, A. (2021). Student class behavior dataset: A video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. *Neural Computing & Applications, 33*(14), 8335–8354. https://doi.org/10.1007/s00521-020-05587-y

Thompson, J., Windschitl, M., & Braaten, M. (2013). Developing a theory of ambitious early-career teacher practice. *American Educational Research Journal, 50*(3), 574–615. https://doi.org/10.3102/0002831213476334

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics, 85*(1), 109–128. https://doi.org/10.1007/s10649-013-9499-x

Walkowiak, T. A., Berry, R. Q., Pinter, H. H., & Jacobson, E. D. (2018). Utilizing the M-Scan to measure standards-based mathematics teaching practices: Affordances and limitations. *ZDM, 50*(3), 461–474. https://doi.org/10.1007/s11858-018-0931-7

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

Webb, N. M., Franke, M. L., Ing, M., Wong, J., Fernandez, C. H., Shin, N., & Turrou, A. C. (2014). Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *International Journal of Educational Research, 63*, 79–93. https://doi.org/10.1016/j.ijer.2013.02.001

Webb, N. M., Franke, M. L., Johnson, N. C., Ing, M., & Zimmerman, J. (2021). Learning through explaining and engaging with others' mathematical ideas. *Mathematical Thinking and Learning, 0*(0), 1–27. https://doi.org/10.1080/10986065.2021.1990744

Wiley, D. E., & Harnischfeger, A. (1974). Explosion of a myth: Quantity of schooling and exposure to instruction, major educational vehicles. *Educational Researcher, 3*(4), 7–12. https://doi.org/10.3102/0013189X003004007

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the fifth international conference on language resources and evaluation*. Genoa, Italy: LREC 2006. http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf.

Xu, L., Aranda, G., Widjaja, W., & Clarke, D. (Eds.). (2018). *Video-based research in education: Cross-disciplinary perspectives* (1st ed.). Routledge. https://doi.org/10.4324/9781315109213.

Yan, L., Martinez-Maldonado, R., Zhao, L., Deppeler, J., Corrigan, D., & Gasevic, D. (2022). How do teachers use open learning spaces? Mapping from teachers' socio-spatial data to spatial pedagogy. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 87–97). https://doi.org/10.1145/3506860.3506872

Youngs, P., Elreda, L. M., Anagnostopoulos, D., Cohen, J., Drake, C., & Konstantopoulos, S. (2022). The development of ambitious instruction: How beginning elementary teachers' preparation experiences are associated with their mathematics and English language arts instructional practices. *Teaching and Teacher Education, 110*, Article 103576. https://doi.org/10.1016/j.tate.2021.103576