JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Early Access Articles, 1–18. https://doi.org/10.18608/jla.2024.8323

Automated Classification of Elementary Instructional Activities: Analyzing the Consistency of Human Annotations

Jonathan K. Foster[1], Peter Youngs[2], Rachel van Aswegen[3], Samarth Singh[4], Ginger S. Watson[5] and Scott T. Acton[6]

Abstract

Despite a tremendous increase in the use of video for conducting research in classrooms as well as preparing and evaluating teachers, there remain notable challenges to using classroom videos at scale, including time and financial costs. Recent advances in artificial intelligence could make the process of analyzing, scoring, and cataloguing videos more efficient. These advances include natural language processing, automated speech recognition, and deep neural networks. To train artificial intelligence to accurately classify activities in classroom videos, humans must first annotate a set of videos in a consistent way. This paper describes our investigation of the degree of inter-annotator reliability regarding identification of and duration of activities among annotators with and without experience analyzing classroom videos. Validity of human annotations is crucial for research involving temporal analysis within classroom video research. The study reported here represents an important step towards applying methods developed in other fields to validate temporal analytics within learning analytics research for classifying time- and event-based activities in classroom videos.

Notes for Practice

- We describe our configuration of an annotation tool and our use of the tool to identify and label instructional activities in videos of elementary mathematics and reading instruction.
- We offer methods for validating the classification of time- and event-based activities in classroom videos by human annotators. Validation efforts, such as these, are crucial in learning analytics research involving human annotations for temporal analysis.
- We report how we trained individuals who lacked experience analyzing videos of K–12 instruction to accurately annotate some low-inference labels of classroom activity.
- We report on inter-annotator reliability using several statistics including raw agreements, time- and event-based kappas, and positive and negative agreements.

Keywords: Video annotation, temporal analysis, elementary instruction, validation

Submitted: 12/10/2023 — **Accepted:** 07/21/2024 — **Published:** 29/09/2024

[1]*Email: jkfoster@albany.edu Address: Department of Educational Theory & Practice, University of Albany, Catskill 263, 1400 Washington Avenue, Albany, NY, 12222, USA. ORCID iD: https://orcid.org/0000-0002-7842-6277*
Corresponding author [2]Email: pay2n@virginia.edu Address: Department of Curriculum, Instruction & Special Education, University of Virginia, PO Box 400273, Charlottesville, VA, 22904, USA. ORCID iD: https://orcid.org/0000-0002-1711-1749
[3]Email: rachel.van.aswegen18@gmail.com Address: 2100 Lakeview Avenue, Richmond, VA, 23220, USA.
[4]Email: ss5xp@virginia.edu Address: Department of Electrical and Computer Engineering, University of Virginia, Thornton Hall, 351 McCormick Road, Suite C210, Charlottesville, VA, 22904, USA.
[5]Email: gswatson@odu.edu Address: Old Dominion University, Virginia Modeling, Analysis, & Simulation Center, 1030 University Boulevard, Suffolk, VA, 23435, USA. ORCID iD: https://orcid.org/0000-0001-7197-1654
[6]Email: acton@virginia.edu Address: Department of Electrical and Computer Engineering, University of Virginia, Thornton Hall, 351 McCormick Road, Suite C210, Charlottesville, VA, 22904, USA. ORCID iD: https://orcid.org/0000-0003-3288-1255

1. Introduction

In the last decade in the U.S., there has been a tremendous increase in the use of video for preparing and evaluating teachers in addition to conducting research on teaching quality (see, e.g., Kane et al., 2013; Hamre et al., 2012; SCALE, 2015). Despite this increased use, there remain notable challenges to using classroom videos at scale. These challenges include the time and financial cost involved in training humans to view and assess video recorded lessons and the time involved in labelling large volumes of video for later viewing. Recent advances in artificial intelligence may provide solutions to these challenges and could make the process of analyzing, scoring, and cataloguing videos more efficient. These advances include deep neural networks in applications such as natural language processing and computer vision. A deep neural network is a hierarchical learning structure that tends to learn more complex and abstract features of a given set of data in its deeper layers. With this structure, a deep neural network can approximate complicated decision functions that directly map input data to output labels. In addition, with deep neural networks, the learning process can improve the level of accuracy in classification as more annotated data are provided.

To train neural networks and other forms of artificial intelligence to accurately classify activities in classroom videos, it is necessary for humans to first annotate a set of videos in a consistent way (i.e., identify and label activities in videos with moderate to high levels of reliability). This paper describes our configuration of an annotation tool and our use of the tool to label instructional activities in videos of mathematics and reading instruction at the elementary level (i.e., children 5–12 years of age). The paper also reports on the degree of inter-annotator reliability regarding (a) identification of activities and (b) duration of annotated activities; (a) and (b) are prominent challenges for temporal analysis (Chen et al., 2018; Knight et al., 2017; Molenaar & Wise, 2022). Most temporal analysis in the field of learning analytics has been based on computer-captured metrics (Epp et al., 2017; Riel et al., 2018). Unlike these studies, this study featured a different type of data: human annotations of videos of elementary instruction. Validation efforts are crucial for research using temporal analytics generated by human annotations of videos.

In this study, we found that individuals without experience analyzing classroom videos were able to annotate the activity labels in videos of elementary instruction in a manner consistent with individuals with such experience. In particular, we reported raw agreement scores of about 80–90% or greater among all our annotators. For some low-inference activity labels, pairs of annotators were able to reach substantial agreements that were statistically significant. The validation study reported here represents an important step towards presenting methods to validate the classification of time- and event-based activities in classroom videos for the purpose of temporal analysis research for the learning analytics community.

2. The Role of Artificial Intelligence in Classifying Instructional Activities in Classroom Videos

During the past ten years, classroom videos have become a prominent feature of efforts to support, assess, and conduct research on classroom interactions. Using video in these ways has many advantages. It allows for a teacher's lesson(s) to be observed by multiple individuals and for more than one observation instrument to be employed in analyzing a given lesson. For teacher candidates whose clinical placements are distant from their universities and practising teachers in geographically isolated settings, using video can enable teacher educators to provide regular feedback on their teaching. In terms of teacher assessment, video is a key part of the initial licensure process in many states, through edTPA (SCALE, 2015); and it is integral to the National Board for Professional Teaching Standards' (NBPTS) process for awarding teachers advanced certification (Cowan & Goldhaber, 2016). In addition, using video has made it much more feasible to conduct research on teachers' instructional practices at scale. This is evident in the Measures of Effective Teaching (MET) study, which collected video-recorded data on the instructional practices of more than 1,000 4th- through 8th-grade teachers in six large U.S. school districts (Kane et al., 2013); as well as research featuring classroom observation instruments such as the Classroom Assessment Scoring System (CLASS; Hamre et al., 2012).

Using classroom videos has great potential to support teacher development as well as the processes of licensing teachers and conducting research on teacher quality. But there are notable costs associated with training raters and ensuring that they rate videos with adequate levels of reliability. These include financial costs associated with purchasing video equipment and storing videos in secure online locations; costs associated with identifying qualified raters, training and compensating them, and assessing the reliability of their rating efforts; and costs associated with the actual process of scoring itself. For many teacher educators, researchers, and policy-makers, these costs may lead them to conclude that using classroom videos is too expensive.

In response to these challenges, growing numbers of researchers are exploring the potential role of artificial intelligence in classifying instructional practices and student engagement in recordings of instructional activities. For example, Kelly et al. (2018) used automatic speech recognition, natural language processing, and machine learning to train computers to identify authentic questioning in secondary English language arts (ELA) classrooms based on audio recordings of classroom

interaction. Drawing on 583 transcripts and recordings from 139 classrooms, the researchers reported correlations between human-coded and computer-coded questions of $r = .602$ for transcripts and $r = .687$ for recordings (Kelly et al., 2018). Whitehill et al. (2014) drew on data from 34 undergraduates to compare human-detected versus computer-detected recognition of student engagement; the researchers used machine learning to develop automatic engagement detectors. They found that for distinguishing between high and low engagement, the automatic engagement detectors performed at a comparable level as humans. Jacoby et al. (2018) analyzed 43 videos of secondary school students engaged in an activity that involved talking, writing, and typing. Employing deep neural networks, Jacoby et al. (2018) reported classification accuracy of 99.7% (talking), 92.5% (writing), and 82.5% (typing).

In summary, researchers have provided evidence that it is possible to use automated speech recognition, natural language processing, machine learning, and even deep neural networks to classify instructional activities in audio and video recordings of classrooms from the early childhood and elementary levels to the secondary and undergraduate levels. At the same time, it is important for human annotators to first be able to label activities in videos with moderate to high levels of reliability to then train these forms of artificial intelligence to accurately classify such activities (Shaffer, 2017).

2.1 Temporal Analysis in Learning Analytics Research

Learning analytics researchers have identified several methodological choices and questions relevant to conducting temporal analysis in general and analyzing activities in videos of classroom instruction in particular (Chen et al., 2018; Knight et al., 2017; Molenaar & Wise, 2022). For example, Knight et al. (2017) made an important distinction between *time windows* and *analytic time units.* A time window refers to the entire period during which instructional activities are analyzed. Regarding elementary school teaching, this can refer to an entire reading or mathematics lesson or to segments of time within a lesson. In contrast, an analytic time unit is defined as "a way to aggregate up from the maximum granularity at which the data is collected" (Knight et al., 2017, p. 9). For example, researchers might select the analytic time unit of 30 seconds and iteratively label each 30-second unit from the video. Time windows and analytic time units are two examples of the analytic decisions researchers make when conducting temporal analysis.

Molenaar and Wise (2022) explicate key aspects of the construct passage of time. These include *position*, or when an activity takes place within a time window; *duration*, or the length of the activity; *frequency*, of how often the activity occurs; and *rate*, or the pace at which the activity takes place over time. For example, in an elementary mathematics lesson, a researcher might investigate the teacher's feedback in response to student explanations of how they solved mathematics problems. This could include exploring when the teacher shares feedback, the duration of the feedback, how often they provide it, and the rate at which it occurs during the lesson. Chen et al. (2018) articulated several key questions for researchers when conducting temporal analysis. These questions include considering how teaching activities are conceptualized in terms of time, where they are observed in time, and what insights into teaching and learning can be provided by the analytics. In terms of conceptualizing instructional activities with respect to time, one may choose to focus on the *passage of time* (e.g., when and how long an activity occurs) or the *sequence of different activities* (e.g., which activities come before/after other activities).

In terms of these possible choices for temporal analysis, Chen et al. (2018) "include concerns regarding evaluation of the analytics (including the preceding questions) and their validity" (p. 7). In the learning analytics field at large, researchers have been working on issues of evaluating analytics and their validity, especially as more analytics are being used for machine learning (Eagan et al., 2020; Kitto et al., 2023). However, there has been little attention to validity for temporal analytics in the learning analytics community. Further, the validation of temporal analytics is a unique problem; typical methods of inter-rater reliability from content analysis are problematic for timed-event data (Bakeman & Quera, 2011; 2023), which we will later discuss in more depth. In this study, we draw on methods from observational behaviour, ethnography, and video research to analyze the consistency of temporal analytics generated by human annotations of videos of elementary instruction. By using these methods, we were able to substantiate claims about whether annotators without experience in classroom video analysis were able to annotate videos of elementary instruction in a manner consistent with individuals with such experience.

2.2 Prior Research on the Reliability of Human Annotation in Videos of Instruction

Several researchers have examined the reliability of human annotation in videos of instruction. For example, Curby et al. (2016) examined whether live observation ratings of lessons in early childhood classrooms were consistent with ratings from video recordings of the same lessons. The researchers used the Classroom Assessment Scoring System-PreK (CLASS-PreK; Hamre et al., 2012) to analyze instructional support, emotional support, and classroom management in each class. The authors reported evidence that scorers could use CLASS-PreK to score both live and video-recorded lessons reliably, but that the scores for the live ratings were generally a little higher than those for the video ratings of the same lessons. In another study, Tucker et al. (2016) used video recordings from university-level physics lessons to compare annotations using video data only with those using both video and audio data. The annotators were asked to assign labels for the following constructs: worksheet, discussion, teaching assistant, joking, and other. For these constructs, the authors found that inter-rater reliability was as high when annotators used video data only as when they employed audio and video data.

In the studies by Curby et al. (2016) and Tucker et al. (2016), the researchers made the decision to compare annotators' labelling of the same analytic time units. In these cases, the studies featured 5-second and 20-minute time units, respectively. A benefit of this approach is that all the annotators will make the same number of timed-event decisions and researchers can compare these decisions by each analytic time unit. Further, annotators do not have to make decisions regarding position or duration for an activity within the video. The absence of these decisions for annotators contrasts with our study, which takes a time window approach.

In a related study, Prusak et al. (2010) investigated the ability of pre-service physical education teachers to annotate videos of physical education instruction using a time window approach. Pre-service teachers apply the following labels to segments of the video: instruction, student transitions, equipment transitions, freeze position, and discipline plan. The pre-service teacher's annotated performance was compared by two experts (i.e., university instructors). The authors reported that after limited training, the study participants were accurate in their content labelling and moderately reliable in their labelling ability. To support their claims, the authors reported two percentages: percent agreement of labels and percent of time overlap for labelled instance. Reporting these percentages alone, however, did not account for whether these agreements occurred by chance; it is recommended, in cases like this one, to use an event-matching algorithm and report an event-based kappa (Bakeman & Quera, 2023; Holle & Rein, 2015). Others also recommend reporting positive agreement (Feinstein & Cicchetti, 1990) and Shaffer's (2017) rho statistics with the event-based kappa.

In the study presented here, we build on previous studies by taking a time window approach to compare annotator performance for labelling classroom videos with activity structures and teaching and learning activities. In addition, this study is unique compared to prior temporal analysis in learning analytics research by applying an event-matching algorithm from observational behaviour research to make claims about the reliability of temporal analytics generated by human annotators with varying levels of experience analyzing classroom videos.

2.3 Using Multiple Statistics in Studies of Inter-Annotator Reliability

Researchers have argued that multiple statistics are needed to assess inter-annotator reliability among humans who are trained to label videos and/or text of human emotion and activity (e.g., Kitto et al., 2023; Tong et al., 2020). One statistic widely used in studies of inter-annotator reliability is percent raw agreement, which simply reports the total number of agreed-upon annotation labels in comparison to the total number of annotations. Researchers caution though against using percent raw agreement as the only measure of inter-annotator agreement since it does not consider agreements that may occur by chance. For example, Kitto et al. (2023) reported on a study of how four annotators applied labels in 2019 and again in 2022 to turns of talk that featured exploratory talk during a session at a professional conference. The authors used percent raw agreement, Cohen's kappa, Fleiss' kappa, and Krippendorff's alpha statistics to report on levels of inter-annotator agreement. Kitto et al. showed that high raw agreement scores did not necessarily correspond to high levels of agreement when the other statistics were used. While the annotators were provided with clearer directions when they engaged in content analysis the second time (i.e., in 2022), the authors reported only minor improvements in agreement using the kappa and alpha statistics between 2019 and 2022.

Other researchers warn against relying exclusively on kappa-based statistics. For example, Tong et al. (2020) analyzed the use of coefficient indices for inter-annotator reliability of classroom observation instruments used to assess implementation fidelity regarding educational programs for English learners. The authors describe a phenomenon in which Fleiss' kappa values were falsely conservative in instances when annotators "demonstrate a high percentage of agreement not due to chance" (Tong et al., 2020, p. 4). This phenomenon occurred when there was high agreement between annotators while some dimensions of the observation instruments were selected infrequently ($n \leq 5$). For such cases, the authors recommended reporting both non-chance-corrected (i.e., raw agreement) and chance-corrected (i.e., kappa) statistics. Similarly, others have recommended reporting positive and negative agreements with kappa-based statistics (Feinstein & Cicchetti, 1990).

Researchers have noted other considerations regarding the use of kappa statistics to assess inter-annotator agreement. For example, D'Mello (2016) examined the degree to which humans assigned the same ratings to short videos of adults expressing a range of emotions. In this study, eight undergraduates served as annotators and each of them participated in three scoring sessions. During each session, they worked in pairs to view nine five-minute video segments of nine different adults, independently annotate emotions in the video data, discuss their annotations in pairs, and subsequently annotate additional data. D'Mello reported that the average inter-annotator agreement measured as a Cohen's kappa statistic increased from 0.23 after the first iteration to 0.36 after the third iteration and then levelled off between the third and the ninth iterations (with inter-annotator agreement of 0.38 after the ninth iteration). The authors acknowledged that their results were lower than the recommended level of substantial agreement (i.e., 0.6 kappa), but noted that "lower kappa scores are to be expected when annotating affect since it is a latent state that is ill-defined and possibly indeterminate" (D'Mello, 2016, p. 144).

The studies reviewed thus far in this section have been situated in contexts where the raters or annotators are labelling the same unit of analysis, and researchers can compare the agreements between them easily through matching. For example, some of these studies involved labelling or rating a video by 10-second segments or turns of talk as the predefined unit of

analysis. When annotators must also make decisions about the position and duration of an activity within the video, they are making two decisions. First, they must decide the onset and offset times of the activity; second, they must decide which label to apply to the activity. For example, one rater may label an activity as one long segment with an onset and offset time while a second rater may label that same interval with that activity but decide to split it into two shorter segments. Therefore, inter-rater agreement in this context needs to consider both the categorization agreement for the activity labels between annotators and the segmentation agreement (Bakeman & Quera, 2011; Holle & Rein, 2015).

This inter-rater agreement problem is unique for *timed-event sequential data* (Bakeman & Quera, 2011). There are a few event-matching algorithms (e.g., Bakeman & Quera, 2023; Holle & Rein, 2015) that allow researchers to calculate categorization and segmentation agreements separately or derive an overall agreement that jointly considers categorization and segmentation. In this study, we used an event-matching algorithm (i.e., EasyDIAg) developed to provide a modified kappa that jointly considers categorization and segmentation (Holle & Rein, 2015). A strength of EasyDIAg in comparison to other event-matching algorithms is that it only requires one user-specified parameter — percentage of required temporal overlap — to be considered a match. Next, we review how sample size and segmentation may impact the event-matching algorithm used in this study.

Sample size can have a direct influence on inter-annotator reliability statistics for timed-event sequential data. Holle and Rein (2013) compared two research teams' applications of observation labels using EasyDIAg. For the first team, they reported that a sample size of 15 videos with 333 labels yielded large 95% confidence intervals around kappa scores. For this sample size, the authors noted that "conclusions about the tool's reliability" would be difficult to obtain (Holle & Rein, 2013, p. 6). The second team had a sample size of 44 videos and their 95% confidence intervals around the kappa scores were significantly narrower. Based on this analysis, the authors recommended that each category for labelling has at least 200 occurrences in the data. Alternatively, one can calculate Shaffer's rho (ρ) statistics to determine the generalizability of the kappa score (Eagan et al., 2020; Shaffer, 2017). For instance, if $\rho < 0.05$, then one may conclude kappa is over a predetermined threshold (e.g. > 0.65) with a maximum error rate of less than 5% for their entire dataset.

Holle and Rein (2015) also noticed that their event-matching algorithm EasyDIAg struggled to match in certain cases. One case of note was when one annotator creates one long segment, but the other annotates the same interval as multiple sort sequences. Often, EasyDIAg will fail to match these annotations on this interval because the overlap criterion tends not to be fulfilled. However, if EasyDIAg is used during the training of annotators, then these differences in grain size can be detected and can be specified in training materials to help increase the likelihood of matching with the algorithm.

2.4 Experience With Analyzing Activities in Classroom Videos

Another question that researchers have addressed is whether individuals who lack experience analyzing and interpreting images or activities within videos can classify phenomena in video data in a manner similar to those with such experience. In the area of medical imaging, researchers have shown that those without experience analyzing medical data in videos can apply labels in a comparable manner as those with more experience. Budd et al. (2021) compared individuals with no experience annotating U.S. medical data with a cardiologist and three sonographers; they reported that those lacking experience analyzing medical data were able to conduct complex medical image segmentation activities at the same high level as the more experienced analysts in their study. Kwitt et al. (2014) trained eight individuals without experience analyzing medical data to assess the presence of celiac disease in medical images; they concluded that their "large corpus of non-expert labelled (i.e., noisy) training data can in fact be used to build a classification system that performs equally well as a system trained solely on a limited number of pristine labels" (p. 460).

Scholars have documented similar results in other areas of video analysis. Jones et al. (2018) examined the ability of ten individuals who lacked experience analyzing video to label instances where a person with a cane took steps with the cane touching or not touching the ground. They showed that these individuals were able to label events in the cane dataset with high levels of agreement. Pustu-Iren et al. (2019) investigated the ability of five individuals without experience analyzing historical material and five with such experience to recognize concepts and historical figures in video; those lacking experience were able to annotate concepts at the same level of performance as their counterparts, but they had difficulty correctly identifying historical figures. Finally, in terms of natural language processing, Snow et al. (2008) compared individuals without experience analyzing text to those who had such experience regarding five tasks: affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation. The authors provided evidence of high levels of agreement between both groups for each of the five tasks.

In summary, annotation research in medical imaging, video analysis, and natural language processing has shown that individuals without experience analyzing video or text can classify some phenomena in various types of datasets at comparable levels of agreement as those with such experience. In this study, we build on this work by examining whether individuals without experience analyzing classroom video could classify activities in videos of elementary mathematics and reading instruction in a manner similar to those with such experience.

3. Research Questions

As a first step towards developing a dataset to train artificial intelligence to classify activities within classroom videos, the validation study reported here focused on assessing the agreement of human annotators in classifying a set of activities in such videos specifically in the absence of audio information. This study addressed the following research questions:

1. Can individuals without experience analyzing videos of elementary mathematics and language arts instruction annotate activities and their duration at the same level of agreement as those with such experience?
2. How reliably do these individuals label activities and their duration in classroom video?

4. Methods

4.1 Annotators

The six annotators included a postdoctoral researcher (A1) and two PhD students in Curriculum and Instruction (A2, A3) and three undergraduates in Youth Development and Policy at a research university in the Mid-Atlantic U.S. The first three annotators had taught full-time for two to five years; one taught high school mathematics for two years in South Carolina, one taught elementary ELA and mathematics for five years in Washington state, and one taught secondary ELA and science for five years in Namibia and Ethiopia. All three were white; two were male and one was female. All three had prior experience analyzing video recordings of K–12 instruction and providing feedback to teaching candidates in the context of university courses on teaching methods and classroom management. The three undergraduate students had prior experience tutoring youth, but none of them had experience analyzing videos of elementary instruction. One was Asian-American and two were white; all three were female.

4.2 Sample of Classroom Videos

This validation study drew on a randomly selected sample of videos of elementary instruction from a prior research study known as the Development of Ambitious Instruction study (DAI; Youngs et al., 2022). The DAI study focused on 83 beginning elementary teachers who graduated from teacher preparation programs at five universities across three states in the United States. The study participants completed their final year of preparation in either 2015–16 or 2016–17 and subsequently began teaching full-time in grades K–5 in general education settings. Each teacher was observed teaching mathematics and ELA up to six times in each subject during their first two years of full-time teaching; each video-recorded lesson was about 45 minutes to an hour in length.

For this validation study, the six annotators first used 24 activity labels to annotate a total of 50 hours of video (i.e., approximately 25 hours of mathematics instruction and 25 hours of ELA instruction); the number of hours of video annotated by each annotator ranged from 3 to 10. Next, the annotators each used the 24 activity labels to label four 15-minute "validation" segments (two in mathematics and two in ELA) that had previously been labelled by a primary annotator. The 15-minute validation segments were intentionally chosen to include a notable shift in instructional activities; for example, each segment included a shift from whole group instruction to small group activity or individual activity. The annotators' labels for these four 15-minute segments were used as the data source for this study. Each annotator labelled four 15-minute segments (i.e., one hour) of video because this represented 10% or more of the amount of video.

4.3 Annotation Tool

The annotators used ELAN (2021) computer software to annotate the 15-minute segments for the two mathematics videos and two ELA videos without listening to audio.[1] Figure 1 provides an example of the ELAN tool configured for annotating in this study. The video player is in the upper left while a timeline with classroom-based activity labels is annotated in the lower half of the screen. To label activities, the annotators selected onset and offset times and annotated the selected duration as "y" or yes to indicate the presence of that activity. The absence of an annotated time duration indicates that the activity was not present. Sound was disabled during all annotation using a mute control function within ELAN.

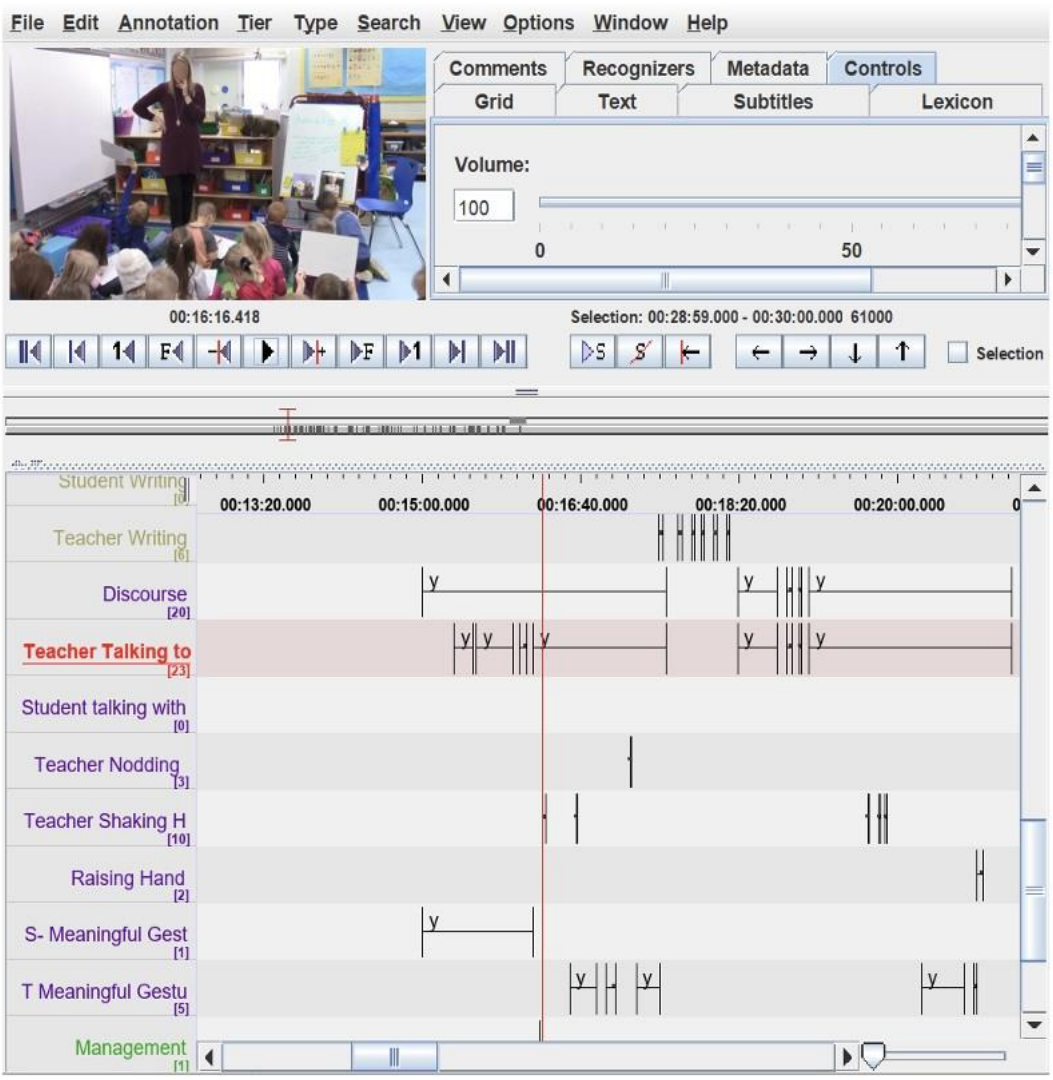

Figure 1. ELAN tool example.

4.4 Pilot of Classroom-based Activity Labels

Annotators used a set of 24 activity labels when annotating the 50 hours of video and the four 15-minute validation segments. The activity labels are based on two classroom observation instruments widely used in instructional observation research: the Mathematics-Scan (M-Scan; Berry et al., 2013) and the Protocol for Language Arts Observations (PLATO; Grossman et al., 2013). Using domains present in the M-Scan and PLATO rubrics, the research team developed a pilot set of labels focused on the following six parent areas: activity types (e.g., whole class, small group, individual); classroom discourse (e.g., student talking to student, nodding and shaking heads, meaningful gesture); teacher location (e.g., movement and position); representations of content (e.g., books, mathematics manipulatives, technologies, worksheets, writing); student location (e.g., movement and position); and management (e.g., signalling, movement of students, and other time management). Some of these pilot labels were low inference, such as whole group activities and using or holding books, while others were high inference, such as meaningful gesture. To annotate high-inference labels, the annotators made judgments and inferences featuring complex psychological constructs that inevitably led to a certain level of disagreement, as has been found in other studies (e.g., D'Mello, 2016).

Following an iterative labelling process, the annotators arrived at a finalized list of 24 labels prior to applying them to the 50 hours of video recordings and the four 15-minute validation segments in this study. This list of labels is in Table 1 with the text of parent labels in bold and placed in a highlighted cell and then their child labels listed below. Child label segmentations are either the same length as the parent label or as multiple smaller segmentations within the longer segment

of the parent label. For example, discourse (as the parent label) may occur for three minutes and students may raise their hand for 20 seconds during this three-minute period.

Table 1. List of Annotation Labels

Index	Annotation Parent Labels	Annotation Child Labels
1	Activity Types	Whole Class Activity
2		Small Group Activity
3		Individual Activity
4		Transition
5	Teacher Supporting	One Student
6		Multiple students with SS interaction
7		Multiple students without SS interaction
8	Teacher Location	Sitting (T)
9		Standing (T)
10		Walking (T)
11	Student Location	Sitting on the carpet or floor
12		Sitting at group tables
13		Sitting at desk
14		Student(s) walking or standing
15	Discourse	On task student talking with student
16		Student raising hand
17	Representing Content	Book — Using or Holding Book
18		Worksheet — Using or Holding
19		Notebook — Using or Holding
20		Instructional Tool — Using or Holding
21		Presentation with Technology
22		Individual technology — Using or Holding
23		Student Writing
24		Teacher Writing

The final set of activity labels used in this study was derived from the broader list of activities addressed in the M-Scan and PLATO observation rubrics; this study focused on a more limited set of labels for two reasons. First, for this validation study the annotators used videos without audio; thus, in early piloting work, annotators were unable to accurately annotate some high-inference instructional activities addressed by these rubrics such as behaviour management, meaning of interaction, student engagement, or differentiation of instruction. Second, this study is part of an initial effort to annotate classroom videos for training deep neural networks to recognize a limited set of activities; our long-term goal is to use both audio and video features from videos to train neural networks to recognize most, if not all, activities addressed in the M-Scan and PLATO.

4.5 Training of Annotators

Training of the annotators was based on procedures recommended for raters of performance assessments (Lane & Stone, 2006). Specifically, training focused on consistently recognizing activities present in the classroom videos and annotating the presence or absence of when each activity began and ended within the time frame for each video. Thus, annotators focused on position and duration, two notable features in temporal analysis, within the passage of time construct (Molenaar & Wise, 2022). Also, annotators only viewed video data of classroom instruction without listening to audio data. The pilot set of activity labels was created by the research team and shared with the annotators. The annotators, along with other members of the research team, discussed the purpose of the annotations, further defined the activities, and refined the process used for annotating a pilot set of classroom videos using the ELAN tool.

Two lead annotators, the PhD students, trained the other four annotators new to the project over four sessions that lasted between one-and-a-half to two hours each. The annotator training spanned five weeks, with assignments of practice annotations between each session. The practice annotations were three to four 15-minute video segments from the larger DAI dataset. Following each training session and practise annotation assignment, the lead annotators calculated the raw agreement between lead annotator-scored segments and the annotations of the trainees. Training was completed when all the annotators had raw agreement scores over 70% agreement for each activity label. Once they reached this threshold, annotators were assigned to complete individual annotation work over the course of two months. One annotator required an additional discussion regarding the teacher writing label; afterwards, she was assigned independent annotation work as well.

After the team annotated a total of 50 hours of video, this validation study was conducted. This study relied on three lead annotators due to their experience analyzing videos of K–12 instruction: the two PhD students and the postdoctoral researcher. Each PhD student was the lead annotator for one 15-minute validation segment and the postdoctoral researcher was the lead annotator for two 15-minute segments. All six annotators annotated four 15-minute segments of video, and their annotations were compared to those of the lead annotators. A small percentage of videos (1 hour out of 50 hours, split over four 15-minute segments) was considered sufficient because these annotations were being used to train neural networks and not being used to make evaluative decisions about individual teachers or classrooms (Lane & Stone, 2006).

4.6 Analytic Strategies

Annotations were treated as categorical (dichotomous) data. Agreement between two annotators was the extent to which both identified elementary instructional activities. Therefore, to measure the degree of agreement between the lead annotator and the other annotators, percent agreement (also called raw agreement) statistics and modified Cohen's (1960) kappa (κ) statistics (see Holle & Rein, 2013, 2015) were derived. We denote κ^* to distinguish the kappa-based statistics we used in this study from Cohen's κ statistics. Percent agreement was calculated as the total number of cases where both annotators agree divided by the total number of cases considered. On one hand, κ is generally thought to be a more robust measure than simple percent agreement calculation as it considers the possibility of the agreement between annotators occurring by chance. It can vary between –1 to 1 with the strength of the agreement interpreted based on the ranges presented in Table 2.

Table 2. Kappa (κ) Score Interpretation (Cohen, 1960)

Kappa (κ) Score	Implication
0	Agreement equivalent to chance
0.01–0.20	None to slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Near-perfect agreement
1.00	Perfect agreement

On the other hand, as noted in our literature review section, kappa-based statistics can be falsely low because they involve the use of a stringent correction to control for the possibility of chance agreement (Feinstein & Cicchetti, 1990). They are also less useful in making inferences about inter-annotator reliability when small samples of label instances are used (Shaffer, 2017). Therefore, we also computed Shaffer's rho (ρ) statistics to determine the percentage of samples in the empirical distribution of κ^* that are less than or equal to 0.65; that is, the likelihood of making a Type I error for assuming $\kappa^* > 0.65$ for the entire dataset. We set the expected base rate for activity labels to 0.01 and used the Calculate Rho online application (https://app.calcrho.org/). Thus, we report raw agreement, κ^*, and Shaffer's ρ statistics in the findings section.

A challenge in determining inter-annotator agreement for timed-event sequential data is to develop clear objective criteria to determine whether two annotators' judgments relate to the same event. This has been termed the "linking" problem (Holle & Rein, 2015), as depicted in Figure 2. There are at least three different ways in which the events could be linked. First, the short unit seen by Annotator 2 could be linked with the long unit of Annotator 1, based on their very similar onset time, whereas the second unit of Annotator 2 would remain unlinked. A second possibility might be to link the long unit of Annotator 2 with the unit seen by Annotator 1 (based on their substantial overlap), leaving the first, shorter unit of Annotator 2 unlinked. Finally, one could allow multiple linking, such that a unit from one annotator could be linked to multiple units from a second annotator. Figure 3 depicts an actual example from our dataset showcasing how two different annotators assigned the "Discourse" label across time. The inconsistent temporal breaks in the label adversely affect the κ^* statistic as well as the raw agreement.

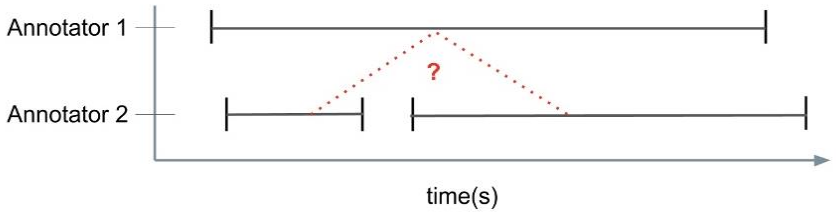

Figure 2. Illustration of the linking problem (Holle & Rein, 2015). Two annotators have independently identified onset (start) and offset (end) of an event of interest.

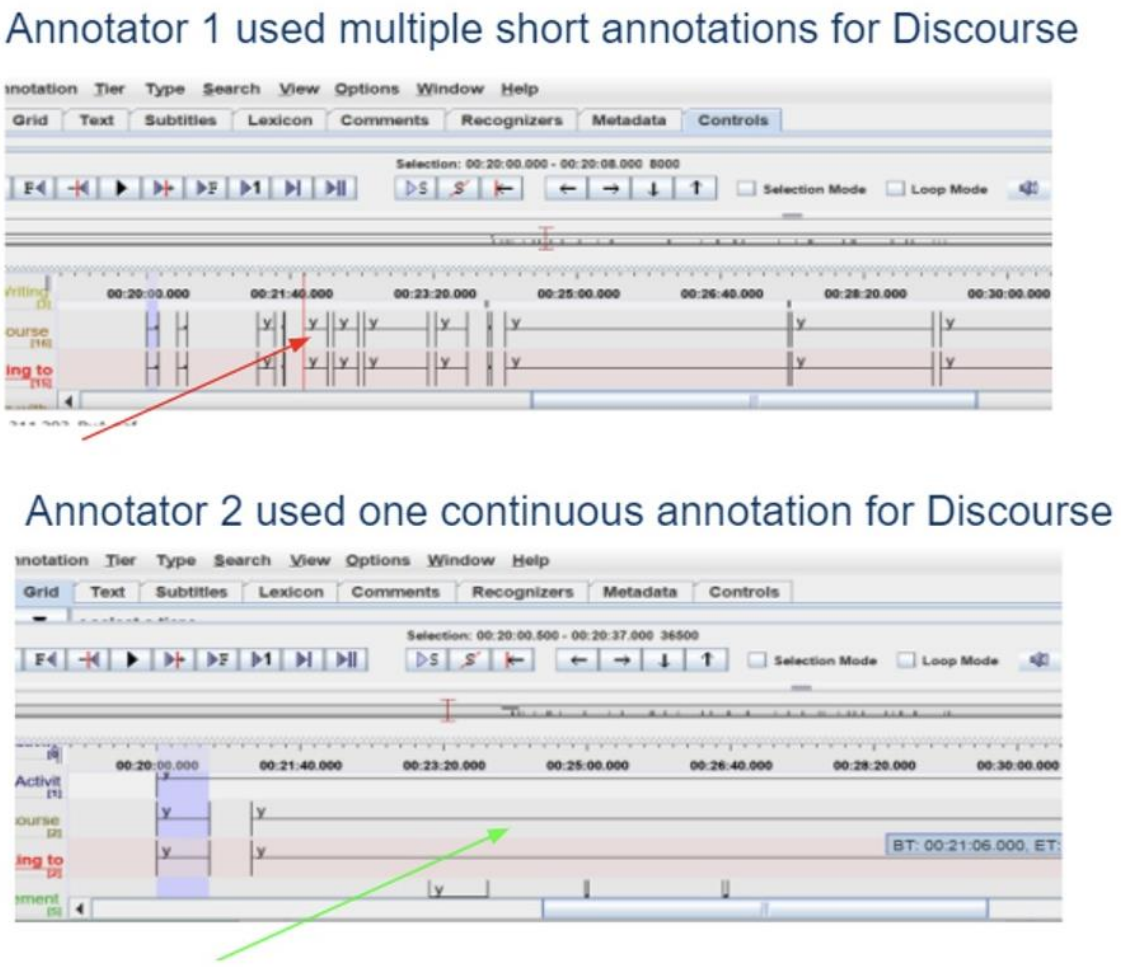

Figure 3. An example of the "underlapping" issue that arose from the data and created problems with determining the statistical overlap and inter-rater agreement versus the real-world agreement of the label.

To overcome the "linking" problem, while also providing chance corrected estimates (raw agreement statistics and kappa statistics) for inter-annotator agreement, this study utilized a free and open-source toolbox called EasyDIAg, developed by Holle and Rein (2015). This toolbox was implemented in MATLAB for analysis. EasyDIAg assumes that two annotators have independently annotated one or more video clips. Each segment of interest has a defined onset and offset time as well as an assigned category (i.e., activity label). The categories for our study are nominally scaled, mutually exclusive, and exhaustive.

EasyDIAg uses an event-matching algorithm to overcome the linking problem based on a user-defined threshold. For this study, the lowest threshold at 60% overlap was used. The aim of the matching algorithm is to generate an agreement table based on timed-event sequential rating data from two annotators. The resulting agreement table will have as many rows and columns as there are categories in the coding system, plus one additional column and row for commission/omission errors (Bakeman & Quera, 2011). Agreements are tallied on the main diagonal of the table, from top left to bottom right. Disagreements are tallied on the respective off-diagonal cells. Once the agreement matrix is populated, the EasyDIAg toolbox uses an iterative proportional fitting algorithm (Deming & Stephan, 1940) to obtain a modified Cohen κ statistic.

Both percent agreement statistics and kappa-based statistics have strengths and limitations. Raw agreement statistics are easily calculated and directly interpretable. Their key limitation is that they do not take account of the possibility that annotators guessed when applying labels or that agreement among annotators is by chance. Thus, they may overestimate true levels of agreement among annotators. Kappa-based statistics are designed to control for the possibility of guessing and random agreement, but the assumptions that they make about annotator independence and other factors are not well supported, and therefore they may excessively lower estimates of agreement among annotators. Further, they cannot be directly interpreted, and thus it has become common for researchers to accept low kappa values in studies of inter-annotator reliability (McHugh, 2012). Because both raw agreement statistics and kappa statistics provide meaningful information for clarifying annotation procedures (Holle & Rein, 2013), both statistics were included in this study.

4.7 Data Availability

The dataset generated and analyzed in this study are available from the corresponding author on reasonable request.

5. Findings

5.1 Agreements Among Annotators With Experience Analyzing Classroom Videos

Each of the three lead annotators had experience analyzing classroom videos. In this section, we report raw agreement statistics for these annotators using the four 15-minute video segments. The pairwise lead annotator raw agreements across all 24 activity labels ranged between 0.889 to 1.000. The average pairwise lead annotator raw agreements was 0.976. From these raw agreements, we can conclude that annotators with experience analyzing classroom videos had high agreement. However, we are unable to conclude yet if these high agreements were largely due to positive agreement, negative agreement, or both. Understanding the contributions of positive and negative agreements to the raw agreements will indicate whether annotators are agreeing mostly because they have identified the same activity occurring or agreeing mostly because they have identified the same activity as not occurring in video sequences. Next, we report the positive and negative agreements between the lead annotators.

Whereas the pairwise raw agreements between lead annotators for the 24 activities were high, there was quite a bit of variability for the pairwise positive agreements. Table 3 is a summary of the pairwise agreement statistics between lead annotators.[2] For some of the activities, the positive agreement ranges were high, such as teacher sitting (0.744–0.923) and raising hand (0.810–0.838), but for other activities, the positive agreement ranges were low, such as using or holding an instructional tool (0.000–0.254) and using or holding a book (0.311–0.390). For certain activities, the range of the pairwise positive agreements was broad such as in the case of teacher supporting multiple students without student interaction (0.000–0.643) and individual activity (0.000–0.750). The pairwise negative agreements for the 24 activities, in contrast, were high with little variability. The negative agreement values fell within the 0.939–1.000 range.

5.2 Agreements Between Those With and Without Experience Analyzing Classroom Videos

In this section, we report agreement statistics for the annotators without experience analyzing classroom videos in comparison to those annotators with experience. For each activity, we provide the range of the agreement statistics for those without experience in comparison to a lead annotator (see Table 4). Similar to the pairwise raw agreements for the annotators with experience, those without experience ranged from 0.857 to 1 across all 24 activity labels. We can conclude that annotators without experience had high agreements with those annotators with experience. Again, we turn to examining the contributions of positive and negative agreements to determine whether annotators without experience are agreeing with the lead annotators mostly because they have identified the same activity occurring or not occurring in video sequences.

The spread of the positive agreements of the annotators without experience to the lead annotators was highly variable. For eight of the activity labels, the positive agreements ranged between 0 and 1 (e.g., small group activity); this indicates no agreement to perfect agreement when the lead annotators identified an activity. For a few activities, the range of the positive agreements was much narrower such as raising hand (0.714–1.000) and student(s) standing or walking (0.533–0.881). In contrast, the negative agreements were high, and the spread of negative agreements was much narrower. The negative agreements between annotators with and without experience fell between 0.863 and 1.000 for all activity labels. There were three activity labels where the max negative agreement was less than 1, indicating that an annotator without experience labelled a video segment has having the activity present, but the lead annotator did not.

So far, we have examined the agreements among the annotators with experience and how the agreements compared between annotators without experience. Overall, the raw agreements across the 24 activity labels in these two cases were similar. For instance, in the case of whole class activity, the raw agreements between annotators with experience fell within the range of 0.949 to 0.967; comparably, the raw agreements for annotators without experience analyzing videos compared to those with experience was within the range of 0.938 to 1.000. Even though the agreements between annotators with and without experience seem comparable, we have yet to account for whether some of these agreements may be due to chance. In other words, we need to provide a measure for how reliable these annotators are performing, which we turn to next.

The spread of the positive agreements of the annotators without experience to the lead annotators was highly variable. For eight of the activity labels, the positive agreements ranged between 0 and 1 (e.g., small group activity); this indicates no agreement to perfect agreement when the lead annotators identified an activity. For a few activities, the range of positive agreements was much narrower, such as raising hand (0.714–1.000) and student(s) standing or walking (0.533–0.881). In contrast, the negative agreements were high, and the spread of negative agreements was much narrower. The negative agreements between annotators without experience and those with experience for all activity labels fell between 0.863 and 1.000. There were three activity labels where the max negative agreement was less than 1; this indicates that an annotator without experience labelled a video segment has having the activity present, but the lead annotator did not label it as such.

Table 3. Summary of the Pairwise Agreement Statistics Between Lead Annotators

Activity Labels	A1–A2 Agreements			A1–A3 Agreements			A2–A3 Agreements		
	Raw	Pos	Neg	Raw	Pos	Neg	Raw	Pos	Neg
Book-Using or Holding	0.987	0.000	0.994	0.980	0.225	0.990	0.987	0.600	0.993
Carpet or Floor-Sitting	0.989	0.000	0.994	0.990	0.536	0.995	0.993	0.750	0.997
Desk	0.987	0.753	0.993	0.997	0.958	0.999	0.982	0.686	0.991
Group Tables-Sitting	0.990	0.622	0.995	0.993	0.764	0.996	0.995	0.833	0.997
Individual Activity	0.997	0.000	0.998	0.997	0.750	0.999	0.992	0.708	0.996
Individual Technology	0.997	0.857	0.998	0.997	0.875	0.999	0.998	NA	0.999
Instructional Tool-Using or Holding	0.981	0.254	0.990	0.988	0.000	0.994	0.980	0.000	0.990
Multiple with SS Interaction	0.987	NA	0.994	0.974	0.210	0.987	0.972	0.233	0.986
Multiple without SS Interaction	0.987	0.000	0.994	0.996	0.000	0.998	0.992	0.643	0.996
Notebook	0.994	NA	0.997	0.994	NA	0.997	1.000	NA	1.000
On Task Student Talking with Student	0.992	0.000	0.996	0.993	0.000	0.996	0.992	0.450	0.996
One Student	0.981	0.440	0.990	0.986	0.450	0.993	0.985	0.361	0.992
Presentation with Technology	0.989	0.697	0.994	0.994	0.878	0.997	0.988	0.697	0.994
Raising Hand	0.976	0.830	0.987	0.978	0.838	0.988	0.973	0.810	0.986
Sitting	0.990	0.774	0.995	0.991	0.774	0.996	0.997	0.923	0.998
Small Group Activity	0.949	0.415	0.974	0.988	0.500	0.994	0.993	0.750	0.997
Standing (T)	0.919	0.526	0.956	0.916	0.577	0.954	0.963	0.794	0.980
Student Writing	0.951	0.133	0.975	0.929	0.459	0.962	0.942	0.450	0.969
Student(s) Standing or Walking	0.920	0.596	0.957	0.947	0.771	0.970	0.920	0.556	0.956
Teacher Writing	0.978	0.681	0.988	0.977	0.584	0.988	0.980	0.574	0.990
Transition	0.981	0.000	0.990	0.983	0.571	0.991	0.980	0.254	0.990
Walking	0.892	0.536	0.939	0.889	0.519	0.939	0.937	0.691	0.965
Whole Class Activity	0.989	0.633	0.994	0.997	0.900	0.999	0.988	0.556	0.994
Worksheet-Using or Holding	0.949	0.390	0.973	0.960	0.376	0.979	0.967	0.311	0.983

Note. NA indicates that either one annotator or both did not identify the activity within the 15 min. segments.

Table 4. Summary of the Pairwise Agreement Statistics Between Lead Annotators to Secondary Annotators

Activity Labels	Raw Agreement		Positive Agreement		Negative Agreement	
	Min	Max	Min	Max	Min	Max
Book-Using or Holding	0.935	1.000	0.000	0.000	0.988	1.000
Carpet or Floor-Sitting	0.969	1.000	0.333	1.000	0.975	1.000
Desk	0.962	1.000	0.333	1.000	0.969	1.000
Group Tables-Sitting	0.985	1.000	0.667	1.000	0.992	1.000
Individual Activity	0.985	1.000	0.000	0.833	0.985	1.000
Individual Technology	0.975	1.000	0.000	1.000	0.984	1.000
Instructional Tool-Using or Holding	0.944	1.000	0.000	0.000	0.944	1.000
Multiple with SS Interaction	0.950	0.992	0.000	1.000	0.962	1.000
Multiple without SS Interaction	0.976	1.000	0.000	0.750	0.977	1.000
Notebook	0.992	1.000	0.000	1.000	0.992	1.000
On Task Student Talking with Student	0.969	1.000	0.000	1.000	0.987	1.000
One Student	0.944	1.000	0.000	1.000	0.962	1.000
Presentation with Technology	0.938	1.000	0.400	1.000	0.942	1.000
Raising Hand	0.965	1.000	0.714	1.000	0.981	1.000
Sitting	0.989	1.000	0.667	1.000	0.989	1.000
Small Group Activity	0.985	1.000	0.000	1.000	0.985	1.000
Standing (T)	0.922	0.962	0.504	0.875	0.958	0.978
Student Writing	0.890	0.992	0.000	0.660	0.938	1.000
Student(s) Standing or Walking	0.920	0.978	0.533	0.881	0.928	0.979
Teacher Writing	0.929	1.000	0.588	1.000	0.939	1.000
Transition	0.977	1.000	0.400	1.000	0.977	1.000
Walking	0.857	0.954	0.222	0.754	0.863	0.976
Whole Class Activity	0.969	1.000	0.000	1.000	0.977	1.000
Worksheet-Using or Holding	0.938	1.000	0.000	1.000	0.960	1.000

5.3 Reliability of Annotators With Experience Analyzing Classroom Videos

Across the 24 activity labels, there was considerable variability in the reliability measures for the lead annotators (see Table 5). For the following activity labels, there was no to slight agreement among the annotators with experience: 1) teacher supporting multiple students with student interaction and 2) using or holding notebook. There were a few activities where the reliability between one pair of lead annotators was no to slight agreement and another pair was fair agreement, such as using or holding a book and using or holding an instructional tool. Across all three pairings of lead annotators, there was fair agreement when annotating for using or holding a worksheet. For the following six activity labels, the agreements between lead annotators ranged from no to slight agreement to moderate agreement: 1) sitting on carpet or floor, 2) teacher supporting multiple students without interaction, 3) on task student talking with student, 4) teacher supporting one student, 5) student writing, and 6) transition. All three pairings of the lead annotators achieved moderate agreement for teacher writing. In the case of 1) students sitting at group tables, 2) teacher standing, 3) student(s) standing or walking, and 4) teacher walking, the pairwise agreements between annotators with experience ranged from moderate to substantial agreement. Only the activity of raising hand achieved near-perfect agreement between all three pairs of lead annotators. Three activity labels ranged from substantial to near-perfect agreement between the lead annotators: 1) students sitting at desks, 2) presentation with technology, and 3) teacher sitting.

For some activities, there was quite a spread in the agreements between annotators with experience. There was no agreement to substantial agreement among annotators for students sitting on carpet or floor and individual activity. For individual technology, there was no agreement to near-perfect agreement for individual technology among lead annotators. The agreement for small group activity ranged from fair agreement to substantial agreement, whereas the agreement for whole class activity ranged from substantial to near-perfect agreement.

Table 5. Summary of the Pairwise Reliability Statistics Between Lead Annotators

Activity Labels	A1–A2			A1–A3			A2–A3		
	κ	Pos	Neg	κ	Pos	Neg	κ	Pos	Neg
Book-Using or Holding	0.00	0.000	0.994	0.21	0.225	0.990	0.33	0.600	0.993
Carpet or Floor-Sitting	0.00	0.000	0.994	0.53	0.536	0.995	0.66	0.750	0.997
Desk	0.74*	0.753	0.993	0.96*	0.958	0.999	0.68	0.686	0.991
Group Tables-Sitting	0.57	0.622	0.995	0.76*	0.764	0.996	0.80*	0.833	0.997
Individual Activity	0.00	0.000	0.998	0.67	0.750	0.999	0.70	0.708	0.996
Individual Technology	0.86*	0.857	0.998	0.86*	0.875	0.999	0.00	NA	0.999
Instructional Tool-Using or Holding	0.24	0.254	0.990	0.00	0.000	0.994	0.00	0.000	0.990
Multiple with SS Interaction	0.00	NA	0.994	0.17	0.210	0.987	0.18	0.233	0.986
Multiple without SS Interaction	0.00	0.000	0.994	0.00	0.000	0.998	0.44	0.643	0.996
Notebook	0.00	NA	0.997	0.00	NA	0.997	0.00	NA	1.000
On Task Student Talking with Student	0.00	0.000	0.996	0.00	0.000	0.996	0.44	0.450	0.996
One Student	0.39	0.440	0.990	0.44	0.450	0.993	0.30	0.361	0.992
Presentation with Technology	0.69	0.697	0.994	0.87*	0.878	0.997	0.69	0.697	0.994
Raising Hand	0.81*	0.830	0.987	0.82*	0.838	0.988	0.80*	0.810	0.986
Sitting	0.75*	0.774	0.995	0.76*	0.774	0.996	0.92*	0.923	0.998
Small Group Activity	0.36	0.415	0.974	0.49	0.500	0.994	0.66	0.750	0.997
Standing (T)	0.48	0.526	0.956	0.51	0.577	0.954	0.77*	0.794	0.980
Student Writing	0.09	0.133	0.975	0.41	0.459	0.962	0.41	0.450	0.969
Student(s) Standing or Walking	0.51	0.596	0.957	0.73*	0.771	0.970	0.51	0.556	0.956
Teacher Writing	0.66*	0.681	0.988	0.54	0.584	0.988	0.56	0.574	0.990
Transition	0.00	0.000	0.990	0.56	0.571	0.991	0.24	0.254	0.990
Walking	0.47	0.536	0.939	0.43	0.519	0.939	0.65	0.691	0.965
Whole Class Activity	0.63	0.633	0.994	0.89*	0.900	0.999	0.53	0.556	0.994
Worksheet-Using or Holding	0.36	0.390	0.973	0.28	0.376	0.979	0.27	0.311	0.983

Note. (*) indicates that the Shaffer's $\rho < 0.05$ and so we can conclude that for any video the kappa-based statistic between two lead annotators' labelling would most likely be ≥ 0.65.

ISSN 1929-7750 (online).

5.4 Reliability Between Those With and Without Experiencing Analyzing Classroom Videos

There was also considerable variability in the reliability of the annotators without experience in comparison to the lead annotators (see Table 6). No annotators without experience were able to reliably label the activities of using or holding a book and using or holding an instructional tool with any lead annotator. This is not too surprising given that even among the lead annotators with experience the agreements were fair at best and at worst no agreement.

Some of the annotators without experience were found to label some activities in a reliable manner compared to one — and often more than one — of the lead annotators. Those annotators without experience were found to have substantial to (near) perfect agreement ($0.60 < \kappa \leq 1.00$) with the lead annotators for the following activity labels: 1) students sitting at group tables, 2) raising hand, 3) teacher sitting, and 4) student(s) standing or walking. For these activity labels, some of these substantial agreements were statistically significant ($\rho < 0.05$). Also, those annotators without experience were found to have moderate to near-perfect agreement ($0.40 < \kappa \leq 1.00$) with lead annotators for students sitting on the carpet or floor, students sitting at desk, teacher standing, and transition; further, some of those with substantial agreements were also statistically significant ($\rho < 0.05$).

For the remaining activity labels, the reliability of the annotators without experience compared to lead annotators varied, with agreements ranging between no agreement to perfect agreement. The agreement outcomes for these labels usually fell into one of two outcomes. In terms of the first outcome, at least one of the lead annotators had substantial agreements with more than one of the annotators without experience (e.g., individual technology); however, these agreements were often not statistically significant or only significant for one of the annotators without experience. For the other outcome, two of the lead annotators reached substantial agreement with at most one of the annotators without experience — not necessarily the same annotator. In addition, this substantial agreement was often not statistically significant (e.g., teacher supporting one student).

Table 6. Summary of Secondary Annotators' Pairwise Agreement Statistics to Lead Annotators

Activity Labels	Min			Max			Number of Secondary		
	κ	Pos	Neg	κ	Pos	Neg	A1	A2	A3
Book-Using or Holding	0.00	0.000	0.988	0.00	0.000	0.988	0, 0	NA	0, 0
Carpet or Floor-Sitting	0.43	0.333	0.975	1.00	1.000	0.975	3, 2	2, 1	NA
Desk	0.43	0.333	0.969	1.00	1.000	0.969	2, 2	2, 0	3, 2
Group Tables-Sitting	0.66	0.667	0.992	1.00	1.000	0.992	3, 2	3, 1	3, 1
Individual Activity	0.00	0.000	0.985	0.80	0.833	0.985	2, 0	0, 0	NA
Individual Technology	0.00	0.000	0.984	1.00	1.000	0.984	1, 0	2, 0	3, 1
Instructional Tool-Using or Holding	0.00	0.000	0.944	0.00	0.000	0.944	0, 0	NA	0, 0
Multiple with SS Interaction	0.00	0.000	0.962	0.80	1.000	0.962	1, 0	0, 0	1, 0
Multiple without SS Interaction	0.00	0.000	0.977	0.66	0.750	0.977	1, 0	0, 0	0, 0
Notebook	0.00	0.000	0.992	1.00	1.000	0.992	1, 1	0, 0	0, 0
On Task Student Talking with Student	0.00	0.000	0.987	1.00	1.000	0.987	1, 0	1, 1	NA
One Student	0.00	0.000	0.962	0.80	1.000	0.962	1, 0	1, 0	NA
Presentation with Technology	0.00	0.400	0.942	1.00	1.000	0.942	1, 1	2, 0	3, 1
Raising Hand	0.69	0.714	0.981	1.00	1.000	0.981	3, 2	3, 3	3, 3
Sitting	0.79	0.667	0.989	1.00	1.000	0.989	3, 3	3, 3	3, 1
Small Group Activity	0.00	0.000	0.985	1.00	1.000	0.985	2, 0	0, 0	3, 2
Standing (T)	0.45	0.504	0.958	0.86	0.875	0.958	0, 0	2, 1	3, 3
Student Writing	0.00	0.000	0.938	0.61	0.660	0.938	0, 0	0, 0	NA
Student(s) Standing or Walking	0.62	0.533	0.928	0.90	0.881	0.928	3, 3	2, 1	2, 2
Teacher Writing	0.00	0.588	0.939	1.00	1.000	0.939	0, 0	3, 3	2, 1
Transition	0.56	0.400	0.977	1.00	1.000	0.977	3, 3	1, 0	3, 0
Walking	0.29	0.222	0.863	0.70	0.754	0.863	2, 1	0, 0	0, 0
Whole Class Activity	0.00	0.000	0.977	1.00	1.000	0.977	3, 3	2, 2	3, 1
Worksheet-Using or Holding	0.00	0.000	0.960	1.00	1.000	0.960	0, 0	NA	2, 0

Note. NA signifies that neither the lead annotator nor any of the secondary annotators identified instances of the activity in the corresponding row of the video annotation. For the A1, A2, and A3 columns, the left value indicates the number of secondary annotators with whom the lead annotator reached substantial agreement ($\kappa^* \geq 0.65$); the right value indicates the how many agreements with the secondary annotators were statistically significant ($\rho < 0.05$).

6. Discussion

The findings from this study confirm that labelling activities from classroom videos using visual recordings in the absence of audio is challenging, even for annotators with experience. Even though the raw agreements among the lead annotators were greater than 0.900 for almost all of the activity labels, the kappa* scores for some activity labels were low ($\kappa^* < 0.40$). This is a well-known paradox, as described by Feinstein and Cicchetti (1990) and observed in classroom observation research (e.g., Tong et al., 2020), between high raw agreement and low kappa values. Feinstein and Cicchetti (1990) recommended examining the positive and negative agreements to explain any discrepancies between high raw agreements and low kappa values. In Table 5, most of these discrepancies can be explained by low positive agreements among lead annotators. For instance, lead annotators A1 and A2 had a positive agreement of 0.133 and a kappa* value of 0.09 for the activity of student writing. Low positive agreements and kappas among the lead annotators suggest that further refinement of the activity definition may be needed. This can be seen in the case of the activity label called teacher supporting multiple students without student interaction. Lead annotator A1 had no positive agreements with A2 and A3, but there was moderate positive agreement between A2 and A3 for this activity. Therefore, it may be useful for A2 and A3 to further discuss with A1 how they are labelling this activity and refine the definition from that discussion.

When we examined whether the agreement among lead annotators for certain activity labels was statistically significant (i.e., $\rho < 0.05$), we found only 10 labels had at least one pair of lead annotators meeting this condition. These were typically low-inference activities such as students sitting at group tables or desks or students raising their hands. High-inference activities, such as on task student talking with student, were unlikely to reach substantial levels of agreement among the lead annotators. This outcome for low- and high-inference labels is consistent with the research literature (D'Mello, 2016). However, there were some low-inference activities that were worse off in agreement than we expected, such as student writing, teacher standing, and teacher walking. Refining these definitions for activities did not seem necessary. On further examination of the annotated data, we found that segmentations for these activities were at different grain sizes. Some lead annotators had longer durations of student writing whereas others had the same time window but had created multiple fine-grained onsets and offsets within that window. In the case of teacher walking and teacher standing, some annotators had turned off the teacher walking label if the teacher was walking around the room but then stopped for a brief period of time (< 5 seconds); others kept the teacher walking label turned on for the entire time. One annotator labelling one long segment and another labelling multiple, short sequences within the same interval is a known limitation of the linking algorithm in EasyDiag (Holle & Rein, 2015). Therefore, our selected algorithm may have led to lower agreements between some lead annotators for certain activities.

One important finding from this study was that the three undergraduate annotators, none of whom had experience analyzing videos of K–12 instruction prior to working on this research project, were able to annotate 16 of the 24 labels at a statistically significant high level of consistency with at least one of the lead annotators. This result is similar to those reported in studies comparing annotators with and without experience in medical imaging, video analysis, and natural language processing (Budd et al., 2021; Jones et al., 2018; Snow et al., 2008). Nearly all of these activities were low inference. This finding suggests that it is possible to train individuals to accurately annotate some low-inference labels of classroom activity. At the same time, as our study adds high-inference labels (such as the quality of teacher questioning, teacher scaffolding, and classroom discussion) in the future based on using both audio and video data, it is unclear whether individuals who lack experience analyzing instruction video will be able to accurately annotate more high-inference labels. This is an empirical question that should be addressed in future research.

This study offers some implications for validating time-based and event-based analytics to be used for temporal analysis. The annotators in the study analyzed instructional activities regarding two important aspects within the passage of time: position and duration (Molenaar & Wise, 2022). Typical methods of validating passage of time analytics regarding inter-rater reliability are problematic (Bakeman & Quera, 2011, 2023). To create time-based and event-based kappa statistics, we had to draw on a linking algorithm from behaviour research (Holle & Rein, 2015), which may have limitations in determining the agreements for certain activity labels. It will be important going forward to investigate which event-matching algorithms may be more appropriate for different temporal analytics and for researchers to offer a rationale for the algorithm they select for validation purposes. It seems critical to understand how the algorithm handles temporal overlapping and what are the thresholds for matching among annotations. This is useful for bolstering validity arguments, and it seems useful, especially during the early phases of label development and annotator training, for understanding the reported raw agreements, time- and event-based kappas, and positive and negative agreements to examine consistent performance.

Finally, based on recent trends in MLOps (Machine Learning DevOps), as discussed by Ng (in DeepLearningAI, 2021) and Jiang (2021), it is becoming evident that it is more important to have clean and reliable data compared to large amounts of data. Depending on the dataset, the training accuracy of a neural network will eventually stagnate after a fixed number of iterations and can even go down. Small, accurately labelled datasets can prove more meaningful for research compared to large-scale datasets with inconsistent labels. When the dataset is small, it is imperative to ensure that it is accurate to generate reliable artificial intelligence (AI) models. Consistency of labels present is one of the most important criteria for a dataset.

By ensuring that consistent labels are made available to the neural network in the training and testing phases, our research project aims to ensure the generation of a high-quality dataset that synergizes the impact of the neural networks for agnostic classroom observations that could assist in teacher preparation and evaluation in K–12 education in the United States. The label list in Table 1 will most definitely be refined based on results from this validation study and as neural networks start providing our research team with feedback on the accuracy of detection of these labels and how these labels may offer value to researchers as an agnostic tool to assist in pedagogical performance. However, the 50-hour dataset generated with these 24 labels provided a reasonable starting point for our neural network analysis (see Foster et al., 2024).

Endnotes

1. ELAN is a software tool designed to support manual annotation, transcription, visualization, linking, and searching of video and audio recordings. Using ELAN, annotations can be created on multiple layers, referred to as tiers; thus, ELAN supports multi-level, multi-annotator labelling of videos of instruction. ELAN has been used in several disciplines, including education, psychology, and medicine (Wittenburg et al., 2006); it has been applied to topics such as human–computer interaction, nonverbal communication, and gesture analysis (Giuliani et al., 2015; Kong et al., 2015). ELAN is available as free and open-source software under the GNU General Public License.
2. The list of labels in Tables 3–6 is the same as the list in Table 1 except for the fact that the labels are listed in alphabetical order in Tables 3–6 while in Table 1 they are grouped by category (e.g., activity types, representing content).

Declaration of Conflict of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This material is based upon work financially supported by the National Science Foundation under Grant No. 2000487 and the Robertson Foundation under Grant No. 9909875. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

References

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences.* Cambridge University Press. https://doi.org/10.1017/CBO9781139017343

Bakeman, R., & Quera, V. (2023). Behavioral observation. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics*, 2nd ed. (pp. 251–274). American Psychological Association. https://doi.org/10.1037/0000318-013

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

Budd, S., Day, T., Simpson, J., Lloyd, K., Matthew, J., Skelton, E., Razavi, R., & Kainz, B. (2021). Can non-specialists provide high quality gold standard labels in challenging modalities? In S. Albarqouni, M. J. Cardoso, Q. Dou, K. Kamnitsas, B. Khanal, I. Rekik, N. Rieke, D. Sheet, S. Tsaftaris, D. Xu, & Z. Xu (Eds.), *Domain adaptation and representation transfer, and affordable healthcare and AI for resource diverse global health* (pp. 251–262). Springer International Publishing. https://doi.org/10.1007/978-3-030-87722-4_23

Chen, B., Knight, S., & Wise, A. F. (2018). Critical issues in designing and implementing temporal analytics. *Journal of Learning Analytics, 5*(1), 1–9. https://doi.org/10.18608/jla.2018.53.1

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cowan, J., & Goldhaber, D. (2016). National Board Certification and teacher effectiveness: Evidence from Washington State. *Journal of Research on Educational Effectiveness*, *9*(3), 233–258. https://doi.org/10.1080/19345747.2015.1099768

Curby, T. W., Johnson, P., Mashburn, A. J., & Carlis, L. (2016). Live versus video observations: Comparing the reliability and validity of two methods of assessing classroom quality. *Journal of Psychoeducational Assessment, 34*(8), 765–781. https://doi.org/10.1177/073428291562711

DeepLearningAI. (2021, March 21). *A chat with Andrew on MLOps: From model-centric to data-centric AI* [Video]. YouTube. https://www.youtube.com/watch?v=06-AZXmwHjo

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*(4), 427–444. https://doi.org/10.1214/aoms/1177731829

D'Mello, S. K. (2016). On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing*, *7*(2), 136–149. https://doi.org/10.1109/TAFFC.2015.2457413

Eagan, B., Brohinsky, J., Wang, J., & Shaffer, D. W. (2020). Testing the reliability of inter-rater reliability. *Proceedings of the 10th International Conference on Learning Analytics and Knowledge* (LAK '20), 23–27 March 2020, Frankfurt, Germany (pp. 454–461). ACM Press. https://doi.org/10.1145/3375462.3375508

Max Planck Institute for Psycholinguistics, The Language Archive. (2021). *ELAN* (Version 6.2) [Computer software]. https://archive.mpi.nl/tla/elan

Epp, C. D., Phirangee, K., & Hewitt, J. (2017). Talk with me: Student pronoun use as an indicator of discourse health. *Journal of Learning Analytics*, *4*(3), 47–75. http://dx.doi.org/10.18608/jla.2017.43.4

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Foster, J. K., Korban, M., Youngs, P., Watson, G. S., & Acton, S. T. (2024). Classification of instructional activities in classroom videos using neural networks. In X. Zhai & J. Krajcik (Eds.), *Uses of artificial intelligence in STEM education*. Oxford University Press.

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

Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., & Tscheligi, M. (2015). Systematic analysis of video data from different human–robot interaction studies: A categorization of social signals during error situations. *Frontiers in Psychology*, *6*, 931. https://doi.org/10.3389/fpsyg.2015.00931

Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., Howes, C., LaParo, K., & Scott-Little, C. (2012). A course on effective teacher–child interactions: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal*, *49*(1), 88–123. https://doi.org/10.3102/0002831211434596

Holle, H., & Rein, R. (2013). The modified Cohen's kappa: Calculating interrater agreement for segmentation and annotation. In H. Lausberg (Ed.), *Understanding body movement: A guide to empirical research on nonverbal behaviour* (pp. 261–277). Peter Lang.

Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavioral Research Methods*. *47*(3), 837–847. https://doi.org/10.3758/s13428-014-0506-7

Jacoby, A. R., Pattichis, M. S., Celedón-Pattichis, S., & LópezLeiva, C. (2018). Context-sensitive human activity classification in collaborative learning environments. *Proceedings of the 2018 IEEE Southwest Symposium on Image Analysis and Interpretation* (SSIAI 2018), 8–10 April 2018, Las Vegas, NV, USA (pp. 141–144). IEEE. https://doi.org/10.1109/SSIAI.2018.8470331

Jiang, J. (2021, May 3). *What is MLOps and why we should care*. Medium. https://towardsdatascience.com/what-is-mlops-and-why-we-should-care-9b2d79a29e75

Jones, M. D., Johnson, N., Seppi, K., & Thatcher, L. (2018). Understanding how non-experts collect and annotate activity data. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (UbiComp '18), 8–12 October 2018, Singapore (pp. 1424–1433). https://doi.org/10.1145/3267305.3267507

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Bill & Melinda Gates Foundation. https://files.eric.ed.gov/fulltext/ED540959.pdf

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

Kitto, K., Manly, C. A., Ferguson, R., & Poquet, O. (2023). Towards more replicable content analysis for learning analytics. *Proceedings of the 13th International Conference on Learning Analytics and Knowledge* (LAK '23), 13–17 March 2023, Arlington, TX, USA (pp. 303–314). ACM Press. https://dl.acm.org/doi/10.1145/3576050.3576096

Knight, S., Wise, A. F., & Chen, B. (2017). Time for change: Why learning analytics needs temporal analysis. *Journal of Learning Analytics, 4*(3), 7–17. https://doi.org/10.18608/jla.2017.43.2

Kong, A. P.-H., Law, S.-P., Kwan, C. C.-Y., Lai, C., & Lam, V. (2015). A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a Database of Speech and Gesture (DoSaGE). *Journal of Nonverbal Behavior*, *39*(1), 93–111. https://doi.org/10.1007/s10919-014-0200-6

Kwitt, R., Hegenbart, S., Rasiwasia, N., Vécsei, A., & Uhl, A. (2014). Do we need annotation experts? A case study in celiac disease classification. In P. Golland, N. Hata, C. Barillot, J. Hornegger, & R. Howe (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014* (pp. 454–461). Springer International Publishing. https://doi.org/10.1007/978-3-319-10470-6_57

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement*, 4th ed. (pp. 387–431). Roman & Littlefield Publishers.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Molenaar, I., & Wise, A. F. (2022). Temporal aspects of learning analytics: Grounding analyses in concepts of time. In C. Lang, G. Siemens, A. F. Wise, D. Gašević, & A. Merceron (Eds.), *The handbook of learning analytics*, 2nd ed. (pp. 66–76). SoLAR. https://doi.org/10.18608/hla22.007

Prusak, K., Dye, B., Graham, C. R., & Graser, S. (2010). Reliability of pre-service physical education teachers' coding of teaching videos using studiocode analysis software. *Journal of Technology and Teacher Education, 18*(1), 131–159. http://hdl.lib.byu.edu/1877/2846

Pustu-Iren, K., Mühling, M., Korfhage, N., Bars, J., Bernhöft, S., Hörth, A., Freisleben, B., & Ewerth, R. (2019). Investigating correlations of inter-coder agreement and machine annotation performance for historical video data. In A. Doucet, A. Isaac, K. Golub, T. Aalberg, & A. Jatowt (Eds.), *Digital Libraries for Open Knowledge* (pp. 107–114). Springer International Publishing. https://doi.org/10.1007/978-3-030-30760-8_9

Riel, J., Lawless, K. A., & Brown, S. W. (2018). Timing matters: Approaches for measuring and visualizing behaviours of timing and spacing of work in self-paced online teacher professional development courses. *Journal of Learning Analytics*, *5*(1), 25–40. http://dx.doi.org/10.18608/jla.2018.51.3

SCALE. (2015). *Elementary education: Assessment handbook.* edTPA. https://www.colorado.edu/education/sites/default/files/attached-files/edtpaelehandbook.pdf

Shaffer, D. W. (2017). *Quantitative ethnography*. Cathcart Press.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (EMNLP '08), 25–27 October 2008, Honolulu, Hawaii (pp. 254–263). Association for Computational Linguistics. https://aclanthology.org/D08-1027

Tong, F., Tang, S., Irby, B. J., Lara-Alecio, R., & Guerrero, C. (2020). The determination of appropriate coefficient indices for inter-rater reliability: Using classroom observation instruments as fidelity measures in large-scale randomized research. *International Journal of Educational Research, 99,* 101514. https://doi.org/10.1016/j.ijer.2019.101514

Tucker, L., Scherr, R. E., Zickler, T., & Mazur, E. (2016). Exclusively visual analysis of classroom group interactions. *Physical Review Physics Education Research, 12*(2), 020142. https://doi.org/10.1103/PhysRevPhysEducRes.12.020142

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86–98. https://doi.org/10.1109/TAFFC.2014.2316163

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings of 5th International Conference on Language Resources and Evaluation* (LREC '06), 24–26 May 2006, Genoa, Italy (pp. 1556–1559). http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf

Youngs, P., Molloy Elreda, L., 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*, 103576. https://doi.org/10.1016/j.tate.2021.103576