Classroom Observation: Evaluating Instructional Support Automatically in Classroom for Young Children

Jiani Wang JWANG21@WPI.EDU
Worcester, United States
Kamil Hankour HANKOURK@VCU.EDU
Richmond, United States
Yuqi Zhang ZHANGY19@VCU.EDU
Richmond, United States
Jennifer LoCasale-Crouch LOCASALECRJ@VCU.EDU
Richmond, United States
Jacob Whitehill JRWHITEHILL@WPI.EDU
Worcester, United States

Abstract

To improve efficiency of evaluating classroom Instructional Support (IS) and enhance the reliability of the IS score evaluation system, we proposed a novel annotation protocol based on classroom discourse types and a framework which employed large language models (LLM) as the core component to estimate IS score automatically. We constructed the SentTag dataset which was annotated by the proposed annotation protocol. The Fleiss' Kappa among all annotators was 0.7120. Additionally Llama 3.1 models were fine-tuned on this dataset, achieving an accuracy of 0.7864 in classifying discourse types. While these features were not able to predict IS scores accurately ($RMSE = 2.6584$ and $PCC = 0.1197$), they could potentially serve as useful qualitative feedback to teachers on their classroom discourse. Future research will explore the integration of multimodal features, local session characteristics, and the generalization of the framework to diverse classroom settings.

Keywords: instructional support, automatic classroom analysis, large language models, classroom observation, classroom assessment scoring system

1. Introduction

Classroom is a significant component in education, and its quality directly impacts students' holistic development and academic performance (Engida et al., 2024; Datnow et al., 2022; Ekmekci and Serrano, 2022). Measuring classroom quality has become a focal point for many researchers (Francisco and Celon, 2020; Nisar et al., 2019). One commonly used classroom observation protocol is the CLassroom Assessment Scoring System (CLASS) (Pianta, 2008). In this study, we focus on one of its key dimensions–Instructional Support.

Instructional Support is a teaching approach characterized by engaging students in diverse and meaningful learning activities to enhance their understanding of concepts and language. It involves connecting new information to students' existing knowledge and real-world experiences while providing timely, specific, and constructive feedback to guide learning effectively (Hamre et al., 2013; Pianta, 2008). Instructional Support is a cornerstone of effective early childhood education. It not only enhances academic and cognitive

outcomes but also contributes significantly to the development of essential social and emotional skills. Extensive research has demonstrated that high-quality instructional support, delivered through evidence-based practices and effective teacher-child interactions, fosters language acquisition, social-emotional competence, and early cognitive development (Collie, 2022; Egert et al., 2020; Prediger et al., 2024; Schmerse et al., 2024; Yang et al., 2021).

Automated classroom analysis: There are numerous established methodologies that use machine learning techniques to assist classroom analysis in various aspects. Ramakrishnan et al. (2023) developed an Automatic Classroom Observation Recognition Network (ACORN), a multi-modal machine learning-based system designed to evaluate the positive climate (PC) and the negative climate (NC) of the CLASS. Wang et al. (2024) and Wang et al. (2025) introduced a speaker diarization framework to estimate how much each student speaks in the classroom, providing feedback to teachers to improve classroom quality. Sümer et al. (2021) proposed an engagement classification system that leverages video information, primarily using head pose estimation and facial expression recognition. They trained Attention-Net and Affect-Net to extract attentional and affective features instead of utilizing handcrafted features. He et al. (2024) presented a classroom person re-id system aimed at tracking the movements of both students and teachers during class. Their system efficiently addressed the "who is where when" problem while automatically capturing interactions between students and teachers. Kelly et al. (2018) proposed an automatic tool to measure question authenticity in classrooms which is a vital dimension of classroom quality. By combining automatic speech recognition (ASR), natural language processing (NLP) as well as machine learning methods, their approach eliminated the reliance on manual processes such as human observation or coding of teacher discourse.

In this paper, we investigate how we can use an LLM to recognize key aspects of Instructional Support. In support of this, we devise a codebook to define the standard for classifying classroom discourse based on content and annotate a dataset of classroom transcripts. Then, using the fine-tuned model, we assess whether we can improve upon previous work on automatically estimating IS scores. Finally, we compare whether adding new features based on this fine-tuned LLM can improve the prediction accuracy of IS scores compared to the approach in the work of Whitehill and LoCasale-Crouch (2023).

2. Dataset

To evaluate IS score in classroom for young children, we employed the National Center for Research on Early Childhood Education Pre-Kindergarten (NCRECE) in this work (Pianta and Burchinal, 2016; Pianta et al., 2017). The NCRECE dataset comprises videos recorded in kindergarten classrooms alongside IS scores manually rated by a team of 43 expert annotators. Teachers are the primary speakers in the classroom whose speech occupied over 75% of the total length of the classroom session. The number of students in each classroom varies, but their ages are consistently around four years old. The entire NCRECE dataset contains thousands of classroom videos, and for this study, we selected the same 561 videos as in the work of Whitehill and LoCasale-Crouch (2023) to organize the mini-NCRECE dataset. The averaged length of the videos in the mini-NCRECE dataset is about 14 minutes. There are 41 unique teachers in this dataset.

SentTag Dataset: Based on the mini-NCRECE dataset, we further selected 50 classroom sessions for additional annotation, forming a dataset named SentTag dataset. For each classroom session in this dataset, audio was first extracted from the video recordings. Subsequently, Whisper large-v2 (Radford et al., 2023) was employed to generate transcripts for each audio recording. The generated transcripts capture the session's dialogue content but do not recognize speakers for sentences. For each transcript, 50 consecutive sentences were randomly selected and annotated based on the type of the sentence. Details on the annotation categories and the annotation protocol can be found in Section 3. When annotators labeled the sentences, their decisions were solely based on the content of the current sentence. They did not reference video or audio information, nor did they consider the surrounding context. Each annotator worked independently and did not engage in discussions with other annotators during the annotation process. Among the four annotators, one has teaching experience as a preschool teacher, one has undergone CLASS training, one is a Ph.D. student in education, and another is a Ph.D. student in computer science. Therefore, when generating the codebook and annotating the data, the annotators can not only provide practical classification suggestions from a teacher's perspective but also take into account the requirements of applying deep learning methods to solve classification problems.

3. Qualitative Coding Process

To train models capable of detecting instructional support, we first developed and implemented a qualitative coding scheme to capture the various ways instructional support occurs in early childhood education settings. The process began with one annotator inductively coding multiple transcripts to identify common themes and draft potential coding categories. Subsequently, the annotators convened to discuss these preliminary codes and refine them into a working codebook. Each annotator independently coded several transcripts, after which the team reconvened to compare codes, resolve discrepancies, and further refine the coding scheme. This iterative process was repeated multiple times, focusing on clarifying coding criteria, addressing challenging cases through exemplar quotes, and improving procedural guidelines.

Ultimately, a finalized codebook was established, categorizing each utterance into one of five classifications: student talk, general talk, exceptionally stimulating teacher talk, questions or prompts, and teacher responses. Student talk included all the student responses or speech in the transcripts. General talk was defined as teacher speech unrelated to instructional support, such as behavior management, classroom management, reading aloud, or lecturing (e.g., "You can sit over here."). Exceptionally stimulating teacher talk referred to teacher speech that, while typically not indicative of instructional support (e.g., lecturing), was distinguished by its informative or thought-provoking nature (e.g., "You know, sometimes when we put big clumps of food in our mouth, when it goes down in our throat, we have some acids and stuff down in there. It helps break it down, too. It can help us grow just like big old dinosaurs."). Questions or prompts included any teacher utterance intended to elicit student responses (e.g., "What will happen if we eat those big clumps of food?"). Teacher responses comprised follow-ups to student contributions, such as praise, feedback, revoicing, or extending a student's idea (e.g., "Now, see, Tyrone said he [storybook character] ate the leaf because it would make him feel better. Sometimes there is

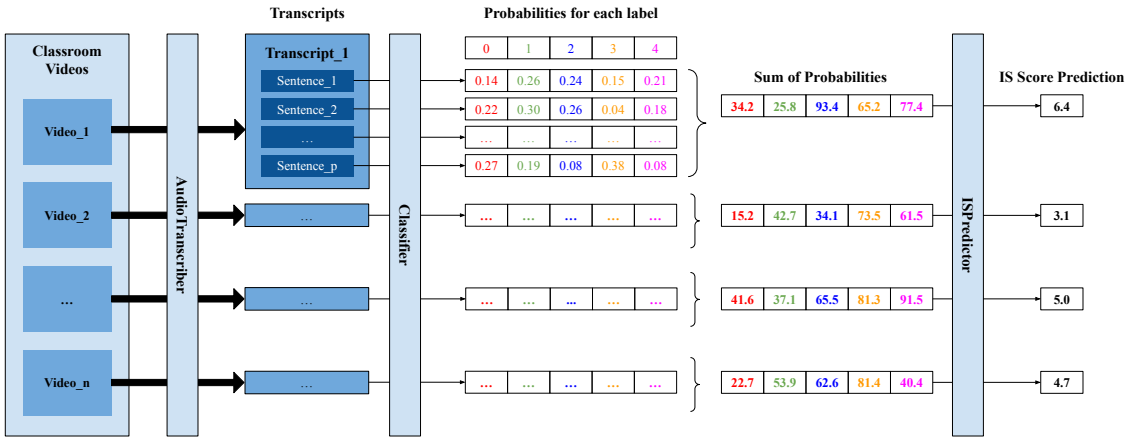

Figure 1: Instructional Support Evaluation Framework

medicine inside of the leaf, and it can make you feel better. Very good, Tyrone."). A table format codebook with definition and examples for each category is provided in Appendix A. We narrowed down to these five codes because by doing so, we could capture the vast majority of the instructional support occurring in the classroom, primarily via questions and feedback, while keeping the coding scheme relatively simple and easy for an LLM to apply across different classroom contexts.

After finalizing the codebook, the annotators coded several transcript segments, and inter-rater reliability was calculated. Discrepancies were discussed and resolved until all annotators achieved a Cohen's Kappa of at least 0.6 with each other, indicating substantial agreement (Hallgren, 2012). Following this, each annotator coded a designated number of transcript segments, consisting of 50 lines each, with a subset of segments coded by multiple annotators to ensure reliability. This process resulted in a total of 2500 lines coded to be used as training data for the models. The final Fleiss' Kappa among all annotators was 0.7120 which indicated there was substantial agreement among all annotators.

4. Framework

In this work, we proposed a framework to automatically assess IS scores. The framework consists of three components: AudioTranscriber, Classifier, and ISPredictor.

AudioTranscriber: This component is used to transcribe spoken dialogue from audio recordings into text. We employed Whisper large-v2 to achieve this component in this work.

Classifier: This component is designed to get the probability of each sentence belonging to each label. We feed sentences extracted from AudioTranscriber as the input to the classifier. For sentence S_i, the classifier provides the probabilities ($P_j^i, j \in [0, 4]$) of its association with each of the five categories. The sum of these probabilities equals 1 ($\sum_{j=0}^{4} P_j^i = 1$). For a given classroom session, the expected number of occurrences for each label in the transcripts, O_j, is calculated by summing the probabilities of all sentences in the session for each label ($O_j = \sum_{i=0}^{n} P_j^i$). The resulting values $O_j, j \in [0, 4]$ are standardized by calculating their z-scores which will be the features to infer IS scores in later procedure. In this

study, we implemented this component using the collection of Llama 3.1 models (Dubey et al., 2024).

ISPredictor: This component is used to predict the final IS scores based on the feature data produced by the classifier. We employed linear regression model to achieve this component in this work.

5. Experiments

The aim of this work is to investigate the best classifier to classify the category for each sentence. Thus, we designed experiments to compare the performance of different classifiers, rather than the entire framework. Accuracy is employed as the evaluation metric in the following experiments. It should be noted that although the Fleiss' Kappa among all annotators is 0.7120, indicating a substantial level of agreement, there are still instances where different annotators assigned different labels to the same sentence. For such sentences, we retain all the annotated labels. However, since the model produces a single prediction for each input, it inevitably results in a misclassification whenever a sentence has multiple valid labels. Consequently, even a "perfect" classifier cannot achieve 100% accuracy in this study. Based on statistical analysis, the maximum achievable accuracy is 97.85%.

5.1. With A Fine-tuned Classification Head Or Not

In this experiment, we aim to explore the impact of fine-tuning a classification head on recognizing classroom discourse categories. We compared the Llama 3.1-8B models loaded from Hugging Face's automated model loaders, AutoModel and AutoModelForSequenceClassification (Wolf, 2020). The key relationship and differences between them are as follows: The model loaded from AutoModel is the base model, which does not include task-specific heads, such as a classification head. It outputs the hidden states, which are the feature vectors extracted from the last transformer layer. In contrast, the model loaded from AutoModelForSequenceClassification is specifically designed for sequence classification tasks. It builds upon the base model above by incorporating an additional classification head, enabling it to perform classification tasks. In stead of feature vectors, it outputs logits for each class, which are then used to determine the final classification category. However, in this study, before using this model, the classification head must be fine-tuned to adapt to our specific classification task, ensuring better alignment with the categories in this work.

To obtain classification results from the pre-trained Llama 3.1-8B model loaded via AutoModel, we first extracted feature vectors from model's last transformer layer and then applied K-Means clustering (k=5), resulting in five clusters. We found the optimal permutation via brute-force search and got prediction results. The final accuracy is 40.41%. For the model with fine-tuned classification head, the accuracy is 78.64%, which is much higher. Thus, the results indicate that a fine-tuned classification head significantly improves classification accuracy for this task. In subsequent experiments, unless otherwise specified, models with classification as the task type will refer to models equipped with a classification head.

5.2. Previous Sentence Information

In this experiment, we explore whether the LLM can recognize the type of classroom discourse more accurate if it receives the previous sentence as additional context. Due to the kindergarten scenario in this dataset, many utterances are very short, and some of them consisting of only one word. The information contained in such sentences is limited, making it difficult to determine their categories. For instance, consider a sentence like "Green.". On its own, this single word provides insufficient context for category determination. However, when the previous sentence, "Underneath, what color is broccoli?" is provided, it becomes clear that this is a interaction between teacher and student and the sentence "Green." is the student's answer for the teacher's question which should belong to student talk.

Similarly, the model may face the same challenges when dealing with such cases that short utterances may not provide sufficient information for classification. Consequently, instead of inputting only the current sentence into the model, we additionally provide the content of the preceding sentence to enhance the information available for classification.

In this experiment, we utilized two models: the pre-trained Llama 3.1-8B model and the pre-trained Llama 3.1-8B-instruct model. Both models require fine-tuning on the last layer for classification. Results are detailed in Table 1. For the Llama 3.1-8B model, adding the previous sentence did not improve classification accuracy. In contrast, for Llama 3.1-8B-instruct model, including the previous sentence enhanced classification accuracy.

Model Name	With previous Sentence	Accuracy
Llama 3.1-8B	No	0.7864
	Yes	0.7340
Llama 3.1-8B-instruct	No	0.6822
	Yes	0.7099

Table 1: Accuracy results for with or without previous sentence

5.3. Text Classification vs. Text Generation

In this experiment, we aim to study the different performances of the same model when recognizing classroom discourse categories using two different modes: test generation and text classification. The Llama 3.1 collection of models is not specifically designed for classification tasks but is more suitable for answering questions and generating specific text according to given requirements. So what would the performance be like if we ask the model to give answers in text generation way instead of using a classification layer? We designed this experiment to investigate its performance.

In the text classification mode, the input to the model is the sentence to be classified, and the output is the sentence's category. In the text generation mode, the input is a prompt, and the output is the model's generated response. The prompts used in this experiment were designed as follows: The prompts included two roles, *system* and *user*, each with corresponding content. The types of content and their details were: **Task description:** Explains what we expect the model to do and describes the task. Three variations were used, ranging from concise to detailed descriptions. **Label description:** Describes the meaning of each numerical label. Two variations were employed: the first one includes

only the label numbers and their corresponding names; the second additionally provides examples for each category. **Examples:** Includes five sentences and their corresponding labels. Two variations were designed, including randomly selected sentences and chosen representative sentences for each category. **Question:** The sentence to be classified. Only one variation is used.

Besides, three templates were used to determine which types of content were assigned to each role. Some other additional hyperparameters include: $max\_new\_tokens \in [5, 10, 20, 50, 100]$, $temperature \in [0.1, 0.3, 0.5, 0.7, 0.9, 1]$, $do\_sample \in [True, False]$. A mini-dataset consisting of 10 sentences whose ground truth labels match the distribution of the overall dataset was used for hyperparameter search (including the selection of the best prompt). The top 17 hyperparameter combinations with top accuracies were further applied to test on the full dataset. We selected the first numerical value appearing in a model's response as the predicted category of the sentence. The best accuracy results are detailed in Table 2. From the results, it is evident that using the text classification mode of the model demonstrates significant advantages compared to the text generation mode, regardless of which model is used or whether the previous sentence is additionally provided.

Model Name	**Task Type**	**Accuracy**	
		without previous sentence	with previous sentence
Llama 3.1-8B	Classification	0.7864	0.7340
	Generation	0.4512	0.5117
Llama 3.1-8B-instruct	Classification	0.6822	0.7099
	Generation	0.6596	0.4781

Table 2: Accuracy results for different task types

6. Application

According to the previous experiments, the classifier that achieved the best performance was: Llama 3.1-8B, with a fine-tuned classification head, without previous sentence, and use text classification mode. We then applied this best-performing classifier to the entire framework to estimate IS scores. To evaluate the effectiveness of the framework, we employed the root mean square error (RMSE) as a measure of predictive accuracy and the Pearson Correlation Coefficient (PCC) to assess the relationship between the predicted IS scores and human-assigned IS scores. The results are as follows: $RMSE = 2.6584$, $PCC = 0.1197$. These results indicate that while classifying classroom discourse is not entirely ineffective for predicting IS scores, it contains limited relevant information. Compared to the previous work by Whitehill and LoCasale-Crouch (2023) ($RMSE = 2.32$, $PCC = 0.48$), relying solely on discourse classification features is insufficient to predict IS scores that closely align with human annotations.

To further investigate potential improvements, we concatenated the features extracted in this study with those from Whitehill and LoCasale-Crouch (2023), retrained a linear regressor, and predicted the IS scores. The results were $RMSE = 2.3675$ and $PCC = 0.4658$. These findings suggest that, incorporating classroom discourse categories as additional fea-

tures did not yield meaningful improvements in predicting IS scores. One possible explanation is that the features extracted in Whitehill and LoCasale-Crouch (2023) already contain certain aspects of classroom discourse categories. For example, their work considers whether a detected classroom utterance expresses "encourage and affirms" which is a characteristic often found in teacher feedback. In our study, teacher feedback is classified as a form of teacher responses, a predefined category in our codebook. Therefore, their extracted features may implicitly encode information that overlaps with the textual features used in this study. As a result, combining the two feature sets did not lead to a substantial improvement in prediction accuracy.

Nevertheless, the text classifiers trained in this study may still be useful for providing teachers with more specific and qualitative feedback on their classroom discourse, above and beyond what a pre-trained LLM can offer in zero-shot mode. In future work, we will explore methods for extracting more informative features and integrating them more effectively to enhance IS score prediction.

7. Conclusion

In this work, we introduced an annotation protocol based on classroom discourse types to serve as an auxiliary dimension for evaluating IS scores. The classroom discourse is categorized into five types: student talk, general talk, exceptionally simulating teacher talk, questions or prompts, and teacher responses. Using the NCRECE dataset, we constructed the SentTag dataset and annotated it following the proposed annotation protocol. Subsequently, we fine-tuned Llama 3.1 models on this dataset to emulate human annotators for this task, achieving an accuracy of 0.7864. Furthermore, we developed an automated framework which leveraging the predictions of the fine-tuned LLM to estimate IS scores. The predicted results yielded an RMSE of 2.6584, with a PCC of 0.1197 compared with the human assigned scores.

Future Work: In future research, we plan to incorporate multimodal information to enrich feature representations (Si et al., 2022). For example, extracting emotional features from speech, which is also a critical indicator for evaluating classroom quality (Gazawy et al., 2023), will allow for a more comprehensive assessment of IS scores. Additionally, we will explore how to integrate local features of a classroom session, such as the type of a single sentence, to derive more reliable overall characteristics. Furthermore, we will investigate the generalization performance of the proposed framework, including its applicability to other types of classroom settings and whether its components are compatible with other models.

Acknowledgement

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL #2019805, and also from an NSF CAREER grant #2046505. The opinions expressed are those of the authors and do not represent views of the NSF.

References

Rebecca J Collie. Instructional support, perceived social-emotional competence, and students' behavioral and emotional well-being outcomes. *Educational Psychology*, 42(1):

4–22, 2022.

Amanda Datnow, Vicki Park, Donald J Peurach, and James P Spillane. Transforming education for holistic student development: Learning from education system (re) building around the world. report. *Center for Universal Education at The Brookings Institution*, 2022.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Franziska Egert, Verena Dederer, and Ruben G Fukkink. The impact of in-service professional development on the quality of teacher-child interactions in early education and care: A meta-analysis. *Educational Research Review*, 29:100309, 2020.

Adem Ekmekci and Danya Marie Serrano. The impact of teacher quality on student motivation, achievement, and persistence in science and mathematics. *Education Sciences*, 12(10):649, 2022.

Mengistu Anagaw Engida, Ashagrie Sharew Iyasu, and Yalemwork Mossu Fentie. Impact of teaching quality on student achievement: student evidence. In *Frontiers in Education*, volume 9, page 1367317. Frontiers Media SA, 2024.

Ch DC Francisco and LC Celon. Teachers' instructional practices and its effects on students' academic performance. *Online Submission*, 6(7):64–71, 2020.

Qusai Gazawy, Selim Buyrukoglu, and Ayhan Akbas. Deep learning for enhanced education quality: Assessing student engagement and emotional states. In *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–8. IEEE, 2023.

Kevin A Hallgren. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23, 2012.

Bridget K Hamre, Robert C Pianta, Jason T Downer, Jamie DeCoster, Andrew J Mashburn, Stephanie M Jones, Joshua L Brown, Elise Cappella, Marc Atkins, Susan E Rivers, et al. Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The elementary school journal*, 113(4):461–487, 2013.

Xinlu He, Jiani Wang, Viet Anh Trinh, Andrew McReynolds, and Jacob Whitehill. Tracking classroom movement patterns with person re-id. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 679–685, 2024.

Sean Kelly, Andrew M Olney, Patrick Donnelly, Martin Nystrand, and Sidney K D'Mello. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47(7):451–464, 2018.

Muhammad Nisar, Iqbal Amin Khan, and Faridullah Khan. Relationship between classroom management and students academic achievement. *Pakistan Journal of Distance and Online Learning*, 5(1):209–220, 2019.

RC Pianta. Classroom assessment scoring system™: Manual k-3. *Paul H Brookes Publishing*, 2008.

Robert Pianta and Margaret Burchinal. National center for research on early childhood education teacher professional development study (2007-2011), April 12 2016. https://doi.org/10.3886/ICPSR34848.v2.

Robert Pianta, Bridget Hamre, Jason Downer, Margaret Burchinal, Amanda Williford, Jennifer Locasale-Crouch, Carollee Howes, Karen La Paro, and Catherine Scott-Little. Early childhood professional development: Coaching and coursework effects on indicators of children's school readiness. *Early Education and Development*, 28(8):956–975, 2017.

Susanne Prediger, Kirstin Erath, Kim Quabeck, and Rebekka Stahnke. Effects of interaction qualities beyond task quality: Disentangling instructional support and cognitive demands. *International Journal of Science and Mathematics Education*, 22(4):885–909, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

Anand Ramakrishnan, Brian Zylich, Erin Ottmar, Jennifer LoCasale-Crouch, and Jacob Whitehill. Toward automated classroom observation: Multimodal machine learning to estimate class positive climate and negative climate. *IEEE Transactions on Affective Computing*, 14(1):664–679, 2023. doi: 10.1109/TAFFC.2021.3059209.

Daniel Schmerse, Henning Dominke, Jana Mohr, and Mirjam Steffensky. Children's understanding of scientific inquiry: The role of instructional support and comparison making. *Journal of Educational Psychology*, 116(2):233, 2024.

Qi Si, Tracey S Hodges, and Julianne M Coleman. Multimodal literacies classroom instruction for k-12 students: a review of research. *Literacy Research and Instruction*, 61(3): 276–297, 2022.

Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 14(2):1012–1027, 2021.

Jiani Wang, Shiran Dudy, Xinlu He, Zhiyong Wang, Rosy Southwell, and Jacob Whitehill. Speaker diarization in the classroom: How much does each student speak in group discussions? In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 360–367, 2024.

Jiani Wang, Shiran Dudy, Xinlu He, Zhiyong Wang, Rosy Southwell, and Jacob Whitehill. Optimizing speaker diarization for the classroom: Applications in timing student speech and distinguishing teachers from children. *Journal of Educational Data Mining*, 17(1):98–125, Feb. 2025. doi: 10.5281/zenodo.14871875. URL https://jedm.educationaldatamining.org/index.php/JEDM/article/view/841.

Jacob Whitehill and Jennifer LoCasale-Crouch. Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *arXiv preprint arXiv:2310.01132*, 2023.

Thomas Wolf. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2020.

Ning Yang, Jiuqian Shi, Jinjin Lu, and Yi Huang. Language development in early childhood: Quality of teacher-child interaction and children's receptive vocabulary competency. *Frontiers in Psychology*, 12:649680, 2021.

Appendix A. Codebook

Category Name	**Definition**	**Example**
Student talk	All the student responses or speech.	A bunny rabbit. And dogs can run fast.
General talk	Teacher speech unrelated to instructional support, such as behavior management, classroom management, reading aloud, or lecturing.	You can sit over here. Okay, we're going to take questions. Hands down, my friends. Let's move on.
Exceptionally stimulating teacher talk	Teacher speech that, while typically not indicative of instructional support (e.g., lecturing), was distinguished by its informative or thought-provoking nature.	You know, sometimes when we put big clumps of food in our mouth, when it goes down in our throat, we have some acids and stuff down in there. It helps break it down, too. It can help us grow just like big old dinosaurs.
Questions or prompts	Any teacher utterance intended to elicit student responses.	Emily, do you know some words that rhyme with bat? Well, what are some words that rhyme with tree? What about rain?
Teacher responses	Follow-ups to student contributions, such as praise, feedback, revoicing, or extending a student's idea.	Very good. January, good job, Janine. Yes, but this house is of straw right here.

Table 3: Codebook with definition and examples for each category