Computational Modeling of Collaborative Discourse to Enable Feedback and Reflection in Middle School Classrooms

Chelsea Chandler
chelsea.chandler@colorado.edu
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

Thomas Breideband
Thomas.Breideband@colorado.edu
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

Jason G. Reitman
Jason.Reitman@colorado.edu
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

Marissa Chitwood
Marissa.Chitwood@colorado.edu
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

Jeffrey B. Bush
Jeffrey.Bush@colorado.edu
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

Amanda Howard
Amanda.Howard@colorado.edu
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

Sarah Leonhart
Sarah.Leonhart@colorado.edu
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

Peter W. Foltz
foltzp@colorado.edu
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

William R. Penuel
William.Penuel@colorado.edu
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

Sidney K. D'Mello
sidney.dmello@gmail.com
Institute of Cognitive Science,
University of Colorado Boulder
Boulder, Colorado, USA

ABSTRACT

Collaboration analytics has the potential to empower teachers and students with valuable insights to facilitate more meaningful and engaging collaborative learning experiences. Towards this end, we developed computational models of student speech during small group work, identifying instances of uplifting behavior related to three Community Agreements: *community building*, *moving thinking forward*, and *being respectful*. Pre-trained RoBERTa language models were fine-tuned and evaluated on human annotated data (N = 9,607 student utterances from 100 unique 5-minute classroom recordings). The models achieved moderate accuracies (AUROCs between 0.67-0.84) and were robust to speech recognition errors. Preliminary generalizability studies indicated that the models generalized well to two other domains (transfer ratios between 0.46-0.85; with 1.0 indicating perfect transfer). We also developed four approaches to provide qualitative feedback in the form of *noticings* (i.e., specific exemplars) of positive instances of the Community Agreements, finding moderate alignment with human ratings. This research contributes to the computational modeling of the relationship dimension of collaboration from noisy classroom data, selection of positive examples for qualitative feedback, and towards the empowerment of teachers to support diverse learners during collaborative learning.

CCS CONCEPTS

• **Applied computing** → **Collaborative learning**; • **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

Natural language processing, Collaboration analytics

ACM Reference Format:
Chelsea Chandler, Thomas Breideband, Jason G. Reitman, Marissa Chitwood, Jeffrey B. Bush, Amanda Howard, Sarah Leonhart, Peter W. Foltz, William R. Penuel, and Sidney K. D'Mello. 2024. Computational Modeling of Collaborative Discourse to Enable Feedback and Reflection in Middle School Classrooms. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3636555.3636917

1 INTRODUCTION

Small group collaborative learning is becoming a hallmark of 21st century K-12 pedagogy, owing to its effectiveness in nurturing skills such as critical thinking, problem-solving, and social interaction in addition to domain knowledge and disciplinary practices [16, 24, 35]. However, managing effective collaborative learning experiences can be challenging for teachers as they also need to monitor progress, provide guidance on learning activities, and support groups in productive, knowledge-building conversations [38, 50]. Past research

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636917

has highlighted a persistent deficit in students' collaboration skills [14, 16, 18, 51] and students often report demotivation and frustration with collaborative learning activities when they feel their peers' social conduct during activities is disruptive [5]. These issues have been exacerbated by the long period of remote learning in the wake of the COVID-19 pandemic, where students were required to engage in independent study, leading to challenges related to self-regulation, including low motivation and ineffective communication with peers, ultimately leaving students less prepared for collaborative learning [13, 36].

The fields of Computer-Supported Collaborative Learning and Collaborative Learning Analytics have proposed several technological solutions to serve as resources for teachers and learners to address some of these issues through capturing, analyzing, and visualizing insights from group interactions [24, 29, 40]. Yet, a recent review has highlighted several issues with existing approaches to collaborative analytics relating to features, scope, and presentation [26]. "Collaborative learning dashboards," they posit, "should be designed not to simply show a group's interactive behavior, but rather to inform and motivate future decisions" (p. 177).

Taking a step in this direction, the present paper focuses on the development of computational models of student discourse to enable collaborative analytics technologies for teachers, students, and researchers. Our focus on discourse was motivated by the fact that sharing ideas and building off of others' is the hallmark of collaboration [37, 53]. Accordingly, language emerges as a significant indicator of many aspects of collaborative learning [2, 16]. Prior research has leveraged natural language processing (NLP) techniques to assess group communication to identify indicators of Collaborative Problem Solving (CPS) proficiency, showcasing their ability to accurately detect such skills [20, 32, 33, 39, 44, 46]. NLP models can then be used to provide feedback, scaffolding skill development during collaboration [9].

Whereas prior work has largely focused on the task dimension of collaboration, emphasizing behaviors that promote success on the task at hand, we were interested in measuring and supporting the relationship dimension of collaboration as a valued measure in its own right [14, 16, 22]. Accordingly, a goal of our work was to support students in developing skills to have more accountable and uplifting interactions with one other. Based on extensive co-design with youth [7], we are developing technology that automatically identifies and visualizes expressions of socially uplifting discourse in student group speech across three dimensions, which we call Community Agreements (CAs): *community building*, *moving thinking forward*, and *being respectful*. A major component was the analytics needed to model CAs from real-world classroom speech as our system is intended for use in situations where multiple groups are simultaneously interacting. Another important aspect was model generalization, which poses a substantial challenge in the deployment of models within educational environments. Collaboration inherently spans many contexts, and the availability of pre-existing data in new contexts is uncertain. Finally, integral to the system was the extraction of *noticings* (i.e., exemplars) of CAs from student speech. We envisioned *noticings* to be presented as feedback to teachers and students, providing non-evaluative, qualitative instances of affirmative discourse to inspire discussion and bolster transparency of the underlying analytics. By exploring this novel type of feedback, we sought to help students understand how their own community-building talk manifests in collaborative learning settings.

1.1 Related Work

Collaboration entails telling and doing, implicating verbal, paraverbal, and nonverbal modalities [30, 44]. While others have have investigated the use of nonverbal signals like eye gaze, facial expression, body movement, and acoustic-prosodic features of speech to model aspects of collaboration [45, 52], we focus on linguistic approaches. Considerable research has analyzed collaborative discourse via the application of advanced NLP techniques. This has been operationalized by obtaining language data (e.g., from text chats or transcribed spoken interactions) to model CPS skills such as negotiation, information sharing, regulation, and argumentation [12, 20, 39, 44, 46]. Insights have been used to understand emerging social networks and collaborative patterns [31] and to predict outcomes like learning improvement [41] and task performance [15]. Prevalent NLP methodologies have involved using words, phrases, or part-of-speech tags as features in classification, however recently there has been a growing trend in utilizing pre-trained neural networks, with several studies showing their effectiveness with collaborative discourse [25, 33].

We focus on two key computational issues: model generalization and speech recognition in real-world settings. While generalization, or the ability of models to transfer knowledge from domains in which they were trained to new ones, is one of the primary desiderata of NLP, few collaboration analytics studies have addressed this [8, 32]. Many focus on a single domain or classroom curriculum, leading to models with highly specific knowledge, often lacking the high-level representations needed for broader learning and applications. Extending the scope of collaborative analytics research to encompass diverse contexts and real-world educational settings would enhance their relevance. However, analyzing student speech engenders significant computational challenges. Automated Speech Recognition (ASR) accuracy is substantially lower for children's speech than adults' [42], and typical signal-to-noise ratios in the classroom can range from −7 dB to +5 dB [23], further impeding data quality. Thus, pertinent questions are the extent to which ASR errors cascade to affect downstream classification and how to increase robustness to noisy speech [4].

Beyond computational modeling, studies have investigated tools for providing feedback for classroom collaboration skills. One example, *CPSCoach*, employed ASR and NLP to offer college students personalized feedback on their CPS skills during video conferencing sessions, supplemented by learning materials for skill enhancement [47]. A similar system, *IneqDetect*, visualizes individual speaking time to help students reflect on team communication [27]. Feedback solely rooted in speaking time may overlook contributions that advance the team's goals in different ways. In a study conducted to evaluate the impact of feedback on students' collaboration skills, findings indicated that although there was no significant improvement in collaboration skills, it was recommended that feedback be prompt, include important metrics, and offer explanations, along with personalized guidance on enhancing collaboration [10]. We found feedback in the form of model *noticings* to be an underrepresented approach. AI explainability techniques such as saliency maps, attention mechanisms, and tools like LIME or SHAP [28] are often employed to visualize and explain the model's decision-making process to end users, however in the context of a classroom, providing concrete examples may allow students to better grasp and trust the model's behavior.

1.2 Current Study

We focused on the development of CA coding schemes, training and evaluation of NLP models, and selection of *noticings* for feedback. Specifically, we sought to answer the following research questions:

(1) What are effective indicators of Community Agreements in small group collaborative learning discourse?
(2) How can ASR challenges in noisy classroom environments be addressed for classifying Community Agreements?
(3) To what extent do the models generalize to new curricula without incorporating data from the target contexts?
(4) What are the most effective ways to identify *noticings* of student discourse for qualitative feedback?

To address (1), we employed a multi-faceted approach to developing a robust CA coding scheme, aligning indicators from a validated CPS framework to CAs, applying this mapping to classroom discourse. For (2), we fine-tuned RoBERTa language models on noisy classroom data, involving both ASR and human transcripts. For (3), we evaluated our models on labeled data from small group work from an educational physics game and a block programming game. Finally, for (4), we implemented four computational strategies for identifying and ranking examples of positive instances of CAs in student speech. We collected human ratings on the usefulness and validity of a subset of these *noticings*. Together, we make progress in advancing collaborative discourse analytics to facilitate formative feedback in real-world classrooms.

Key novel aspects of our work in light of prior research includes fully-automated modeling of Community Agreements in child speech from noisy classroom environments, and the development of methods for selecting *noticings* to promote reflection and transparency.

2 DATA

Approval for all procedures was obtained from the designated Institutional Research Boards, and analyzed data involved students who provided their assent and whose parents or legal guardians provided consent.

2.1 Data Collection

Data was collected from urban, rural, and suburban public middle school classrooms in Colorado, USA during the preceding two years. Students participated in small group work during a curriculum unit called "Sensor Immersion", where they programmed and wired environmental sensors to collect data about their surroundings. The curriculum revolved around an interactive display called the Data Sensor Hub (DaSH), which was a central point of exploration for the students [6]. Students explored the system, constructed scientific models, and acquired the skills to replicate its functionality in the scope of their own investigations, involving authentic debugging and engineering practices as they relate to sensor technology and pair programming within group settings [11].

A tabletop omnidirectional microphone (Yeti Blue) was placed at each group table during recorded Sensor Immersion sessions. This microphone was selected after considering audio quality, affordability, power source, form factor, and ease of use. Due to the reliance on a single omnidirectional microphone to capture the conversations of multiple students, the collected data was inherently noisy [4, 42]. Video data was also collected with an iPAD camera.

Within each recording, we identified five 5-minute segments from the group work portion of the lesson, typically confined to the middle of the recording, as the initial and final portions tended to have less relevant discussion. We systematically listened to each segment. If it met a 20-word threshold, it was included as a sample, otherwise the next segment was evaluated, and so on. If none of the segments met the criteria, the recording was excluded from analyses.

The dataset consisted of 100 5-minute excerpts of small group work collected from 164 unique students (73 dyads, 7 triads, and 6 tetrads) under the guidance of 14 teachers. Demographic information from individual students was not collected, however the demographic composition of the school districts indicated that the sample was diverse.

2.2 Human and Automated Speech Recognition Transcription

Recordings were transcribed manually by three humans resulting in a total of 16,515 transcribed utterances. In cases where speech was too noisy, transcribers denoted some or all of the utterance as "[inaudible]". Human transcriptions included notes such as laughter, singing, or crosstalk (e.g., "[laughter]", "[singing]", "[crosstalk]"), as well as who the student was addressing (e.g., "[addressing group]", "[addressing other]"). Individual utterances were automatically extracted from recordings using the human annotated timestamps and transcribed with Whisper, a state-of-the-art open-access ASR model trained on a substantial dataset of 680,000 hours of speech [34].

We computed the word error rate (WER) of each utterance, defined as $\frac{substitutions+deletions+insertions}{total words}$. The mean student WER was 68.9%, highlighting challenges of working with real classroom speech. There was high variability in WERs, however, analyses suggested that this can be improved by filtering on ASR confidence value. For example, focusing only on utterances with confidence values greater than the 40th percentile reduced the student WER to 32.6%.

2.3 Human Coding of CA Labels

Utterance-level CA Annotations. To assess CAs within student group conversations, we devised a novel mapping from CPS indicators (derived by [48, 49]). This generalizable CPS coding framework consists of three facets of CPS established by literature (constructing shared knowledge, negotiation/coordination, and maintaining team functions) and 18 indicators observable in group conversations. The scheme was validated through empirical studies of triads across contexts including differences in participant age, co-locality and virtuality, and task type [48], and was further validated as a

Table 1: Collaborative Problem Solving (CPS) indicator mappings to Community Agreements and their mean frequency in the final filtered dataset. Corresponding examples from the Sensor Immersion dataset are given.

Community Agreement (Mean Freq.) CPS Indicators	**Corresponding Examples from Sensor Immersion Data**
Being Respectful (9%)	
Responds to others' questions or ideas	"Yeah, you spelt it right."
Asks others for suggestions	"Okay, what should we do now?"
Compliments or encourages others	"Oh that's just dope. I love that color."
Apologizes for one's mistakes	"[...] Sorry. My bad. Scroll over to where you can see the whole thing."
Community Building (15%)	
Talk about the challenge situation	"No I think you wanna get rid of that one."
Confirms understanding	"Like that?"
Discusses the results	"Uh, it's not really working."
Provides instructional support	"If your pool has a heater, the bar will go up a lot."
Asks others for suggestions	"Okay. What do you wanna, what do you want me to draw?"
Moving Thinking Forward (9%)	
Proposes (in)correct solutions	"Place on shake."
Strategizes to accomplish task goals	"So now I think we push download and see what happens."
Asks others for suggestions	"How do I get rid of this?"
Provides reasons to support a solution	"[...] we have to make that lower because its always gonna play if its 50."
Questions/corrects others' mistakes	"No no. We already said Hello. [...]"

predictor of CPS performance both as individual indicators [49] or as temporal clusters of indicators [54].

We adapted the original CPS scheme to the present study wherein four coders, including one expert coder who helped develop the original scheme, iteratively annotated a subset of Sensor Immersion transcripts, noted points of disagreement, added clarifications and examples to the coding scheme, and discussed inconsistencies until consensus. After multiple training sessions, transcripts were divided among coders who individually coded each utterance for the presence of indicators (alongside the video for context). To ensure reliability, the expert coder reviewed each coded transcript. We then created a novel mapping of CPS indicators to CAs (Table 1), using definitions from OpenSciEd (a free curriculum from which the CA framework was derived) and by consulting with experts in OpenSciEd. As shown in the table, approximately 9% to 15% of the utterances contained a CA, and the labels are not mutually exclusive.

Recording-level Expert CA Ratings. For further validation, we adopted a high-level approach to CA rating, where two experts (education and language researchers with experience in observing classroom activity) applied subjective ratings of CAs at the recording-level. They were asked to "rate each video from 1-5 for each [CA] (or indicate if it was not scorable)." Inter-rater reliability as assessed by quadratic kappa was high for *being respectful* (0.90) and *community building* (0.72) but lower for *moving thinking forward* (0.34). Mean scores were 3.58, 3.14, and 3.47, respectively.

Utterance-level CPS mapped human annotations were averaged to thhe recording-level and Spearman correlations with the expert ratings were ρ = 0.44 (*moving thinking forward*), ρ = 0.25 (*community building*), and ρ = 0.42 (*being respectful*). Together, this is a theoretical and methodological extension of the CPS framework to a novel construct.

2.4 Data Processing

Teacher and non-consenting student utterances, and those directed at teachers or students from other groups were excluded. Human and ASR transcripts were normalized to ensure consistency and accuracy for future classroom use. Normalization involved the replacement of hyphens with spaces, removal of transcriber notes (e.g., "[inaudible]") and punctuation, and conversion to lowercase. After preprocessing, the final filtered dataset comprised 9,607 utterances. On average there were 2.3 students per recording (SD = 0.6, range = 2-4), 1.4 recordings per student (SD = 1.0, range = 1-5), 67.6 utterances per recording (SD = 29.8, range = 20-194), 41.1 utterances per student (SD = 33.8, range = 4-228), 4.6 words per utterance (SD = 4.0, range = 1-47), and 311.6 words per recording (SD = 142.4, range = 94-906).

3 METHODS

3.1 Deep Transfer Learning with RoBERTa

We developed three computational models: one each for *community building*, *moving thinking forward*, and *being respectful*. The models were pre-trained RoBERTa language models - a variant of the BERT language model with a multilayer bidirectional transformer architecture. The RoBERTa models were individually fine-tuned on the filtered Sensor Immersion dataset annotated with binary labels for each CA. Fine tuning pre-trained large language models is a NLP technique that allows adaptation of powerful domain-agnostic models to specific tasks. Utterances were tokenized with the RoBERTa tokenizer which incorporates padding and truncation to ensure that all text sequences are of uniform length. The fine-tuning process comprised a batch size of 32, 50 training epochs, a learning rate of 5e-06, and 50 warmup steps. Hyperparameters were based on previous research on fine tuning RoBERTa models for CPS prediction on a different dataset [32]. Minor adjustments were made but we did not do massive hyperparameter tuning.

We used a stratified recording-level 10-fold cross-validation framework for model training and evaluation. The dataset was divided into 10 folds, where the proportions of positive samples from each CA class were approximately equivalent and utterances from single recordings were not split between folds. All experiments were conducted based on these initial cross-validation splits, ensuring consistency and reproducibility in our analyses. For each round of cross validation, train (8 folds), validate (1 fold), and test (1 fold) sets were created. The train set was used to fine-tune the models, generating checkpoints throughout the process. The best model checkpoint for a round was determined by testing the checkpoints on the validate data. The checkpoint with the best performance on the validate data was then used to test the held out test set and generate final predictions. Our primary evaluation metric was the area under the precision-recall curve (AUPRC), chosen for its robustness in handling class imbalances compared to metrics such as the area under the receiver operating curve (AUROC) [19]. The reported percent above chance represents AUPRCs adjusted to account for baseline occurrence variations. We present results testing on both human and ASR transcripts.

We also determined the extent that utterance-level human labels and model predictions, aggregated to the recording-level, agree with the subjective expert perception of CA usage per recording, which goes beyond language and incorporates video context. Accordingly, we correlated aggregated utterance-to-recording level human annotations and model predictions of the CAs with the recording-level expert rating (1-5 scale). Recording-level aggregations leverage the principle of aggregation to reduce noise for an overall estimate of CA prevalence per session. These correlations were performed on a subset of recordings (N = 31) that had recording-level expert judgments.

3.2 Human and ASR-Augmented Training

While training and evaluating NLP models on human-generated transcripts provides a "gold-standard", it remains critical to consider how resilient models are to data collected in realistic conditions. Thus, in addition to training solely on human transcripts, following [4], we also incorporated ASR-augmentation. In this setting, for each human-transcribed utterance, we added its ASR-transcribed counterpart for training and testing. This new, effectively doubled, ASR-augmented dataset was shuffled within recording such that the human and ASR transcripts of a single student utterance were not consecutive in model training. The same stratified recording-level 10-fold cross-validation folds were retained between the approaches as our emphasis was in evaluating the relative performance shift with the utilization of ASR-augmented training.

3.3 Generalization

We sought to determine the extent that the models fine-tuned on Sensor Immersion data transferred to new tasks and curricula, without the necessity of further training with data from the target context. Specifically, we evaluated the models on two additional data sets: (1) Physics Playground - an educational physics game, and (2) Minecraft Hour of Code - a block programming game. These datasets were collected as part of previous research on remote CPS [43]. These datasets involved remote collaboration among N = 288 university level students (average age = 22). The Physics dataset contained 46,679 utterances from 74 unique groups and the Minecraft dataset contained 10,976 utterances from 32 unique groups. Participants from both datasets self-reported gender as follows: 54% female, 41% as male, 1% as non-binary/third gender, and 4% did not report. Participants self-reported race as follows: 48% Caucasian, 25% Hispanic/Latino, 17% Asian, 3% Black or African American, 1% American Indian or Alaska Native, 3% Other, and 3% did not report. Both datasets were transcribed with IBM Watson, which provided an additional source of variability.

Following [32], the primary metric we used to evaluate the ability of the models to accurately predict CAs in new domains was the Transfer Ratio (TR):

$$TR = \frac{AUROC_{across\ task} - 0.50}{AUROC_{within\ task} - 0.50}$$

The TR measures the relative decline in a model's performance when training and evaluating on data from different tasks (across task evaluation), as compared to within the same task (within task evaluation). A TR value of 1 would signify perfect generalizability, with no decline in performance due to across-task evaluation. Both the numerator and denominator of the TR equation are adjusted by subtracting 0.50 to quantify performance difference over chance.

3.4 CA Noticings

Selecting Noticings. Our models returned a probability between 0 and 1 for each utterance, with 0.5 as a threshold for positive predictions. Probabilities closer to the threshold were low confidence and those closer 1.0 were high confidence positive predictions. Preliminary examinations indicated that it was insufficient to provide highly confident positive predictions as *noticings* as they often lacked context and viability to serve as learning examples. For instance, versions of the phrase "yeah" tended to comprise the high confidence predictions for *being respectful* as this was a highly typical exemplar of the indicator *Responds to others' questions or ideas*, but do not serve as good reflection opportunities. More context and varied examples are needed for model transparency and to provide qualitative feedback.

As such, we explored four computational strategies for identifying and ranking student speech examples that demonstrate community building behaviors. These strategies included (1) rule-based, (2) semantic similarity to student co-negotiated classroom agreements, (3) semantic similarity to CA expert definitions, and (4) topic modeling. The co-negotiated agreements were collected from students in two middle school classrooms where teachers assisted students in categorizing their ideas for definitions of the CAs. A subset of the phrases used in both semantic similarity approaches, as well as psuedocode for the rule based approach and the topics

Table 2: Overview of the four computational approaches for selecting and ranking *noticings*.

CA	Rule-Based (Pseudocode)	Semantic Similarity: Expert Defined Indicators	Semantic Similarity: Co-Negotiated Classroom Agreements	Topic Modeling (Representative Words)
Respect	If the model *prediction probability* exceeds a threshold: Categorize into lists by word count: A. N >= 5, B. 3 <= N < 5, C. 2 >= N < 3 Sort A, B, C individually by *prediction probability.* Concatenate sorted lists A, B, C.	- Offers thanks, apology - Complies with request for help, offers help - Compliments another student - Jokes to build rapport - Asks for help or a question	- Ask others before just doing things - Give everyone time to think - Don't put group down - Use your manners - Treat everyone equal - Don't talk over others	- Helping (*we, work* - Encouraging (*fun, nice*)
Thinking		- Gives reasons, evidence for actions or statements - Realization, insight - Conjecture	- Elaborate on thoughts to enforce clarity - Think outside the box - Give everyone a chance to think - Speak your own truth - Here's what went wrong	- Software (*scroll, download*) - Numbers (*zero, one*) - Visuals (*draw, graph*) - Science (*soil, water*) - Hardware (*wires,* connect) - Sensor (*sound, gator*)
Community		- Asks for input - Asks for ideas - Checks for understanding - Bids, requests to contribute	- Be dependable - Contribute equally - Show that you care - Help with no motive - I can help you - Your part looks good	- Collective ownership of task (*we, do*) - Establishing roles (*driver, navigator*) - Results, next steps (*show, think*) - Reorienting team (*get, should*)

(with corresponding representative words) chosen from the topic modeling approach are detailed in Table 2.

For the **rule-based** approach, initially all positive predictions for utterances comprising more than one word and an ASR confidence score (Spearman ρ correlation to WER = -0.59) greater than 0.5 were identified. The set of two word or larger positive predictions with sufficient ASR confidence were then divided into three separate lists based on their word count (utterances with more than five words, those with three to five words, and those with three words or fewer). Following this categorization, each list was individually sorted by prediction probability. Finally, these three sorted lists were concatenated together, arranged in descending word count order. This prioritized longer utterances, which inherently contain more contextual information.

Both **semantic similarity** techniques follow the same process, but differ in the set of phrases used to sort by. Starting with the same list of utterances as the rule-based approach, the utterances were projected into an embedding space using the BERT language model. Embeddings adhere to the distributional hypothesis theory [21], which posits that texts sharing similar meanings also possess similar representations, and are positioned in closer proximity within the embedding space. Consequently, the greater the cosine similarity between two phrases, the more semantically related they are. One version of this approach sorted *noticings* by semantic proximity to real co-negotiated classroom agreements, whereas the other sorted by proximity to expert definitions of the CAs. A cosine similarity matrix of student utterances to phrases of interest was created. *Noticings* were iteratively chosen from the matrix in order of highest cosine similarity to a phrase. A threshold of 0.80 was chosen such that utterances with cosine similarity to all phrases below 0.80 would not be considered. The algorithm started by considering all possible phrases, and as utterances were selected, the corresponding phrase was no longer considered in that iteration. We considered only phrases that had not yet been chosen so as to diversify the *noticings* rather than overselect from a single phrase. This continued until all utterances with cosine greater than 0.80 to a phrase were chosen.

The **topic modeling** approach began with the utterances filtered by the rule-based approach. We then harnessed topic models to choose utterances that represent explicit topics of interest. Topic modeling is a classic NLP technique that identifies latent topics or themes within a collection of documents, enabling the discovery of underlying patterns and structures [3]. We trained three BERTopic models [17] on the set of all positive instances of each CA from the training data and empirically chose a number of topics per model that carefully balanced over-simplification with over-segmentation, ultimately aiming for one that best captured structure in the data. Once the topic models were created, we identified the topics that aligned with expert definitions and perceptions of the CAs, excluding those that were too broad or irrelevant. The selection and ranking of *noticings* was performed by filtering the utterances that are labeled as a topic of interest by a topic model and this set was then sorted by NLP model prediction probability.

Human Ratings to Validate Noticings. The effectiveness and accuracy of these approaches were evaluated with human ratings. We recruited a total of 33 raters who completed secondary education or above through the decentralized Prolific survey platform [1] to read *noticings* and judge whether they are examples of *community building*, *moving thinking forward*, and *being respectful*. Each *noticing* received a total of 3 human ratings. Raters (Female = 17, Male = 16) were geographically diverse with an age range of 20 - 72 who indicated English as their primary and first language. Raters were compensated US$6.00, with a median completion time of 14 minutes and 40 seconds.

We randomly sampled 200 positive predictions (from human transcripts only to avoid the confound of ASR errors) for each CA and subsequently filtered and ranked them with each approach. We first computed correlations between the rankings given by each approach in order to verify that they were not associated with one another. There was considerable overlap in the selections of the two semantic similarity methods (Spearman correlation ρ = 0.75) so we proceeded only with the semantic similarity to classroom agreements. Correlations among the other methods were between ρ = -0.25 and ρ = 0.18, suggesting considerable variability.

The survey was split into three sections, one for each CA. Each category contained 20 utterances *noticed* by the methods, 5 random utterances that were *not noticed* by any method, as well as one attention check totaling 78 items per participant. Items were individually presented with the following instructions: "Please indicate the extent to which the phrase below is an example of [Respect or Thinking or Community].") using a scale ranging from 1 (not at all) to 5 (extremely). The definition of each CA was accessible for the raters to view on each question. To control for quality, raters completed two different tests before they were considered eligible to participate: (1) a screener validation, consisting of questions replicated from the Prolific internal screening system in order to disqualify those with inconsistent responses and (2) a comprehension check, involving 12 items that tested comprehension of the CA definitions.

4 RESULTS

4.1 Accuracy of RoBERTa Models

Utterance-level Model Accuracy. The results of fine-tuning the RoBERTa models on human and ASR-augmented data are given in Table 3. Overall, models outperformed chance as evidenced by AUROC scores greater than 0.5 (mean = 0.72, SD = 0.07) and AUPRC scores exceeding the base rate (mean percent above base rate = 149%, SD = 87%). In general, testing on human transcripts provided an upper bound, and surpassed random chance by a substantial margin. As expected, we observed performance degradation when testing the models with the noisier ASR transcripts. Specifically, when training on human transcripts only, we found AUPRC decreases of -14.81%, -6.25%, and -38.71% between testing on human vs ASR transcripts for *moving thinking forward*, *community building*, and *being respectful*, respectively.

The incorporation of ASR-augmentation in fine-tuning yielded significant enhancements in testing with both human and ASR transcripts for all CAs. We found a notable improvement in the comparison between testing on human vs ASR transcripts. Specifically, we found AUPRC changes of -3.33%, +9.37%, and -7.69% between the test sets for *moving thinking forward*, *community building*, and *being respectful*, respectively. In addition to reducing the performance gap between testing on human vs ASR transcripts, we found that overall predictions were more accurate for both transcript types. The percent improvements of the AUPRC metric on human transcripts was 11.11%, 0.0%, and 25.8% and the change was greater for the Whisper transcripts, with AUPRC improvements of 26.09%, 16.67%, and 89.47%, for *moving thinking forward*, *community building*, and *being respectful*, respectively. ASR-augmentation proved to be the most beneficial for *being respectful.*

Focusing on the improved ASR-augmentation approach, we found fairly consistent results between the CA models. Testing on human transcripts yielded AUROCs between 0.77 and 0.84, and between 0.67 and 0.71 for ASR transcripts. ASR-augmentation introduced a degree of diversity into the training data, encompassing variations in speech patterns and errors often encountered in real-world, noisy classroom environments. We hypothesized that this diversity allowed our models to adapt better to the intricacies of student discourse, ultimately resulting in improved generalization to both transcript types. These combined effects highlighted the effectiveness of ASR-augmentation in refining the language model's capabilities for this task. While comparing our results with similar research (e.g., [33]) is complex due to different labels, base rates, and datasets, we found our results to fall within previously cited accuracy ranges.

Relationships with Expert Ratings. The correlations (computed with Spearman's rho ρ) between recording-level expert ratings (on a 1-5 scale) of the CAs and aggregated utterance-to-recording level human annotations, human transcript trained model predictions, and ASR-augmented model predictions (utterance-level labels and predictions were averaged to the recording-level) were generally moderate as shown in Table 4. We expected this, as individual perceptions of CAs will inevitably differ from a CPS indicator mapped version. The ASR-augmented model correlated more strongly with the expert ratings than the model trained on human transcripts across the board, presumably because this model was more accurate. Surprisingly, it was also more strongly correlated than the ground-truth human annotations for *moving thinking forward*; correlations were on par for *community building*, and lower for *being respectful.* Overall, the correlations from the ASR-augmented model provide some confidence in the automated measurements.

4.2 Generalizability of Sensor Immersion Models

We found that the models trained on Sensor Immersion data generalize well to the Physics Playground and Minecraft domains. TR comprises the across task AUROC (the AUROC of the model trained on Sensor Immersion and tested on the transfer datasets) and the within task AUROC (the AUROC of the model trained and

Table 3: Performance metrics from the three models (Agreement) tested on Human and Whisper ASR transcripts, averaged over 10 folds of stratified recording-level cross validation. Human and Whisper base rates differ due to missing Whisper transcripts in the presence of noisy or unintelligible utterances.

Train Set	Test Set	CA	AUROC	AUPRC	Base Rate	% Above Base Rate	Precision	Recall
Human	Human	Thinking	0.75	0.27	0.09	200%	0.34	0.37
		Community	0.72	0.32	0.15	113%	0.38	0.34
		Respect	0.77	0.31	0.09	244%	0.42	0.38
	Whisper	Thinking	0.64	0.23	0.12	92%	0.33	0.21
		Community	0.61	0.30	0.22	36%	0.37	0.18
		Respect	0.62	0.19	0.13	46%	0.26	0.30
ASR-Augmented	Human	Thinking	0.82	0.30	0.09	233%	0.40	0.22
		Community	0.77	0.32	0.15	113%	0.37	0.30
		Respect	0.84	0.39	0.09	333%	0.51	0.29
	Whisper	Thinking	0.71	0.29	0.12	142%	0.46	0.17
		Community	0.67	0.35	0.22	59%	0.42	0.15
		Respect	0.71	0.36	0.13	177%	0.38	0.22

Table 4: Correlations between recording-level expert ratings (1-5 scale) and (1) utterance-level human annotations, (2) utterance-level human transcript trained model predictions, and (3) utterance-level ASR-Augmented model predictions, all averaged to the recording-level.

CA	Correlations with Holistic Expert Ratings on 1-5 Scale				
	Human Annotations (Ground Truth)	Human Transcript Model		ASR-Augmented Model	
		Human Test	ASR Test	Human Test	ASR Test
Thinking	ρ = 0.44	ρ = 0.47	ρ = 0.41	ρ = 0.68	ρ = 0.69
Community	ρ = 0.25	ρ = 0.10	ρ = 0.09	ρ = 0.41	ρ = 0.30
Respect	ρ = 0.42	ρ = 0.03	ρ = 0.06	ρ = 0.11	ρ = 0.17

tested within each transfer dataset). For Minecraft, the whithin task AUROCs were 0.88 (*moving thinking forward*), 0.87 (*community building*), and 0.88 (*being respectful*), and the across-task AUROCs were 0.59 (*moving thinking forward*), 0.71 (*community building*), and 0.67 (*being respectful*), resulting in TRs of 0.56 (*moving thinking forward*), 0.73 (*community building*), and 0.46 (*being respectful*). For Physics, the whithin task AUROCs were 0.83 (*moving thinking forward*), 0.84 (*community building*), and 0.86 (*being respectful*), and the across task AUROCs were 0.62 (*moving thinking forward*), 0.68 (*community building*), and 0.75 (*being respectful*), resulting in TRs of 0.67 (*moving thinking forward*), 0.74 (*community building*), and 0.85 (*being respectful*). While the datasets substantially differ in content, student age, ASR type, and CA base occurrence (Minecraft: 0.22, 0.27, 0.25 and Physics: 0.19, 0.34 0.13 for *moving thinking forward*, *community building*, and *being respectful*, respectively), the models produced across task AUROCs well within the range of Sensor Immersion results.

The TRs suggest better generalization overall to Physics (mean TR = 0.75) and less so for Minecraft (mean TR = 0.58). We found good generalization for *community building* across both transfer tasks (TRs of .73 and .74) whereas *being respectful* generalized well for Physics (TR of .85) but not Minecraft (TR of. 56). TRs for *moving thinking forward* (.56 and .67) were intermediate. It appears that the type of words commonly used in positive instances of *community building* had more overlap among the three datasets than in the other CAs as *community building* is less related to the content of group work. With that said, the sensor-related content words in the training data - as opposed to the Physics and Minecraft-related content words in the transfer data - caused a lack of transfer, most noticeably for *moving thinking forward.* Error analysis confirmed that the models specifically suffered in instances with domain-specific verbiage. An example of a false negative due to domain shift (underlined) is, "okay so next time you want to start from the top so that it swings you can hit control right click and it will delete". Further work is necessary to investigate these shortcomings as well as to build robust models that can transfer to new domains with little to no human annotated data.

4.3 Noticings User Study

Our analyses focused on the highest ranked (rank of 4 or 5) 423 utterances by the rule-based (n = 211), semantic similarity (n =

Table 5: Example highly ranked *noticings*. Average human rating of each utterance is given in parentheses.

CA	Rule-Based	Semantic Similarity	Topic Modeling
Respect	Is there anything I can do to help? (*rating 5*)	Oh, I think I know why (*rating 3*)	Here, I can help you (*rating 4.67*)
Community	Wait, what do we do with this though? (*rating 4.25*)	We didn't know where to place that (*rating 3.67*)	Is that all we have to do? (*rating 4.25*)
Thinking	K you press one, I press one. Three, two, one beep (*rating 4*)	Let's see what it does now (*rating 4*)	Should we go with a smile face right here? (*rating 3.7*)

202), and topic modeling (n = 103) approaches. Of these, 336 were selected by a single method, 81 by two, and 6 by all three. The ratings were averaged across raters, which was the main dependent variable in our analyses. On average, the ratings hover around the midpoint (mean = 2.27, SD = 0.84) of the 1-5 scale suggesting that the identified *noticings* were perceived as being reasonable examples of the CA categories, illustrated in Table 5.

For the main analysis, we regressed the mean rating on CA (with *community building* as the reference group) x Method (with rule-based as the reference group) interaction and the recording as a random intercept (more complex random effects structures resulted in convergence errors). There was no significant interaction (p = .39), so we re-ran the model with main effects only. Results indicated a significant main effect for CA (F(2) = 17, p < .001). Post hoc comparisons with false discovery rate corrections for multiple comparisons indicated that ratings were significantly lower (p < .01) for *community building* (M = 2.07) compared to *being respectful* (M = 2.33) and *moving thinking forward* (M = 2.42); the latter two were on par (p = .33). There was also a main effect of method (F(2) = 7.6, p = .02) with semantic similarity (M = 2.16) being rated significantly (p < .01) lower than rule-based (M = 2.38), but not significantly different (p = .32) from topic modeling (M = 2.29), which was on par with rule-based (p = .34).

5 DISCUSSION

Our overall focus was on the computational modeling of the relationship dimension of collaboration in classroom environments. We utilized prior research on Collaborative Problem Solving (CPS) to define three dimensions of collaboration, referred to as Community Agreements (CAs). We investigated the accuracy of three fine-tuned RoBERTa language model classifiers for each CA in noisy classroom data. The classifiers far exceeded chance, though overall accuracy was modest. The use of ASR-augmentation in fine-tuning the models made them more resilient to ASR errors and increased overall robustness, as demonstrated by improved accuracy when testing on human and ASR transcripts. We found that models trained strictly within the Sensor Immersion dataset could modestly generalize to new domains, even those in very different settings. Finally, we found that rule-based and topic modeling approaches to filtering and ranking *noticings* better aligned with human perception of the CAs than a semantic similarity approach.

A major application of this research involves the practical use of these collaboration analytics models within real educational environments. We are in the process of integrating the models developed here into an AI technology that provides automated formative assessments of CAs via visual representations of CA prevalence and model *noticings* during small group collaborative learning. Teachers are then able to facilitate a discussion around the predictions - both in regard to the successful instances of collaboration that were *noticed* in the classroom and also by interrogating current limitations of AI systems. Another application is to provide automated assessments of collaboration as a variable for future research studies on collaborative learning, where manual annotation is a bottleneck.

Like all studies, ours has limitations. With respect to the coding of CAs, the low inter-rater reliability kappa for *moving thinking forward* and the low correlation with expert rating for *community building* are indeed limitations, however the overall convergence across raters and approaches encourages us that this is a productive first step for the robust validation of the CA measure. Next, only one type of NLP model was considered - fine-tuned RoBERTa language models. The absence of a comparison to other NLP architectures restricts our ability to assess relative performance and effectiveness when compared to alternative approaches. We also did not collect demographic data from individual students, which precluded an analysis of bias/fairness of the models. Another limitation pertains to the models' focus on speech-only, whereas CAs may also be expressed nonverbally. Our generalizability assessment was preliminary with mixed results. As such, the applicability to different domains or populations may vary and should be considered with caution. With respect to the *noticings* user study, we only collected feedback from adult raters with limited exposure to the problem space. This choice, while deliberate for certain research objectives, restricts the breadth of insights we can draw from their perspectives. Lastly, whereas the present paper focused on validating the models and selecting *noticings*, we have yet to investigate how these models perform when integrated into future interventions.

Future work includes developing improved approaches to modeling student speech, improving generalization to new domains, investigating and mitigating potential biases, moving towards a multimodal approach, soliciting feedback of *noticings* from users, and assessing overall impact, fairness, and equity. For improving the language models, we plan to investigate other pre-trained language models architectures. The generalization of NLP models to new domains is a fast-evolving research area in NLP, called the cold start problem. This problem can be addressed with techniques such as zero (or few) shot learning. Model bias will be assessed and potentially addressed with techniques such as adversarial debiasing. For further validation and improvement of our *noticing* selections, we will seek feedback from users (e.g., students and teachers), potentially using human ratings for training supervised NLP models that learn to select *noticings*. Finally, while student language is an important indicator of collaborative learning, we plan to incorporate aspects of nonverbal signals such as eye gaze, acoustic-prosodic features of speech, facial expressions, and body movements in order to create a more thorough and robust model of collaboration.

6 CONCLUSIONS

This study leveraged noisy classroom data, a context often underrepresented in research, to explore and model the relationship dimension of collaboration in the form of three Community Agreements, shedding light on the dynamics of collaborative discourse within real-world classroom environments. We successfully modeled the Community Agreements with real-world speech, investigated their generalizability to two other datasets, and we placed special emphasis on fostering deeper insights into the dynamics of collaboration in the form of *noticings* of student discourse to enhance the overall learning experience. By providing concrete illustrations of model predictions, we promoted deeper understanding of the model's decision-making process thereby fostering transparency in AI-augmented educational settings.

ACKNOWLEDGMENTS

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

REFERENCES

[1] [n. d.]. Prolific. https://www.prolific.com. Accessed: 2023-09-27.

[2] Jessica Andrews-Todd and Carol M. Forsyth. 2020. Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior* 104 (2020), 105759. https://doi.org/10.1016/j.chb.2018.10.025

[3] David M. Blei and John D. Lafferty. 2009. *Topic Models.* Chapman and Hall/CRC. 71 – 93 pages.

[4] Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, Margaret E. Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. 2023. A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. Association for Computing Machinery, New York, NY, USA, 250–262. https://doi.org/10.1145/3565472.3595606

[5] Neus Capdeferro and Margarida Romero. 2012. Are online learners frustrated with collaborative learning experiences? *International Review of Research in Open and Distributed Learning* 13, 2 (2012), 26–44.

[6] Alexandra Gendreau Chakarov, Quentin Biddy, Colin Hennessy Elliott, and Mimi Recker. 2021. The Data Sensor Hub (DaSH): A Physical Computing System to Support Middle School Inquiry Science Instruction. *Sensors* 21, 18 (2021). https://www.mdpi.com/1424-8220/21/18/6243

[7] Michael Alan Chang, Thomas M. Philip, Arturo Cortez, Ashieda McKoy, Tamara Sumner, and William R. Penuel. 2022. Engaging Youth in Envisioning Artificial Intelligence in Classrooms: Lessons Learned.. In *Rapid Community Report Series.* Digital Promise and the International Society of the Learning Sciences.

[8] Pankaj Chejara, Luis P. Prieto, Maria Jesus Rodriguez-Triana, Reet Kasepalu, Adolfo Ruiz-Calleja, and Shashi Kant Shankar. 2023. How to Build More Generalizable Models for Collaboration Quality? Lessons Learned from Exploring Multi-Context Audio-Log Datasets Using Multimodal Learning Analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (, Arlington, TX, USA,) *(LAK2023)*. Association for Computing Machinery, New York, NY, USA, 111–121. https://doi.org/10.1145/3576050.3576144

[9] Sidney K. D'Mello, Nicholas D. Duran, Amanda Michaels, and Angela E.B. Stewart. In Press. Improving Collaborative Problem Solving Skills via Automated Feedback and Scaffolding: A Quasi-Experimental Study with CPSCoach 2.0. In *User Modeling and User-Adapted Interaction.*

[10] Vanessa Echeverria, Marisol Wong-Villacres, Xavier Ochoa, and Katherine Chiluiza. 2022. An Exploratory Evaluation of a Collaboration Feedback Report. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (Online, USA) *(LAK22)*. Association for Computing Machinery, New York, NY, USA, 478–484. https://doi.org/10.1145/3506860.3506890

[11] Colin Hennessy Elliott, Jessie Nixon, Jeffrey B. Bush, Alexandra Gendreau Chakarov, and Mimi Recker. 2021. "Do I need to know what I am doing if I am the teacher?" Developing teachers' debugging pedagogies with physical computing. *International Conference of the Learning Sciences* (2021). https://par.nsf.gov/biblio/10340643

[12] Mona Emara, Nicole M. Hutchins, Shuchi Grover, Caitlin Snyder, and Gautam Biswas. 2021. Examining Student Regulation of Collaborative, Computational, Problem-Solving Processes in Open-Ended Learning Environments. *Journal of Learning Analytics* 8, 1 (2021), 49 – 74.

[13] Fareeha Farooq, Farooq Rathore, and Sahibzada Mansoor. 2020. Challenges of Online Medical Education in Pakistan During COVID-19 Pandemic. *Journal of the College of Physicians and Surgeons–Pakistan : JCPSP* 30 (06 2020), 67–69. https://doi.org/10.29271/jcpsp.2020.Supp1.S67

[14] Stephen M. Fiore, Arthur Graesser, and Samuel Greiff. 2018. Collaborative problem-solving education for the twenty-first-century workforce. *Nature human behaviour* 2, 6 (2018), 367–369.

[15] Peter W. Foltz and Melanie J. Martin. 2008. *Automated Communication Analysis of Teams* (1 ed.). Routledge/Taylor and Francis Group. 411–431 pages.

[16] Arthur C. Graesser, Stephen M. Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W. Foltz, and Friedrich W. Hesse. 2018. Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest* 19, 2 (2018), 59–92. https://doi.org/10.1177/1529100618808244 arXiv:https://doi.org/10.1177/1529100618808244 PMID: 30497346.

[17] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[18] Angel Gurria. 2016. PISA 2015 results in focus. *PISA in Focus* 67 (2016), 1.

[19] John Hancock, Taghi M. Khoshgoftaar, and Justin M. Johnson. 2022. Informative Evaluation Metrics for Highly Imbalanced Big Data Classification. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. 1419–1426. https://doi.org/10.1109/ICMLA55696.2022.00224

[20] Jiangang Hao, Lei Chen, Michael Flor, Lei Liu, and Alina A. von Davier. 2017. CPS-Rater: Automated Sequential Annotation for Conversations in Collaborative Problem-Solving Activities. *ETS Research Report Series* 2017, 1 (2017), 1–9. https://doi.org/10.1002/ets2.12184 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ets2.12184

[21] Zellig S. Harris. 1954. Distributional Structure. 10 (1954), 146–162. Issue 2-3. https://doi.org/10.1080/00437956.1954.11659520

[22] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (2007), 81–112. https://doi.org/10.3102/003465430298487 arXiv:https://doi.org/10.3102/003465430298487

[23] Clare S. Howard, Kevin J. Munro, and Christopher J. Plack. 2010. Listening effort at signal-to-noise ratios that are typical of the school classroom. *International Journal of Audiology* 49, 12 (2010), 928–932. https://doi.org/10.3109/14992027.2010.520036 arXiv:https://doi.org/10.3109/14992027.2010.520036 PMID: 21047295.

[24] Heisawn Jeong, Cindy E. Hmelo-Silver, and Kihyun Jo. 2019. Ten years of Computer-Supported Collaborative Learning: A meta-analysis of CSCL in STEM education during 2005–2014. *Educational Research Review* 28 (2019), 100284. https://doi.org/10.1016/j.edurev.2019.100284

[25] Joni Lämsä, Pablo Uribe, Abelino Jiménez, Daniela Caballero, Raija Hämäläinen, and Roberto Araya. 2021. Deep Networks for Collaboration Analytics: Promoting Automatic Analysis of Face-to-Face Interaction in the Context of Inquiry-Based Learning. *Journal of Learning Analytics* 8, 1 (2021), 113 – 125. https://doi.org/10.18608/jla.2021.7118

[26] Arita L. Liu and John C. Nesbit. 2020. *Dashboards for Computer-Supported Collaborative Learning.* Springer International Publishing, Cham, 157–182. https://doi.org/10.1007/978-3-030-13743-4_9

[27] Stephen MacNeil, Kyle Kiefer, Brian Thompson, Dev Takle, and Celine Latulipe. 2019. IneqDetect: A Visual Analytics System to Detect Conversational Inequality and Support Reflection during Active Learning. In *Proceedings of the ACM Conference on Global Computing Education* (Chengdu,Sichuan, China) *(CompEd '19)*. Association for Computing Machinery, New York, NY, USA, 85–91. https://doi.org/10.1145/3300115.3309528

[28] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-Hoc Interpretability for Neural NLP: A Survey. *ACM Comput. Surv.* 55, 8, Article 155 (dec 2022), 42 pages. https://doi.org/10.1145/3546577

[29] Roberto Martinez-Maldonado, Judy Kay, Simon Buckingham Shum, and Kalina Yacef. 2019. Collocated Collaboration Analytics: Principles and Dilemmas for Mining Multimodal Interaction Data. *Human–Computer Interaction* 34, 1 (2019), 1–50. https://doi.org/10.1080/07370024.2017.1338956

arXiv:https://doi.org/10.1080/07370024.2017.1338956

[30] Robert G. Moulder, Nicholas D. Duran, and Sidney K. D'Mello. 2022. Assessing Multimodal Dynamics in Multi-Party Collaborative Interactions with Multi-Level Vector Autoregression. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 615–625. https://doi.org/10.1145/3536221.3556595

[31] Sambit Praharaj, Maren Scheffel, Marcel Schmitz, Marcus Specht, and Hendrik Drachsler. 2022. Towards Collaborative Convergence: Quantifying Collaboration Quality with Automated Co-Located Collaboration Analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (Online, USA) *(LAK22)*. Association for Computing Machinery, New York, NY, USA, 358–369. https://doi.org/10.1145/3506860.3506922

[32] Samuel L. Pugh, Arjun Rao, Angela E.B. Stewart, and Sidney K. D'Mello. 2022. Do Speech-Based Collaboration Analytics Generalize Across Task Contexts?. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (Online, USA) *(LAK22)*. Association for Computing Machinery, New York, NY, USA, 208–218. https://doi.org/10.1145/3506860.3506894

[33] Samuel L. Pugh, Shree Krishna Subburaj, Arjun Ramesh Rao, Angela E.B. Stewart, Jessica Andrews-Todd, and Sidney K. D'Mello. 2021. Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild. In *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021)* (Online, USA) *(EDM21)*. 55–67.

[34] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. *Robust speech recognition via large-scale weak supervision.* Technical Report. OpenAI.

[35] Dani Ramdani, Herawati Susilo, Suhadi Suhadi, and Sueb Sueb. 2022. The Effectiveness of Collaborative Learning on Critical Thinking, Creative Thinking, and Metacognitive Skill Ability: Meta-Analysis on Biological Learning. *European Journal of Educational Research* 11, 3 (2022), 1607–1628.

[36] Meeli Rannastu-Avalos and Leo Aleksander Siiman. 2020. Challenges for distance learning and online collaboration in the time of COVID-19: Interviews with science teachers. In *Collaboration Technologies and Social Computing: 26th International Conference, CollabTech 2020, Tartu, Estonia, September 8–11, 2020, Proceedings 26*. Springer, 128–142.

[37] Jason G. Reitman, Charis Clevenger, Quinton Beck-White, Amanda Howard, Sierra Rose, Jacob Elick, Julianna Harris, Peter Foltz, and Sidney K. D'Mello. 2023. A Multi-theoretic Analysis of Collaborative Discourse: A Step Towards AI-Facilitated Student Collaborations. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Springer Nature Switzerland, 577–589.

[38] Jeremy Roschelle, Yannis Dimitriadis, and Ulrich Hoppe. 2013. Classroom orchestration: Synthesis. *Computers and Education* 69 (2013), 523–526. https://doi.org/10.1016/j.compedu.2013.04.010

[39] Carolyn Penstein Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning* 3 (2008), 237–271. https://api.semanticscholar.org/CorpusID:1438702

[40] Bertrand Schneider, Nia Dowell, and Kate Thompson. 2021. Collaboration Analytics — Current State and Potential Futures. *Journal of Learning Analytics* 8, 1 (April 2021), 1–12. https://doi.org/10.18608/jla.2021.7447

[41] Arabella J. Sinclair and Bertrand Schneider. 2021. Linguistic and Gestural Coordination: Do Learners Converge in Collaborative Dialogue?. In *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021)* (Online, USA) *(EDM21)*. 431 – 438.

[42] Rosy Southwell, Samuel Pugh, E. Margaret Perkoff, Charis Clevenger, Jeffrey Bush, Rachel Lieber, Wayne Ward, Peter Foltz, and Sidney D'Mello. 2022. Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms. In *Proceedings of the 15th International Conference on Educational Data Mining*, Antonija Mitrovic and Nigel Bosch (Eds.). International Educational Data Mining Society, Durham, United Kingdom, 302–315. https://doi.org/10.5281/zenodo.6853109

[43] Angela E.B. Stewart, Mary Jean Amon, Nicholas D. Duran, and Sidney K. D'Mello. 2020. Beyond Team Makeup: Diversity in Teams Predicts Valued Outcomes in Computer-Mediated Collaborations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376279

[44] Angela E.B. Stewart, Zachary Keirn, and Sidney K. D'Mello. 2021. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction* 31, 4 (2021), 713–751. https://doi.org/10.1007/s11257-021-09290-y

[45] Angela E.B. Stewart, Zachary A. Keirn, and Sidney K. D'Mello. 2018. Multimodal Modeling of Coordination and Coregulation Patterns in Speech Rate during Triadic Collaborative Problem Solving. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/3242969.3242989

[46] Angela E.B. Stewart, Hana Vrzakova, Chen Sun, Jade Yonehiro, Cathlyn Adele Stone, Nicholas D. Duran, Valerie Shute, and Sidney K. D'Mello. 2019. I Say, You Say, We Say: Using Spoken Language to Model Socio-Cognitive Processes during Computer-Supported Collaborative Problem Solving. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 194 (nov 2019), 19 pages. https://doi.org/10.1145/3359296

[47] Angela E. B. Stewart, Arjun Rao, Amanda Michaels, Chen Sun, Nicholas D. Duran, Valerie J. Shute, and Sidney K. D'Mello. 2023. CPSCoach: The Design and Implementation of Intelligent Collaborative Problem Solving Feedback. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Springer Nature Switzerland, Cham, 695–700.

[48] Chen Sun, Valerie J. Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers and Education* 143 (2020), 103672. https://doi.org/10.1016/j.compedu.2019.103672

[49] Chen Sun, Valerie J. Shute, Angela E.B. Stewart, Quinton Beck-White, Caroline R. Reinhardt, Guojing Zhou, Nicholas Duran, and Sidney K. D'Mello. 2022. The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Computers in Human Behavior* 128 (2022), 107120. https://doi.org/10.1016/j.chb.2021.107120

[50] Mike Tissenbaum and James D. Slotta. 2015. *Scripting and Orchestration of Learning Across Contexts: A Role for Intelligent Agents and Data Mining.* Springer Singapore, Singapore, 223–257. https://doi.org/10.1007/978-981-287-113-8_12

[51] Sophie Vayssettes et al. 2016. *PISA 2015 assessment and analytical framework: science, reading, mathematic and financial literacy.* OECD publishing.

[52] Hana Vrzakova, Mary Jean Amon, Angela E. B. Stewart, Nicholas D. Duran, and Sidney K. DMello. 2020. Focused or Stuck Together: Multimodal Patterns Reveal Triads' Performance in Collaborative Problem Solving. In *Proceedings of the Tenth International Conference on Learning Analytics and Knowledge* (Frankfurt, Germany) *(LAK '20)*. Association for Computing Machinery, New York, NY, USA, 295–304.

[53] Noreen M. Webb, Megan L. Franke, Nicholas C. Johnson, Marsha Ing, and Joy Zimmerman. 2023. Learning through explaining and engaging with others' mathematical ideas. *Mathematical Thinking and Learning* 25, 4 (2023), 438–464. https://doi.org/10.1080/10986065.2021.1990744 arXiv:https://doi.org/10.1080/10986065.2021.1990744

[54] Guojing Zhou, Robert Moulder, Chen Sun, and Sidney D'Mello. 2022. Investigating Temporal Dynamics Underlying Successful Collaborative Problem Solving Behaviors with Multilevel Vector Autoregression. In *Proceedings of the 15th International Conference on Educational Data Mining*, Antonija Mitrovic and Nigel Bosch (Eds.). International Educational Data Mining Society, Durham, United Kingdom, 290–301.

accepted 18 Dec 2023