Empowering Teacher Learning with AI: Automated Evaluation of Teacher Attention to Student Ideas during Argumentation-focused Discussion

Tanya Nazaretsky*
tanya.nazaretsky@weizmann.ac.il
Weizmann Institute of Science
Rehovot, Israel

Jamie N. Mikeska
jmikeska@ets.org
Educational Testing Service
Princeton, NJ, USA

Beata Beigman Klebanov
bbeigmanklebanov@ets.org
Educational Testing Service
Princeton, NJ, USA

ABSTRACT

Engaging students in argument from evidence is an essential goal of science education. This is a complex skill to develop; recent research in science education proposed the use of simulated classrooms to facilitate the practice of the skill. We use data from one such simulated environment to explore whether automated analysis of the transcripts of the teacher's interaction with the simulated students using Natural Language Processing techniques could yield an accurate evaluation of the teacher's performance. We are especially interested in explainable models that could also support formative feedback. The results are encouraging: Not only can the models score the transcript as well as humans can, but they can also provide justifications for the scores comparable to those provided by human raters.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**; **Natural language processing**; • **Human-centered computing**; • **Applied computing** → **Computer-assisted instruction**;

KEYWORDS

Teacher Discourse, Automated Feedback, Deep Learning, Simulated Teaching, Practice-based Teacher Education,

ACM Reference Format:
Tanya Nazaretsky, Jamie N. Mikeska, and Beata Beigman Klebanov. 2023. Empowering Teacher Learning with AI: Automated Evaluation of Teacher Attention to Student Ideas during Argumentation-focused Discussion. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3576050.3576067

1 INTRODUCTION

Engaging students in argument from evidence is an essential goal of science education. However, effective facilitation of argumentation-focused discussion is a complex, multifaceted activity. It requires a teacher to master leading a coherent conversation while attending to student ideas, encouraging student-to-student communication, and developing students' conceptual understanding of natural phenomena. Previous research on practice-based teacher education suggests that teachers need to experience approximations of practice to master facilitating argumentation-focused discussions and that feedback is essential to support their learning from these approximations. However, providing meaningful feedback is a challenging task requiring substantial human expertise. It is also time-consuming and hard to scale across many teachers. Advances in Artificial Intelligence (AI) and Natural Language Processing (NLP) have great potential for analyzing instructional discourse and providing substantive feedback.

This research is part of our ongoing effort to explore applications of AI and NLP to support teachers in engaging students in productive scientific argumentation. We used transcriptions of 157 discussions conducted in a simulated classroom with five elementary student avatars. The teachers aimed to engage these students in a meaningful argumentation-focused science discussion. Human raters scored the discussions based on a holistic rubric that focused on teachers' ability to elicit significant contributions from all students, to attend to student ideas equitably, and to use student ideas to move the discussion forward. The raters were also required to justify their scores through students' and teachers' discussion moves. We employed a supervised machine learning approach and pre-trained state-of-the-art Transformer models (BERT) to predict holistic scores and identify utterances that could serve as score justifications.

The main contributions of this research are answers to the following two research questions:

- **RQ1**: Can AI-powered models accurately score an elementary teacher's ability to attend to student ideas in a simulated argumentation-focused discussion?
- **RQ2**: Can AI-powered models support automated personalized feedback to the teachers to help them improve in attending to students' ideas?

The rest of this paper is organized as follows. We first give an overview of the related work (Section 2). Next, in Sections 3 and 4 we describe the simulated environment used for data collection and the collected data. In Section 5, we explain the experimental setup. Then, we present the methodology, analysis, and findings for the two research questions (Sections 6 and 7, respectively). Section 8 provides a discussion, followed by the summary and conclusions in Section 9.

*TN was the Educational Testing Service consultant at the time of the research.

LAK 2023, March 13–17, 2023, Arlington, TX, USA

ACM ISBN 978-1-4503-9865-7/23/03…$15.00
https://doi.org/10.1145/3576050.3576067

2 RELATED WORK

2.1 Argumentation in the teaching of science

Engaging students in argument from evidence helps to develop their scientific thinking skills and understanding of scientific phenomena and how scientific knowledge is constructed [2, 17, 37]. Productively engaging in scientific argumentation requires students to participate in both argument construction and argument critique [6]. Argument construction involves students generating and refining scientific claims and evidence-based reasoning, while argument critique focuses on comparing arguments, offering rebuttals, and considering counterarguments. Research has shown that supportive classroom environments where students have frequent opportunities to engage in argument construction and critique are essential for building K-12 students' argumentation skills [5, 47].

However, previous research has shown that learning how to create and maintain classroom environments that support students' engagement in scientific argumentation can be challenging for teachers [32, 40]. Teachers in K-12 settings need opportunities themselves to learn how to engage their students in scientific argumentation. Research in science education across studies has shown positive outcomes in terms of using varied interventions to develop teachers' argumentation skills, knowledge, and perceptions [26, 46]. However, changes to teachers' actual instructional practice have been harder to achieve [14, 18], albeit a few studies have shown promise in this area too [15, 39]. One promising approach is using simulated classrooms to serve as practice spaces where teachers can try out new instructional skills, but in situations of reduced complexity and with no potential for risk to actual students [27].

2.2 Simulated environments for teacher education

During the last decade, approximations of practice have been growing in use within teacher education and professional development settings to provide opportunities for teachers to learn how to engage in core teaching practices, like facilitating argumentation-focused discussions, eliciting and using student ideas, and analyzing student thinking, with some approximations using simulated classrooms to support teacher learning [3, 9, 12, 16, 21]. To support teacher learning using approximations of practice, including those with simulations, research has suggested the critical importance of cycles of preparation for, engagement in, and reflection on the simulated teaching experiences [4, 30]. However, the existing simulation platforms typically do not provide any real-time formative feedback to support teacher reflection and improvement. The feedback still comes from a coach or through written feedback provided by human raters after reviewing video records from the simulated teaching sessions [7] and it is hard to scale across many teachers [38].

2.3 Automated evaluation of classroom discourse

Advances in Artificial Intelligence (AI) and Natural Language Processing (NLP) have great potential for analyzing instructional discourse and providing substantive feedback to support teacher learning. Possible applications include identifying different types of classroom activities [13, 43, 44, 52] and providing automated feedback on various teacher discourse moves, such as moves designed to (a) guide discussion and ensure students' participation [20, 50], (b) ask authentic questions; namely, questions for which the answers are not presupposed by the teacher [1, 22] or (c) restate and use student ideas [10]. Traditional approaches to automated analysis of classroom discourse typically employ supervised machine learning (ML) methods combined with manual feature engineering and human expert annotation of collected data according to evaluation rubrics. As so, these methods usually require the collection of substantial corpora of annotated datasets, which is challenging and both time- and cost-expensive.

Recent advances in NLP (transformer deep-learning architecture) have introduced a new method of *few shot learning*, which refers to the practice of fine-tuning ML models on a very small amount of annotated data [51] and utilizing the power of state-of-art ready-for-use language models (e.g., Google's BERT model and its variations [11, 48]) pre-trained on huge amounts of textual data using substantial computational power. For example, this method was recently applied to identify intent in dialog data [19], model teacher discourse in classrooms [20], and provide teachers with personalized feedback on their classroom discourse [49] and has been shown to outperform traditional ML methods while using relatively small datasets. In this research, we employ the few shot learning approach.

3 A SIMULATED CLASSROOM FOR EDUCATORS

The data used in this study was collected as part of two previously funded National Science Foundation research projects. In each project, elementary science teacher educators integrated the use of simulated teaching experiences into their science methods courses with one class of elementary preservice teachers [28, 33]. To examine whether using the simulated teaching experiences developed the elementary preservice teachers' ability to facilitate argumentation-focused discussions, we used one performance task – the Mystery Powder science task – as a pre/post measure at the beginning and end of the semester. In the Mystery Powder science task, each preservice teacher had an opportunity to lead an argumentation-focused discussion with five upper elementary student avatars in a simulated classroom (see Figure 1). The goal of the discussion was for the preservice teachers to support the students in coming to a consensus about what the mystery powder is and which properties were most useful to make this determination [29]. Prior to leading the discussion in the simulated classroom, each preservice teacher received a written packet that provided them with: (a) information about the discussion's goal, (b) details about what the student avatars did in class prior to this discussion (including the Mystery Powder science investigation they completed), and (c) access to written copies of the students' written work indicating their initial ideas about the mystery powder's identity and the properties the students thought were useful to identifying the mystery powder. Each preservice teacher used the Mystery Powder task packet to plan and then lead an up to 20-minute discussion with the five student avatars in the simulated classroom. Each discussion was video recorded and then transcribed for analysis purposes;

in the transcription, each utterance was time-stamped and identified by the speaker (either the preservice teacher or one of the five student avatars). In an earlier study, findings indicated that the preservice teachers had significant growth in their ability to facilitate discussions from the beginning to the end of the semester across five dimensions of this practice [31].

Figure 1: Teacher Interaction with Upper Elementary Student Avatars in the Simulated Classroom. Image courtesy of Mursion, Inc.

4 DATASETS

4.1 Data

We used transcripts of 157 discussions conducted in a simulated classroom with five elementary student avatars. 81 teachers participated in the study. All teachers (except five) led two simulated sessions each. The data was collected in two previous studies: S1 (n = 88 transcripts) and S2 (n = 69 transcripts). [31, 33]. The distributions of the number of utterances per transcript and number of words per utterance for each discussion participant (Carlos, Emily, Jayla, Mina, and Will are the student avatars) are presented in Table 1.

	Utterances per transcript		Words per utterance	
Participant	median	25-75% range	median	25-75% range
Carlos	13	10-16	12	5-25
Emily	6	4-9	9	3-19
Jayla	10	7-13	9	3-20
Mina	11	8-14	9	3-17
Will	9	6-12	8	4-17
Teacher	42	34-51	16	9-30
Total	94	79-111	12	6-24

Table 1: The distributions of the number of utterances per transcript and number of words per utterance per each discussion participant

4.2 Performance Rubric

One aspect of the performance rubric was focused on teachers' ability to attend to student ideas equitably [45]. It involved being responsive to students and focused on making sure that the discussion was grounded in students' ideas and that all students were engaged in a meaningful component of the discussion. It included three categories: (a) all the key ideas that appear in the students' written work are incorporated into the discussion; (b) all student voices are heard in some non-trivial way, and (c) each of the relevant student ideas is attended by the teacher and made a part of the discussion. These categories are hereafter referred to as **Indicator** *a*, **Indicator** *b*, and **Indicator** *c*, correspondingly.

4.3 Dataset Annotation and Inter-Rater Agreement

To recruit raters, research project team members on the two previous projects reached out to current and retired K-12 teachers through their professional networks. Interested raters with expertise in STEM teaching were offered positions as scorers on the previous projects. Each rater engaged in extensive scoring training to learn about the five dimensions of this teaching practice – one of those dimensions (attending to student ideas) is the focus of this study. Raters watched webinars to learn about the scoring indicators in each dimension and the evidence used to determine the scoring level for each indicator and dimension. In addition, each rater had opportunities to practice scoring using video clips and full discussion videos, as well as received feedback on their scoring prior to starting their scoring assignments. Each rater was required to provide the following:

(1) **Total score.** The S1 project was graded using an ordinal scale with 3 levels (Beginning, Developing and Well-Prepared). For the S2 project, the third level was divided into 2 sub-levels (Well-Prepared and Commendable). The distribution of total scores is presented in Figure 2. We collapsed the two top levels in S2 data; hence all transcripts are rated on a 3-level scale in our experiments.
(2) **Indicator *a*, *b* and *c* scores.** The S1 project was graded using an interval scale from 1 to 3. Similarly to the total score, the highest level was extended for project S2 to the range from 1 to 4. These scores were used as intermediate auxiliary aids for the raters to decide on the total score. See Figure 2 for score distributions.
(3) **Indicator *b* justifications.** The indicator *b* justifications included student utterances that exemplify their substantial contribution to the discussion. Usually, each rater gave one or two examples for each student, such as: "Carlos (09:34): Yeah. So weight is not an important property to identify the mystery powder. It just tells you how much of something you have." Some raters did not provide particular utterance examples while mentioning that a student provided substantial contributions. Raters were asked to write "none" if a student did not significantly contribute to the discussion.
(4) **Indicator *c* justifications.** The indicator *c* justifications included teacher utterances to exemplify teacher usage of student ideas to move the discussion forward. The raters were instructed to briefly describe which student ideas the teacher made use

of during the discussion and how those ideas were used in a way that helped to understand the indicator c score they assigned. In particular, the raters were asked to provide specific examples, quotes, and/or details to justify their indicator score. Usually, each rater gave from one to three separate examples for a teacher, such as: "The teacher frequently used student ideas and responses to move the discussion forward and consistently withheld her own ideas. For example, she noted that Carlos did not need to use weight and said, "And you didn't need to weigh it? So, why didn't you need to weigh it? Can you explain to your classmates?" She later asked the class to explain to Carlos why they thought weight was an important property. Some raters did not provide utterance examples while mentioning that the teacher effectively used student ideas in a meaningful way. In other cases, raters decided that the teacher missed opportunities to use student ideas, and they indicated this by writing "none."

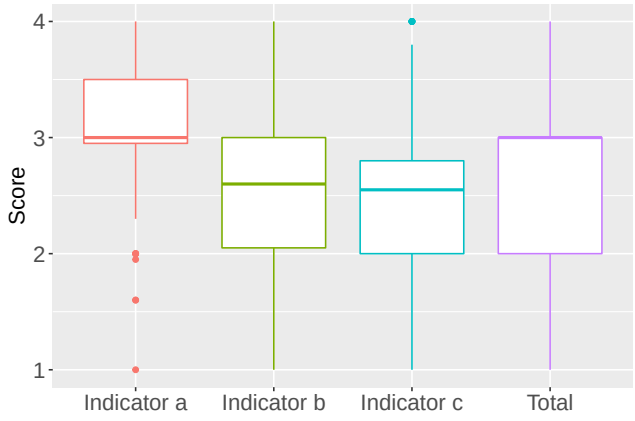

Figure 2: The distributions of evaluation rubric scores: Indicators a, b, c, and total score.

The 157 transcripts were rated in the following way:

- Two transcripts were rated by twelve human raters each. These two transcripts are hereafter referred to as multi-rated transcripts T1 and T2.
- 56 transcripts were rated by two human raters each.
- The rest of the transcripts ($n = 99$) were rated by one rater each.

The inter-rater agreement for the total score and indicators a, b, and c scores was evaluated using 56 double-rated transcripts. The inter-rater Pearson correlation between the total scores was $r = 0.379$, indicating fair level of agreement between the raters [8]. The inter-rater correlations for indicators a, b, and c were $r = 0.127$, $r = 0.516$, and $r = 0.527$, respectively. The latter two values indicate moderate agreement, as suggested in [8]. The low agreement for indicator a guided our decision to not use indicator a in further analyses. In addition, we measured the correlation between the total score and indicator b and c score. The correlations were $r = 0.812$ and $r = 0.835$, correspondingly, indicating a clear alignment of both the intermediate scores with the total score.

The evaluation of inter-rater agreement on justifications provided for indicators b and c was challenging due to the human variation in the selection of examples. The data collected from the 56 double-rated transcripts was not rich enough, as in most cases the two raters chose different examples from the same transcript. Asking people to select a few examples of good engagement with student ideas from a dozen or more utterances per participant is somewhat akin to asking people to pick important information in a document for a summary. In both cases, there are many valid solutions, although one would hope that instances that are *clearly* very good examples or important would be picked more often. These observations suggested the idea of using utterance justification examples collected from multiple human raters (in our case, twelve raters) who rated the same two transcripts and evaluating them using the Pyramid score [36] originally developed for evaluating summaries. We describe the Pyramid evaluation process and the human-human and human-machine agreement result in Sections 5.2.3 and 7 below.

4.4 Pre-processing of the Dataset for Machine Learning

We created four datasets to separately measure our method's ability (1) to predict the total score and the scores of rubric indicators; (2) to identify if a student utterance is an example of a substantial student contribution to the discussion; and (3) to identify if a teacher utterance is an example of a teacher making a meaningful usage of student ideas to move the discussion forward. Below we describe the datasets and corresponding labels in more detail.

- **Dataset DS1.** Unlabeled dataset ($n = 20,813$) of all student and teacher utterances automatically collected by parsing 157 transcript MS Documents. There are 9,833 teacher utterances and 10,974 student utterances.
- **Dataset DS2.** Utterance-level labeled dataset (a subset of DS1, $n = 1,613$) with binary labels indicating if an utterance was chosen by at least one of the raters as an example of a student making a substantial contribution to the discussion (class 1). The examples were collected semi-manually[1] from raters' indicator b justifications. If a rater indicated that a student did not contribute substantially to the discussion by writing "none" for justification, all the student utterances were inserted into the dataset as negative examples (class 0). In total, 1,004 class 1 and 609 class 0 examples were collected. So, only about 15% of all student utterances ($n = 10,974$) were labeled as class 1 or class 0 for indicator b.
- **Dataset DS3.** Utterance-level labeled dataset (a subset of DS1, $n = 693$) with binary labels indicating if an utterance was used by at least one of the raters as an example of a teacher using a student's idea to move the discussion forward (class 1). The examples were collected semi-manually[1] from raters' indicator c justifications. In case a rater indicated that a teacher did not make use of student ideas to move the discussion forward by writing "none" under "Indicator c justifications", all the utterances of the teacher were inserted into the dataset as class 0 instances. In total, 330 class 1 and 363 class 0 instances were collected. So, only about 7% of all teacher utterances ($n = 9,833$) were labeled as class 1 or class 0 for indicator c.

[1]Sometimes parts of an utterance were cited by raters, so automated alignment of examples to transcript utterances had to be manually adjusted.

- **Dataset DS4.** Transcript-level labeled dataset (n = 157) with 3 labels: one for the total score, and two for indicators b and c, respectively. For the 56 double-scored transcripts, we averaged the scores of the two raters.

5 EXPERIMENTAL SETUP

5.1 Data Partitioning

As described in Subsection 4.3, most teachers contributed two transcripts to the study. To make sure that models do not pick up on teacher-specific linguistic idiosyncrasies, we divided all the datasets used to address RQ1 into training and test sets (80% and 20% respectively) randomly by the teacher rather than by transcript.

To address RQ2, for a more comprehensive evaluation of utterance-level predictions on DS2 and DS3 we utilized the two transcripts evaluated by twelve raters. So, all the student and teacher utterances originating from these two transcripts comprised the test sets for datasets DS2 (n = 95) and DS3 (n = 88). The rest of the labeled student and teacher utterances were assigned to the corresponding training sets. There were no transcripts from the two teachers who contributed the test transcripts in the training data. This approach allowed us to utilize "the wisdom of the crowd" for evaluating predictions of our utterance-level models using the Pyramid method [36] designed to evaluate summaries against multiple partially overlapping model summaries. The method is described in more detail in Subsection 5.2.3 below.

5.2 Machine Learning Methods

5.2.1 BERT. We employed a supervised machine learning approach and pre-trained state-of-art deep neural network BERT models [11, 19] to classify utterances (into good examples of indicator scores or not) and generate utterance level embeddings. The BERT architecture was implemented using PyTorch API [41] with TensorFlow [2] backend. Namely, the Hugging Face DistilBERT base model (uncased) [3] [48] was used: (i) for utterance level embeddings generation and (ii) for text classification. In the latter case, we fine-tuned the pre-trained models end-to-end (including all transformer layers, the pooling layer, and the final dense output layer) with the Adam optimizer [23] (learning rate = 1e-5, learning warmup = 600) to minimize the binary cross-entropy loss, which is consistent with typical BERT fine-tuning for text classification [19]. The models were trained with batch size 1 for 5 epochs.

As a typical utterance in our dataset was quite short (median = 12, 25-75% range = 6-24, see Table 1), we decided to enrich BERT input by using the previous utterance concatenated with the name of the speaker of the previous utterance as a context, since the previous utterance most probably contains information important for classification of indicators b and c justifications. In rare cases, if the input was longer than 512 tokens it was truncated to fit the corresponding BERT model restriction.

5.2.2 Regression methods. In the context of RQ1, for total score and indicator b and c scores we used several shallow machine learning architectures, namely Multi-layer Perceptron Regressor (MLP), Linear Regression (LR), Decision Tree Regressor (DT) and Bayesian Ridge Regressor (BR) implemented using scikit-learn [42] Python package [4]. Hyper-parameters of the aforementioned regression models were fine-tuned using 5-folds cross-validation grid search over the parameters relevant to each type of the regressor[5] and the best performing model was chosen for each regression type separately. Then, we compared the mean cross-validation Mean Squared Error (MSE) values of the best MLP, LR, DT, and BR regressor and chose the winning model. Next, we retrained the best regressor over the entire training set.

5.2.3 Evaluation Measures. For RQ1, we used Cohen's Kappa, F1, and Accuracy on validation data for preliminary evaluation of binary models for predicting indicators b and c justifications on utterance level. For evaluating predictions of the total score and indicators b and c scores we used MSE and Pearson correlation (r).

For RQ2, we used a **Pyramid score** for a more comprehensive evaluation of utterance-level predictions of indicators b and c justifications. The Pyramid method can be applied if multiple ratings of the same unit of analysis (in our case, an utterance) are available and it assigns higher weights to utterances used as justifications by more human raters [36].

Applying the Pyramid method includes several steps. First, each utterance is assigned a weight corresponding to the number of raters who used this utterance as a justification. Second, the pyramid is created for each transcript and each indicator separately. The pyramid consists of the number of tiers equal to the maximum weight given in the first step, and each utterance is assigned to the n^{th} tier if it was chosen by n raters. There are usually a few examples with high weights (placed at the top of a pyramid) and a large number of examples used by only one rater (placed at the base of a pyramid) [36], creating the pyramid form. The pyramids created based on the two multi-rated transcripts are presented in Figure 3 (the letters in the circles are a variation we introduced to the 'vanilla' Pyramid method to account for the selection of utterances per participant and will be explained later). Third, to evaluate a new rating, it is compared to the pyramid as follows. Let k be the number of utterances selected as possible justifications by the new rater. The raw score of the new rating is the sum of tiers to which the k utterances belong according to the pyramid; utterances not in the pyramid are assigned zero weight. For example, for the 10-tier pyramid in Figure 3, disregarding the letters, if the new rating has k=4 and picked the two utterances at tier 9, the single utterance at tier 6, and an utterance that is not in the pyramid, its raw pyramid score is 9 + 9 + 6 + 0 = 24. Next, the raw score of the best possible k-utterance selection is computed; in our example, it is 10 + 9 + 9 + 8 = 36. Finally, the Pyramid score of the new rating is calculated as the ratio between its raw score and the best possible score for a k-utterance selection. It is 24/36 = 0.67 in the example. So, the Pyramid score is always between 0 and 1 and higher values indicate higher agreement of the new rater with the pyramid.

In this study, due to the nature of the evaluation rubric that explicitly asked raters to choose justifications for each discussion participant separately, it was important to take into account who

[2] https://www.tensorflow.org/
[3] https://huggingface.co/docs/transformers/model_doc/distilbert
[4] https://scikit-learn.org/stable/index.html
[5] MPL hyperparameters: number of hidden layers, activation function, alpha, DT hyperparameters: max depth, max number of features, and BR hyperparameters: alpha, lambda, tol

delivered each utterance inserted into the pyramid (the letters inside the circles in Figure 3) and consider that while calculating the Pyramid score. Let k be the number of utterances selected as possible justifications by the new rater. For the 10-tier pyramid in Figure 3, taking into account the letters, if the new rating has $k = 4$ and picked the two utterances of Carlos (C) and Mina (M) at tier 9, the single Emily's utterance at tier 6 (E), and Will's utterance that is not in the pyramid, its raw score is $9+9+6+0 = 24$ (same as before). However, the score of the best possible k-utterance selection is different as we now consider the participants. In our example, the utterances of Carlos (at tier 9), Mina (at tier 9), Emily (at tier 8), and Will (at tier 3) are the best possible scores for these participants, which is $9+9+8+3 = 29$, yielding the Pyramid score of $24/29 = 0.83$.

6 RQ1: EXPERIMENTS AND RESULTS

The goal of these experiments is to build and evaluate models for the prediction of total and indicator scores per transcript. We investigate two approaches. The first, *text-to-score*, utilizes solely the text of the transcript for the prediction. The rationale for this method is the track record of the strong performance of out-of-the-box pre-trained language models on a variety of meaning-focused tasks. The second, *evidence-to-score*, first identifies teacher and student utterances that could serve as justifications, or evidence, for indicator scores, then combines the justifications in accordance with the indicator scoring rubric to derive the score. The rationale for this method is two-fold: (a) it uses additional information – justifications provided by raters for their scores, and (b) its scoring mechanism mimics the human rater scoring rubric.

Our experiment design is presented in Figure 4; it consists of four stages.

(1) **Utterance-level BERT classifier fine-tuning.** Train BERT classification models for predicting indicators b and c justifications; these would serve to identify evidence for the evidence-to-score models.
(2) **Utterance-level BERT models prediction.** Generate utterance-level features for the entire DS1. Under the text-to-score design, we generate a 768-dimensional embedding (namely, a contextualized meaning representation) for every student and teacher utterance using the pre-trained BERT. Under the evidence-to-score design, we generate a binary prediction of whether the utterance can serve as evidence of attending to student ideas using the BERT classification models from Stage 1. Labels for student utterances are generated using indicator b BERT classifier, and labels for teacher utterances are generated using indicator c BERT classifier.
(3) **Feature creation from utterance-level to transcript level.** Combine utterance-level features into meaningful transcript-level features. Under the text-to-score design, we average embeddings for all utterances per participant per transcript. Under the evidence-to-score design, we attempted to mimic the evaluation rubric and generated features that would capture, for example, whether there is at least one piece of evidence for each of the students.
(4) **Transcript-level model training and model selection.** Use the transcript-level features crafted in Stage 3 to predict the total score for the transcript. We used several regression models (MLP, LR, DT, and BR, hereafter referred to as Regressors) in two variants of machine learning architecture: (i) direct prediction of the total score from the transcript-level features; and (ii) two-steps prediction: first prediction of indicator b and indicator c scores, followed by prediction of total scores based on the indicators' scores. These approaches were evaluated for both text-to-score and evidence-to-score models.

Below, we present each stage in more detail.

6.1 Stage 1: Utterance-level BERT classifier fine-tuning

Indicator b justifications: We fine-tuned pre-trained BERT classification model end-to-end [19] on the training set of dataset DS2. We performed a small hyper-parameter search over the number of epochs (max epochs = 5) using 7-fold cross-validation and measured the mean values of Cohen's Kappa, F1, and Accuracy measures over a validation set, which led to 7-best performing models trained over 5 epochs.

Indicator c justifications: The same procedure was utilized to train the best predicting models for indicator c over DS3 dataset over 4 epochs.

Validation-set evaluations were conducted to get a sense of the model's ability to identify utterances with substantial contribution (indicator b) and those where the teacher is making use of a student's idea (indicator c). We present results for each of the indicators in Table 2. Results appear sufficiently promising to attempt to use the utterance-level predictions in transcript-level models, as described in the next section. We defer a more comprehensive evaluation of the utterance-level models to the discussion of RQ2.

6.2 Stage 2: Utterance-level BERT models prediction

For the evidence-to-score design, the predictions for indicators b and c were calculated as soft voting (the class with the highest average probability) over the predictions of the corresponding seven cross-validation models. The models predict a binary label for each utterance. Using the indicator b models, we predict one of "a substantial contribution" or "not a substantial contribution" for each student utterance. Using the indicator c models, we predict one of "effective use of a student's idea" or "not an effective use of a student's idea" for each teacher utterance. For the text-to-score design, pre-trained BERT models were used as-is to generate the embeddings (dimension size = 768) for each utterance. Both types of predictions were generated for the entire set of unlabeled utterances DS1.

6.3 Stage 3: Feature creation from utterance-level to transcript level

(1) **Features based on pre-trained embeddings.** To create features at the transcript level for the text-to-score models, we calculated the average values over the embedding of each participant (5 students and a teacher) separately. The resulting dataset contained 6×768 features for each transcript.

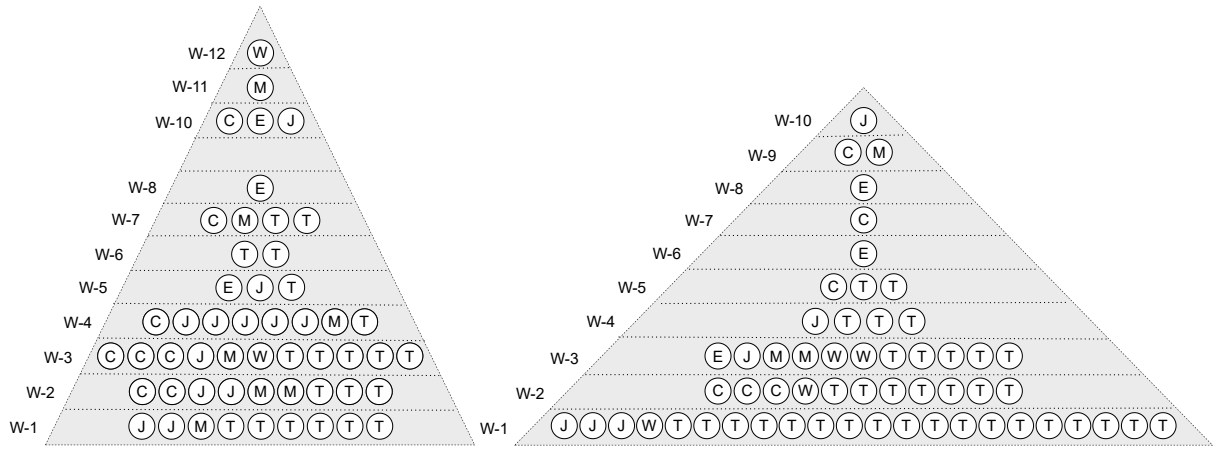

Figure 3: Pyramids for the two multi-transcripts used to evaluate the classifiers that predict whether the given utterance can serve as a justification for the indicator score. Each circle represents an utterance picked as a justification by at least one of the 12 human raters; tier placement is described in the text. The letters denote the first name initials of the five simulated students (Jayla, Will, Emily, Carlos, and Mina) and the teacher (T). Pyramids for indicators *b* and *c* are merged for ease of visualization; the sub-pyramid with just students' utterances is used for indicator *b* and the sub-pyramid with the teacher's utterances is used for indicator *c*.

Dataset	Justification of indicator	Optimal epochs	Accuracy	F1	Kappa
DS2	*b*	5	0.848 ± 0.024	0.845 ± 0.025	0.660 ± 0.057
DS3	*c*	4	0.751 ± 0.039	0.751 ± 0.040	0.503 ± 0.077

Table 2: Validation-set evaluations (7-fold cross validation) for indicator *b* and indicator *c* justification predictions.

(2) **Features based on predictions of the utterance-level models.** For the evidence-to-score models, we used the binary utterance-level predictions from Stage 2 to calculate the total number of utterances per participant and transcript and the number of class 1 labeled utterances ("substantial contribution" for students and "effective use of a student's idea" for the teacher) per participant and transcript. Based on these numbers, we calculated the following 7 features:
- **Student's substantial contribution** (5 features). A binary feature per student with the value set to 1 if the student has at least two class 1 indicator *b* utterances, otherwise it was set to 0.
- **Teacher's ability to elicit substantial contribution from the students** (1 feature). This feature has 3 ordinal levels. Level 3 was assigned if a teacher succeeded in eliciting substantial contribution from all of the students; level 2 – from at least three students; level 1 was assigned in all other cases.
- **Teacher's effective usage of student ideas** (1 feature). A binary teacher feature with the value set to 1 if at least 40% of teacher utterances per transcript were labeled as class 1 for indicator *c*.

These features and the rules for value assignment were based on the corresponding evaluation rubric instructions. For example, the rubric instruction mentioned that the teacher should have elicited substantial contribution from all the students to get a score of 3 on the holistic indicator-level rubric (see item 2 in Subsection 4.3).

6.4 Stage 4: Transcript-level model training and selection of the best model

6.4.1 Experimental design. In this study we used two types of models: (i) models to predict the total score directly, and (ii) models to predict the intermediate scores (indicators *b* and *c* scores) and use these scores as features to predict the total score. The latter was inspired by the very high correlation between human intermediate scores (indicators *b* and *c*) with the total score (Subsection 4.3) and by the desire to create more interpretable models.

We conducted a separate experiment for each pairing of model type (text-to-score and evidence-to-score) and experiment design (direct prediction of total scores and prediction via intermediate indicator scores.) The following models for predicting the total score were trained and evaluated:

- *Auxiliary step.* An auxiliary model for total score prediction using human indicator *b* and *c* scores as features. We will plug predicted indicator scores instead of the human ones for the evaluations with intermediate scores.
- *Experiment 1.* An evidence-to-score model for total score prediction using utterance label-based features.
- *Experiment 2.* Two evidence-to-score models for indicator *b* and *c* scores prediction (as intermediate scores) using utterance label-based features, followed by the auxiliary model.

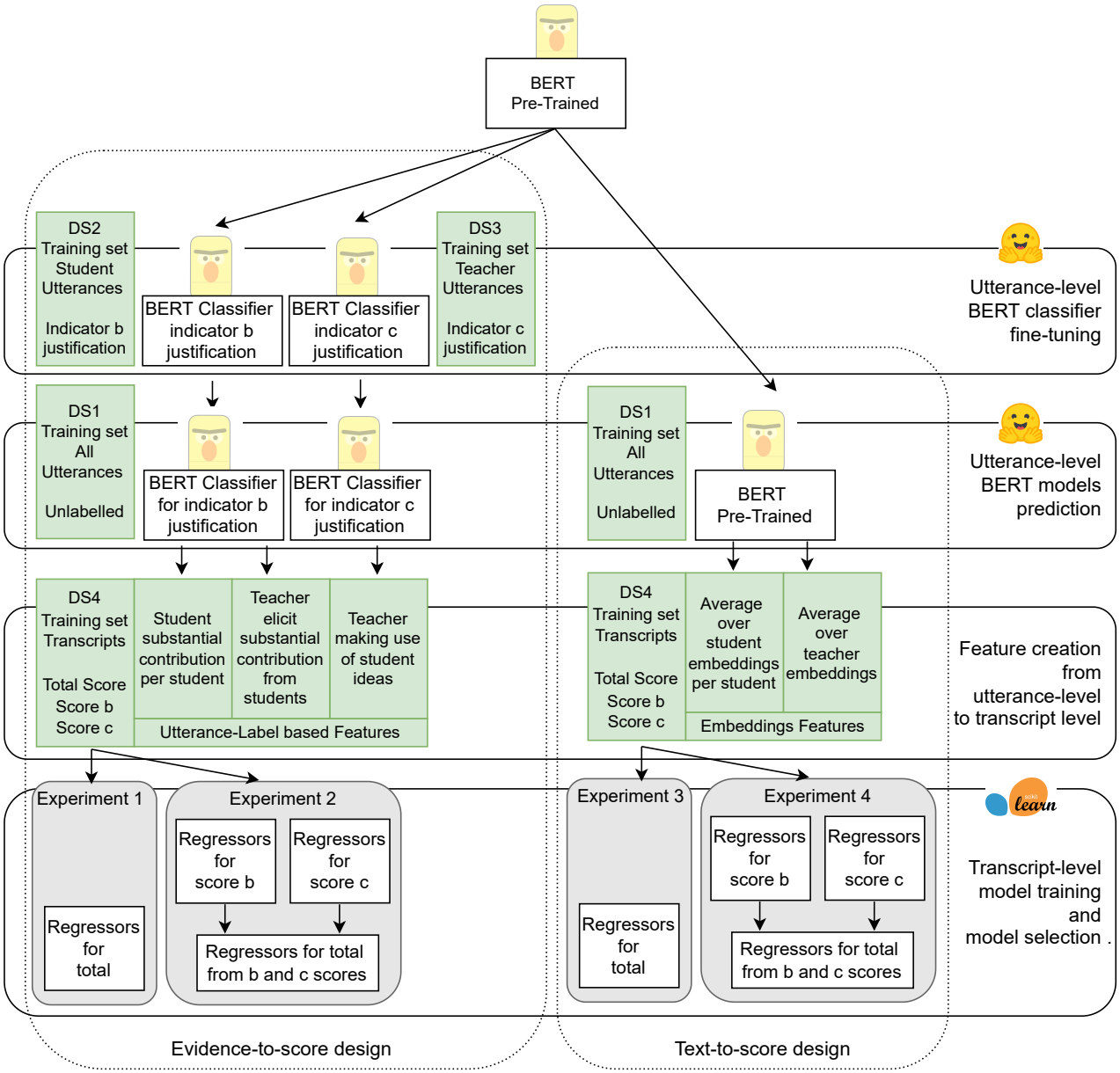

Figure 4: RQ 1 experimental design pipeline. The green boxes represent datasets. The white boxes represent machine learning models.

- *Experiment 3.* A text-to-score model for total score prediction using utterance embedding-based features.
- *Experiment 4.* Two text-to-score models for indicator b and c scores prediction (as intermediate scores) using utterance embedding-based features, followed by the auxiliary model.

All the models (including the auxiliary one) were trained and evaluated on dataset DS4 using the same division into the training and test sets as described in Subsection 5.1. The test set results of the four experiments are presented in Table 3. The results indicate that the two evidence-to-score models based on 7 features generated using automatically predicted utterance-level labels of indicators b and c performed similarly and were able to achieve a fair correlation with human total scores, which was even better than human inter-rater Pearson correlation (r = 0.379, Subsection 4.3).

7 RQ2: EXPERIMENTS AND RESULTS

The goal of this set of experiments is to analyze system performance on the prediction of indicator b and c justifications. Strong performance on this task would both support the validity argument for the evidence-to-score holistic scoring systems developed for

Experiment	Regressor	MSE	Pearson's r
1	BR	0.237	0.439
2	DT	0.287	0.421
3	BR	0.240	0.290
4	DT	0.341	0.016

Table 3: Stage 4 test-set results for evidence-to-score (Experiments 1 and 2) and text-to-score (Experiments 3 and 4) models.

RQ1 and provide a basis for personalized formative feedback to the teacher. The experiment design consisted of three steps.

Utterance-level BERT classifier fine-tuning. We followed the procedure described in Subsection 6.1 (Stage 1); however, the division into the training and test sets was different (see Subsection 5.1). Namely, all the labeled utterances that belong to the multi-rated transcripts were assigned to the test set to allow the Pyramid-based evaluation.

Utterance-level BERT classifier prediction. We calculated the predictions for the indicators *b* and *c* justifications as soft voting (the class with the highest average probability) over the predictions of the seven cross-validation models. For each utterance in the test set, we calculated two numbers: the prediction itself (0 or 1) and its probability. We sorted, per participant, the class 1 ('can serve as a justification') prediction probabilities from highest (the most confident class 1 prediction) to lowest (the least confident class 1 prediction).

Utterance-level BERT classifier Pyramid evaluation. First, we calculated the pyramids for the two multi-rater transcripts T1 and T2 as described in Subsection 5.2.3. The results are presented in Figure 3. Next, we calculated the human raters' Pyramid scores per transcript for each rater and compared them to the Pyramid scores of the classifier. Namely, for each of the 12 raters and each transcript (n=24):

(1) As the pyramids in Figure 3 were constructed using the data of all 12 raters, it was not fair to calculate the rater's score using these pyramids as-is. So, to pretend that a rater under analysis is a 'new' rater, we reduced by one the weights of all the utterances selected by this rater, yielding a slightly modified pyramid $ModP_{rater}$. Then, we calculated the rater's Pyramid score against the pyramid $ModP_{rater}$ as described in Subsection 5.2.3.
(2) Next, we calculated the Pyramid score of our automated prediction against the pyramid $ModP_{rater}$. To make a fair comparison with the 'new' human rater, the automated rater included the same number of utterances per participant as the human. The utterances with the highest class 1 probabilities, by the participant, were selected as automated predictions for the comparison with the 'new' rater.

The process above was run twice, for student utterances (indicator *b*) and for teacher utterances (indicator *c*). For indicator *c* it resulted in 24 pairs of Pyramid scores – one for each rater and each transcript vs the automated score, while for indicator *b* it resulted in 23 pairs because one of the raters did not provide any examples for indicator *b* justifications for one of the transcripts.

The distribution of human Pyramid scores for indicator *b* (M = 0.843, SD = 0.085) does not differ significantly from that of the automated system (M = 0.833, SD = 0.069), paired two-tailed t(23) = 0.959, p = .348. The distribution of human Pyramid scores for indicator *c* (M = 0.671, SD = 0.163) also does not differ significantly from the automated predictions (M = 0.783, SD = 0.129), paired two-tailed t(24) = -2.06, p = .051. We consider these results as an indication that the automated models can pick justifications that are as good as those picked by human raters.

8 DISCUSSION

In response to RQ1, our results show that automating the holistic scoring of teacher-led argumentation practice is feasible, as some of our models achieved human-level performance on the task. We built and evaluated two types of systems: *text-to-score*, where the scores are predicted from the raw text of the transcript using powerful pre-trained language models, and *evidence-to-score*, where the system additionally utilized rater justifications for the scores created during the human rating process. The latter models showed substantially stronger performance. This finding is particularly encouraging from the point of view of effective utilization of human-produced artifacts, and score justifications in this case, even if those were not set up as systematic annotations to support automation. While teams of technologists and science educators might be increasingly common, it is encouraging that rich but only partly systematic human data generated in a science education context can be effectively utilized, *post factum*, for automation purposes.

Previous research has shown that transparency of AI-powered technology (e.g., providing not only a 'bottom line' but also explaining why a decision was made) plays a critical role in practitioners' willingness to use automated recommendations [24, 34, 35]. Our best-performing scoring system also has the best explainability. In particular, a direct evaluation of the scores' evidence identification step suggests that not only can the system find *relevant* evidence, it can pinpoint *the strongest* evidence for the score as well as a human rater. This finding points towards an answer to RQ2 – the system's ability to identify evidence for the score could serve as a basis for formative feedback to the teacher. For example, the system could point out that the teacher has successfully engaged Jayla and Carlos in the discussion but has not done so for Will – either to guide reflection at the end of the interaction or as real-time feedback. Developing a personalized formative feedback mechanism based on these automated models is a major goal of future work.

Limitations. So far we have a relatively small amount of data for training transcript-level models; more data will need to be collected from additional studies with teachers. Second, while there is a strong alignment between total and indicator human scores (r > 0.8), there is only moderate inter-rater agreement on the total and indicator scores (r = 0.4-0.5). We plan to both improve the rubric scoring and to collect more human ratings per transcript to allow distillation of the average human judgment as well as identification of cases that are more or less controversial for human raters; [25] we argue that evaluation on uncontroversial (exemplar) cases provides important information for understanding system performance. Previous research provides evidence that machine learning models are sometimes able to ignore biases and idiosyncrasies of specific

human raters and agree with humans better than humans agree with each other. Providing additional human-scored data labeled by various raters could improve the automated system performance.

9 SUMMARY AND CONCLUSION

Engaging students in argument from evidence is an essential goal of science education. This is a complex skill to develop; recent research in science education proposed the use of simulated classrooms to facilitate the practice of the skill. We used data from one such simulated environment to explore whether automated analysis of the transcripts of the teacher's interaction with the simulated students using Natural Language Processing techniques could yield an accurate evaluation of the teacher's performance.

Teacher performance was assessed as a total score for the teacher's ability to facilitate an argumentation-based science discussion, as well as indicator scores for specific sub-skills, such as 'elicit substantive contributions from all students' or 'engage with students' ideas'. For the latter, raters were asked to provide score justifications. We used these data to create (a) text-to-score models that use state-of-the-art pre-trained Transformer models (BERT) to predict the scores directly from the text of the transcripts, and (b) evidence-to-score models that included an intermediate step of identifying utterances that could serve as justifications/evidence for the scores. We found that the latter models performed better and comparably to humans, both in terms of total scores and in the justifications for these scores.

These findings open up the possibility of generating automated scores paired with formative feedback with concrete evidence for the scores. The unique affordances of being automatically generated provide an opportunity to share such formative feedback at scale with each teacher who uses the simulated classroom environment during or shortly after each session.

ACKNOWLEDGMENTS

The authors are grateful for the raters who worked on scoring these discussion performances, the preservice teachers and teacher educators who participated in previous studies where this data was collected, and for the supportive research teams who helped to collect this data. The data used in this study was collected as part of two research projects funded by the National Science Foundation (grant No. 1621344 and grant No. 2037983). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally Identifying Funneling and Focusing Questions in Classroom Discourse. In *17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. 224–233.

[2] Meghan Bathgate, Amanda Crowell, Christian Schunn, Mac Cannady, and Rena Dorph. 2015. The learning benefits of being willing and able to engage in scientific argumentation. *International Journal of Science Education* 37, 10 (2015), 1590–1612.

[3] Amanda Benedict-Chambers and Roberta Aram. 2017. Tools for teacher noticing: Helping preservice teachers notice and analyze student thinking and scientific practice use. *Journal of Science Teacher Education* 28, 3 (2017), 294–318.

[4] Amanda Benedict-Chambers, Sarah J Fick, and Anna Maria Arias. 2020. Preservice Teachers' Noticing of Instances for Revision during Rehearsals: A Comparison across Three University Contexts. *Journal of Science Teacher Education* 31, 4 (2020), 435–459.

[5] Leema K Berland and David Hammer. 2012. Framing for scientific argumentation. *Journal of Research in Science Teaching* 49, 1 (2012), 68–94.

[6] Leema K Berland and Katherine L McNeill. 2010. A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education* 94, 5 (2010), 765–793.

[7] Julie Cohen, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis* 42, 2 (2020), 208–231.

[8] Christine P Dancey and John Reidy. 2017. *Statistics without maths for psychology*. Pearson London.

[9] Elizabeth A Davis, Matthew Kloser, Andrea Wells, Mark Windschitl, Janet Carlson, and John-Carlos Marino. 2017. Teaching the practice of leading sense-making discussions in science: Science teacher educators using rehearsals. *Journal of Science Teacher Education* 28, 3 (2017), 275–293.

[10] Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. *arXiv preprint arXiv:2106.03873* (2021).

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Lisa A Dieker, Charles E Hughes, Michael C Hynes, and Carrie Straub. 2017. Using simulated virtual environments to improve teacher performance. *School-University Partnerships* 10, 3 (2017), 62–81.

[13] Patrick J Donnelly, Nathan Blanchard, Borhan Samei, Andrew M Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystran, and Sidney K D'Mello. 2016. Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 45–53.

[14] Fayyaz Ahmad Faize, Waqar Husain, and Farhat Nisar. 2017. A critical review of scientific argumentation in science education. *Eurasia Journal of Mathematics, Science and Technology Education* 14, 1 (2017), 475–483.

[15] Evan J Fishman, Hilda Borko, Jonathan Osborne, Florencia Gomez, Stephanie Rafanelli, Emily Reigh, Anita Tseng, Susan Million, and Eric Berson. 2017. A practice-based professional development program to support scientific argumentation from evidence in the elementary classroom. *Journal of Science Teacher Education* 28, 3 (2017), 222–249.

[16] Anthony Tuf Francis, Mark Olson, Paul J Weinberg, and Amanda Stearns-Pfeiffer. 2018. Not just for novices: The programmatic impact of practice-based teacher education. *Action in Teacher Education* 40, 2 (2018), 119–132.

[17] Vetti Giri and MU Paily. 2020. Effect of scientific argumentation on the development of critical thinking. *Science & Education* 29, 3 (2020), 673–690.

[18] J Bryan Henderson, Katherine L McNeill, María González-Howard, Kevin Close, and Mat Evans. 2018. Key challenges and future directions for educational research on scientific argumentation. *Journal of Research in Science Teaching* 55, 1 (2018), 5–18.

[19] Matthew Huggins, Sharifa Alghowinem, Sooyeon Jeong, Pedro Colon-Hernandez, Cynthia Breazeal, and Hae Won Park. 2021. Practical guidelines for intent recognition: Bert with minimal training data evaluated in real-world hri application. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 341–350.

[20] Emily Jensen, Samuel L. Pugh, and Sidney K. D'Mello. 2021. A deep transfer learning approach to modeling teacher discourse in the classroom. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 302–312.

[21] David Kaufman and Alice Ireland. 2016. Enhancing teacher education with simulations. *TechTrends* 60, 3 (2016), 260–267.

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

[23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[24] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2390–2395.

[25] Anastassia Loukina, Klaus Zechner, James Bruno, and Beata Beigman Klebanov. 2018. Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 1–12.

[26] Katherine L McNeill and Amanda M Knight. 2013. Teachers' pedagogical content knowledge of scientific argumentation: The impact of professional development on K–12 teachers. *Science Education* 97, 6 (2013), 936–972.

[27] Jamie N. Mikeska and Heather Howell. 2020. Simulations as practice-based spaces to support elementary teachers in learning how to facilitate argumentation-focused science discussions. *Journal of Research in Science Teaching* 57, 9 (2020), 1356–1399.

[28] Jamie N. Mikeska and Heather Howell. 2021. Authenticity perceptions in virtual environments. *Information and Learning Sciences* 122, 7/8 (2021), 480–502.

[29] Jamie N. Mikeska, Heather Howell, Joseph Ciofalo, Adam Devitt, Elizabeth Orlandi, Kenneth King, Michelle Lipari, and Glenn Simonelli. 2021. Conceptualization and development of a performance task for assessing and building elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics: The mystery powder task (Research Memorandum No. RM-21-06), ETS.

[30] Jamie N. Mikeska, Heather Howell, Lisa Dieker, and Michael Hynes. 2021. Understanding the role of simulations in K-12 mathematics and science teacher education: Outcomes from a teacher education simulation conference. *Contemporary Issues in Technology and Teacher Education* 21, 3 (2021), 781–812.

[31] Jamie N. Mikeska, Heather Howell, and Devon Kinsey. In press. Do simulated teaching experiences impact elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics and science? *Journal of Teacher Education* (In press).

[32] Jamie N. Mikeska, Heather Howell, and Carrie Straub. 2019. Using Performance Tasks within Simulated Environments to Assess Teachers' Ability to Engage in Coordinated, Accumulated, and Dynamic (CAD) Competencies. *International Journal of Testing* 19, 2 (2019), 128–147.

[33] Jamie N. Mikeska, Calli Shekell, Adam V. Maltese, Justin Reich, Meredith Thompson, Heather Howell, Pamela S. Lottero-Perdue, and Meredith Park Rogers. 2022. Exploring the Potential of an Online Suite of Practice-Based Activities for Supporting Preservice Elementary Teachers in Learning How to Facilitate Argumentation-Focused Discussions in Mathematics and Science. In *Proceedings of Society for Information Technology & Teacher Education International Conference 2022*, Elizabeth Langran (Ed.).

[34] Tanya Nazaretsky, Moriah Ariely, Mutlu Cukurova, and Giora Alexandron. 2022. Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology* 53, 4 (2022), 914–931.

[35] Tanya Nazaretsky, Mutlu Cukurova, and Giora Alexandron. 2022. An Instrument for Measuring Teachers' Trust in AI-Based Educational Technology. In *LAK22: 12th international learning analytics and knowledge conference*. 56–66.

[36] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)* 4, 2 (2007), 4–es.

[37] NGSS Lead States. 2013. Next generation science standards: For states, by states (Vol 1) Washington.

[38] Amy Ogan. 2019. Reframing classroom sensing: Promise and peril. *Interactions* 26, 6 (2019), 26–32.

[39] Jonathan F Osborne, Hilda Borko, Evan Fishman, Florencia Gomez Zaccarelli, Eric Berson, KC Busch, Emily Reigh, and Anita Tseng. 2019. Impacts of a practice-based professional development program on elementary teachers' facilitation of and student engagement with scientific argumentation. *American Educational Research Journal* 56, 4 (2019), 1067–1112.

[40] Joe Oyler. 2019. Exploring teacher contributions to student argumentation quality. *Studia Paedagogica* 24, 4 (2019), 173–198.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[43] Luis P Prieto, Kshitij Sharma, Pierre Dillenbourg, and María Jesús. 2016. Teaching analytics: towards automatic extraction of orchestration graphs using wearable sensors. In *LAK16: 6th International Learning Analytics and Knowledge Conference*. 148–157.

[44] Luis Pablo Prieto, Kshitij Sharma, Łukasz Kidzinski, María Jesús Rodríguez-Triana, and Pierre Dillenbourg. 2018. Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data. *Journal of Computer Assisted Learning* 34, 2 (2018), 193–203.

[45] GO Discuss Project. 2021. Scoring. https://doi.org/10.5064/F6NJU10I

[46] Victor Sampson and Margaret R Blanchard. 2012. Science teachers and scientific argumentation: Trends in views and practice. *Journal of Research in Science Teaching* 49, 9 (2012), 1122–1148.

[47] Victor Sampson, Jonathon Grooms, and Joi Phelps Walker. 2011. Argument-Driven Inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education* 95, 2 (2011), 217–257.

[48] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[49] Abhijit Suresh, Jennifer Jacobs, Vivian Lai, Chenhao Tan, Wayne Ward, James H Martin, and Tamara Sumner. 2021. Using transformers to provide teachers with personalized feedback on their classroom discourse: The TalkMoves application. *arXiv preprint arXiv:2105.07949* (2021).

[50] Abhijit Suresh, Tamara Sumner, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2019. Automating analysis and feedback to improve mathematics teachers' classroom discourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9721–9728.

[51] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (csur)* 53, 3 (2020), 1–34.

[52] Zuowei Wang, Kevin Miller, and Kai Cortina. 2013. Using the LENA in Teacher Training: Promoting Student Involement through automated feedback. 4 (2013), 290–305.