Automated Scoring of Argumentation-focused Teaching Transcripts: Added Value of Human Annotations

[1]**Duy Pham,** [2]**Viet Lai,** [1]**Jamie Mikeska,** [1]**Jonathan Steinberg**
[1]**Heather Howell,** [2]**Thien Huu Nguyen**
[2]Educational Testing Service, Princeton, NJ 08540
{dnpham,jmikeska,jsteinberg,hhowell}@ets.org
[2]Department of Computer Science, University of Oregon, Eugene, OR 97403
{vietl,thien}@cs.uoregon.edu

Abstract

This study used human annotations and natural language processing (NLP) features to predict human scores of argumentation-focused teaching transcripts. Results indicated that adding human annotations on top of NLP features helped increase Quadratic Weighted Kappa (QWK) from 0.44 to 0.54 which is very close to the human score QWK of 0.56.

1 Introduction

1.1 Practice Space for Teacher Learning and The Scoring Challenge

Learning how to teach is a complex endeavor. The most recent student learning standards in the United States require that teachers learn how to enact challenging teaching practices, such as being able to elicit, interpret, and use students' ideas and previous experiences to facilitate productive discussions. In the last few decades, practice-based teacher education has been identified as one potential way of supporting teachers who are learning how to engage in such practices within and across content areas, particularly preservice teachers (PSTs) who are enrolled as candidates within teacher education programs. Practice-based teacher education uses teachers' engagement in the work of teaching as the site for learning.

One of the key pedagogies used within practice-based teacher education is approximations of practice whereby PSTs try out a new instructional strategy or skill in settings of reduced complexity. These approximations can involve PSTs in peer rehearsals where they try out a new teaching practice – for example, facilitating a science discussion with a small group of students – but do so with their fellow peers acting as the K-12 students in the practice discussion. Other approximation approaches train adults who are not members of the class to act as K-12 students during face-to-face rehearsals, sometimes relying on other faculty or staff to serve in this role. Due to advances in technology, a more recent trend is the use of online simulated classrooms to serve as the practice space in teaching approximations. For example, Mursion has developed several different K-12 simulated classrooms in which teachers can practice interacting with K-12 student avatars as they try out new instructional strategies. In these simulated classrooms, an adult serves as a human-in-the-loop who acts and responds in real-time to one or more K-12 student avatars during the interaction, although the adult's identity stays obscured from the teacher participating in the approximation.

While previous research has indicated the importance of practice in these various approximations to build teachers' knowledge and skills, findings have suggested that it is not the practice alone that supports teachers' learning, but practice coupled with opportunities for feedback and reflection that are

Preprint. Under review.

the main drivers of change Benedict-Chambers (2016); Mikeska et al. (2021). As such, not only does one need rich environments for engaging PSTs in viable and productive approximations, but one also needs mechanisms to support the provision of substantive feedback and opportunities for reflection. In teacher education, this feedback can come in the form of scores based on rubrics or in the form of written statements highlighting the strengths and potential areas for growth aligned to the PSTs' performance in the approximation. Traditionally, scores and feedback have been human-generated, requiring a significant investment of time from human raters to score and provide feedback. Currently, there is a need in teacher education to explore how machine learning (ML) and natural language processing (NLP) can be leveraged to develop and deploy automated scoring approaches to support PST learning of complex teaching practices. In this study, we examine how ML/NLP approaches can be used to generate automated scoring models that can be applied to argumentation-focused discussions that PSTs facilitated in an online simulated classroom environment consisting of five upper elementary student avatars.

We begin by providing an overview of the potential for using simulated classrooms as productive tools to support PST learning and explaining the importance of providing learning opportunities for PSTs to learn how to engage in one ambitious teaching practice: facilitating argumentation-focused discussions. Then, we provide background about previous studies exploring the role of ML/NLP in automated scoring and recent developments in this research area. After that, we detail this study's research questions and methodology, including describing the two performance tasks that PSTs used to facilitate argumentation-focused discussions in the simulated classroom and the annotation framework that our team developed and applied to 100 individual discussion transcripts. We end by sharing the study's main findings and discussing implications and future directions when applying ML/NLP approaches in teacher education contexts to support teacher learning.

1.2 Using Simulated Classrooms to Support Teacher Learning

In previous studies, researchers focused on using Mursion's upper elementary simulated classroom as a practice space for PSTs to learn how to engage in one core teaching practice: facilitating argumentation-focused discussions Mikeska and Howell (2020); Howell et al. (2021). This previous research involved providing opportunities for PSTs – as part of their elementary mathematics or science methods course – to prepare for, facilitate, and then reflect on their simulated discussion. Each discussion the PSTs facilitated in the simulated classroom focused on a topic in mathematics (e.g., comparing fractions) or science (e.g., properties of matter). In this previous work, we conceptualized the teaching practice of facilitating argumentation-focused discussions via five key dimensions, or features, of the practice. These dimensions included: (a) attending to student ideas, (b) developing a coherent and connected storyline, (c) encouraging peer interactions, (d) developing students' conceptual understanding, and (e) engaging students in argumentation. In this study, we focused our exploratory efforts on developing and applying an annotation framework for the fifth dimension by identifying moves that PSTs use to prompt students to engage in argumentation during the discussion and how the students engage in argumentation with their peers.

Argumentation is a key instructional practice that is cited in both the Common Core State Standards in mathematics and the Next Generation Science Standards as critical to students' success in these content areas Initiative et al. (2010); Council et al. (2013). Productive student engagement usually involves students generating, justifying, and revising arguments; comparing, evaluating, and critiquing arguments; and attempting to persuade each other as they work to come to a consensus. Productive teacher facilitation of argumentation-focused discussions typically involves teachers providing students with opportunities to engage in substantive mathematical or scientific sensemaking; consider and respond to each other's ideas, claims, justifications, and evidence-based reasoning; and offer and respond to counter arguments, challenges, and rebuttals. The importance of this teaching and instructional practice has resulted in some recent research efforts (e.g., Kazemi et al., 2021) to help teachers learn how to productively engage students in argumentation. These efforts, like other efforts that focus on practicing the work of teaching, often depend on time-intensive processes to generate meaningful feedback, representing both a cost barrier and a barrier to the feedback being delivered on time. One area that needs further examination is how ML and NLP can be leveraged productively as part of these learning opportunities, especially when using online environments such as simulated classrooms to support teacher learning.

1.3 Automated Scoring as a Solution for Scalability

As discussed in the earlier section, simulated classrooms can be a helpful tool for PSTs to practice teaching and for teacher educators to assess PSTs' teaching skills Mikeska and Howell (2020). However, scaling up the use of this tool to support formative assessments of PSTs' teaching skills presents challenges. In the authors' prior work, PSTs' performances were scored by human raters which are resource-consuming. On average, it takes each rater two to three hours to score one 20-minute teaching transcript. In addition, it usually takes a few days if not weeks to schedule human raters to score teaching transcripts which makes it unfeasible to provide timely results to PSTs and teacher educators. The use of ML and NLP to automatically score PSTs' performances in simulated classroom environments can therefore address this issue. If we can successfully develop and implement ML and NLP-driven automated scoring solutions, we can then significantly reduce the cost of human scoring and shorten the time needed for scoring such teaching performances. In recent years, advances in the field of ML and NLP such as automated annotation and deep learning have brought hope to the field of educational measurement. Such advances are believed to enable us to use computers to score responses at a level of consistency comparable to that of human raters Kumar and Boulanger (2021).

Researchers have been using ML and NLP to automatically score written essays, short constructed responses, or short speaking transcripts for at least a few decades Burstein et al. (1998); Page (1966). To the best of our knowledge, automated scoring for longer, multi-speaker discussion transcripts is a much newer topic. In this section, we will first summarize the big picture of automated scoring. We will then focus on reviewing the studies we found that reported results of using ML and NLP to score dialogs and discussion transcripts.

One of the first computer applications to do automated scoring of student essays was developed by Ellis Page from Duke University in 1966 Page (1966, 1994). In Page (1966), the author reported using regression models to predict human scores using many surface features such as word counts and punctuation errors. Later, researchers extended the independent variables to include more linguistic, structural, and contextual variables such as the number of complement, subordinate, or infinitive clauses or argument units Burstein et al. (1998); Chen et al. (2016). More recently, with the advancement of deep learning, many more NLP approaches have been explored to automatically score student responses with reference to a scoring rubric or dialogues Kumar and Boulanger (2021); Surya et al. (2019). Using existing NLP tools, these authors extracted linguistic and rubric-based features of student essays or discussions. They then used these features to build their statistical models or train their deep learning engines to predict scores. These studies reported promising results regarding the consistency between human and machine scores measured by Quadratic Weighted Kappa (QWK) Cohen (1968). For example, (Kumar and Boulanger, 2021) reported mean QWKs of 0.65 to 0.68 for their automated scoring engines of written essays, which were comparable to the consistency between human raters in their study. In addition to automated scoring for student essays and writing responses, other studies have reported on automated scoring for argumentation-focused dialogs and classroom discussions.

(Lugini and Litman, 2017)investigated the use of NLP to score a specific argumentation dimension called specificity of classroom discussions by high school students. In their study, specificity refers to the quality (i.e., low, medium, high) of an argument move that is uniquely related to a particular subject. Specificity of an argument move consists of four elements: (i) whether the move is specific to a subject, (ii) whether it contains significant qualifications or elaborations, (iii) whether it uses a specific set of vocabulary related to the content of the discussion, and (iv) whether it provides a series of reasons. The authors used Speciteller, which is a tool developed by a group of researchers at the University of Pennsylvania, to detect the specificity of an utterance by capturing shallow features such as sentence length or number of subjective and polar words, as well as semantic, lexical, and syntactic features. Using the extracted variables, they built several regression models to predict the specificity scores by humans for more than 2,000 turns at talk from 23 classroom discussions. The authors reported QWKs of 0.58 to 0.66 for their models. This research and other studies Lugini and Litman (2018) suggested that existing NLP tools can be used to extract well-defined features such as specificity and ML approaches such as linear regressions can help predict human ratings of argumentation aspects of classroom discussions. While the cited work forms a useful basis for understanding how argumentation features can be extracted, each of the above projects focused exclusively on student argumentation, and not on the teaching moves, or facilitation, that prompted

argumentation to take place. To develop automated scoring models for teaching performances, we set two research questions for our study.

1.4 Research Questions

In this study, we built on the existing literature and methods of automated scoring for written responses and classroom discussion to develop scoring engines to predict "Engaging Student in Argumentation" teaching scores of PSTs. We set out to answer two research questions: (i) were machine scores for the "Engaging Students in Argumentation" dimension generated using NLP features and human annotations consistent with human scores? and (ii) did the use of human annotation data help improve the quality of the prediction model? We evaluated our machine scoring results using QWK, an accuracy rate used in (Cui, 2021), and mean squared error (MSE) of estimation. We also used the percentage of variance explained by regression models (i.e., R-square) to compare and contrast the models.

2 Method

To answer the research questions, we drew on a corpus of data collected by Authors (in review). These data included a pool of 215 transcribed performances from one of two teaching tasks, with associated scores from human raters. The tasks used to generate the transcribed performances were entitled Mystery Powder Mikeska and Howell (2020) and Ordering Fractions Howell et al. (2021). Each task provided an opportunity for PSTs to practice facilitating an argumentation-focused discussion on a science or mathematics topic in Mursion's upper elementary simulated classroom. In the Mystery Powder science task, each PST facilitated a discussion in which they worked to have the five student avatars come to a consensus about the identity of a mystery powder (in this case, baking soda) using evidence about the properties of six known powders and to determine which properties were most useful in making that determination Mikeska and Howell (2020). In the Ordering Fractions mathematics task, the goal of the discussion was for the students to "evaluate, justify, compare, and contrast strategies for ordering fractions with different numerators and different denominators" Howell et al. (2021). Each PST worked to have the five student avatars come to a consensus around the order of three fractions (3/10, 9/10, and 3/4) and to determine if the strategies they used to order these fractions would generalize to any set of fractions. Across mathematics and science, each PST facilitated the discussion in the simulated classroom for up to 20 minutes and these transcripts captured the discussion that the PST facilitated with the five student avatars. In a previous study, trained human raters scored the transcripts using a scoring rubric. Human raters double-scored a portion of these transcripts and the QWK for "Engaging Students in Argumentation" dimension scores was 0.56 indicating that the human scores of three points (i.e., 1, 2, and 3) for this dimension were moderately consistent between two raters. In this study, trained human coders used a coding framework that we developed iteratively through several cycles to assign labels reflecting some aspects of well-developed argumentation for each utterance of a transcript. In the future, we will develop automated annotation engines to assign codes to each utterance of new transcripts. After human annotation was completed, we used Python to aggregate human coding results and extract NLP features to build regression models. In what follows, we describe the details of the data, the annotation process, and the models we investigated during this study.

2.1 Data

Due to limited resources, we could only annotate 100 transcripts. Thus, we carefully selected these from the pool of 215 transcripts. The demographic composition of the PSTs who produced the video or audio files which were transcribed to the 215 documents with human scores showed that the pool was consistent with trends in elementary teacher demographics in the United States Banilower et al. (2018). Eighty-nine percent of the PSTs identified as female, and 11% of them identified as male. Most of the PSTs were White. When PSTs across subjects were taken together, 72% were reported as White. Table 1 reports more details of the demographic and scoring information of the transcript pool.

Subject	Gender		Race		Score		
	Female	Male	White	Non-White	1	2	3
Math	113(89%)	16(11%)	87(72%)	42(28%)	81(63%)	43(33%)	5(4%)
Science	78(91%)	8(9%)	68(79%)	18(21%)	35(41%)	32(37%)	19(22%)
Total	191(89 %)	24(11%)	155(72%)	60(28%)	116(54%)	75(35%)	24(11%)

Table 1: Demographics, Score Point Counts, and Row Percentages (in parentheses) of All Transcripts

2.2 Transcript Selection

Given the demographic and score distributions of the transcript pool, we set the following rules to select 100 transcripts for human annotation, model building, and evaluation. We aimed to (i) maximize the diversity of the selected transcripts by including at least one performance from all non-White and non-female participants, (ii) randomly selecting the remaining transcripts from the White and female participants, (iii) balance the score distribution across three score points to the maximum extent possible. Due to the small number of mathematics transcripts with the highest score points, we selected all with a score point 3 for this subject area. Table 2 below shows the demographic and score information of all the transcripts selected for human annotation and model building.

Subject	Gender		Race		Score		
	Female	Male	White	Non-White	1	2	3
Math	41 (82%)	9 (18%)	28 (56%)	22 (44%)	22 (44%)	23 (46%)	5 (10%)
Science	46 (92%)	4 (8%)	38 (76%)	12 (24%)	17 (34%)	17 (34%)	16 (32%)
Total	87 (87%)	13 (13%)	66 (66%)	34 (34%)	39 (39%)	40 (40%)	21 (21%)

Table 2: Demographic, Score Point Counts, and Row Percentages (in parentheses) of 100 Selected Transcripts

Chi-square analyses were conducted as a way to evaluate the transcript selection process. When examining the effect of selection on gender, there was no significant association overall ($\chi^2(1,1) = .64; p = .43$) or by subject area (Math: $\chi^2(1,1) = 2.35; p = 0.12$; Science: $\chi^2(1,1) = 0.24; p = 0.62$)). When examining the effect of selection on race, there was a non-significant trend in the association overall ($\chi^2(1,1) = 3.45; p = 0.06$) and for science transcripts ($\chi^2(1,1) = 0.68; p = 0.41$), but there was an effect for math transcripts ($\chi^2(1,1) = 4.87; p = 0.03$), such that there was a slight over-representation in transcripts being selected among non-White participants (adjusted residual = 2.2). When examining the effect of selection on dimension scores, given all math transcripts receiving scores of 3 were selected, the only appropriate comparison to report is for science transcripts. A significant effect on selection was detected ($\chi^2(1,1) = 6.95; p = 0.03$), such that transcripts receiving scores of 3 were more prominent in the selected sample (adjusted residual = 2.6). The chi-square results indicated that the selected transcripts were quite consistent with the overall pool with an exception for the selected mathematics transcripts of having a higher percentage of non-White participants and scores of 3. As stated above, this over-representation was intentional to increase the diversity of the data we used in this study.

2.3 Transcript Annotation

Our research team developed an annotation framework to characterize how the PSTs prompted student engagement in mathematical and scientific argumentation and how the students engaged in argumentation during these discussions. We drew upon three main sources to generate the specific codes to include in this annotation framework. First, we reviewed previous empirical and practitioner literature in mathematics and science education focused on how teachers prompt and how students engage in argumentation within these disciplines. Second, we considered the types of argumentation-focused teaching moves and student interactions we have observed across these discussions, based on our previous analysis and scoring of these videos for a previous research project. Third, we leveraged expertise from our group of assessment developers and research scientists who had deep knowledge of the subject matter and/or content teaching in one or both areas. Our goal in this annotation was to bridge a gap in the extant data between the type of information produced by raters in prior work and the type of information needed to support NLP and ML techniques. Prior scores were produced via a holistic scoring procedure that resulted in high-level inferential judgments across the full transcript rather than fine-grained evidence at the utterance level. Our challenge, therefore, was to build on the

extant conceptualization of argumentation, maintaining consistency with the prior scoring efforts, while simultaneously producing codes that were both more specific and more specifically linked to the utterance-level dialogue.

The development of the annotation framework occurred iteratively and collaboratively across team members. In the first step, three of the authors used their review of the literature and previous experience observing these discussions to brainstorm a comprehensive list of possible codes for inclusion in the framework. Seven of these initial codes (e.g., eliciting data; providing reasoning/justification) were designed to characterize the argumentation moves that teachers used to prompt students to provide data and reasoning and the moves that students used to state, explain, and justify their own or others' arguments Erduran et al. (2004); Oyler (2019). Eight of these initial codes (e.g., raising or responding to counter arguments/challenges/rebuttals; evaluating) were developed to capture the argumentation moves that PSTs used to prompt students to debate various arguments and the moves that the students make to evaluate, critique, and persuade one another about specific arguments. The other eight initial codes identified generic moves that one might make during a discussion, such as paraphrasing another person's idea or redirecting another person to return to or address something that has been missed, neglected, or deserves additional attention during the discussion Oyler (2019).

Second, we worked with our larger research team to try applying the initial set of 23 codes to PST and student utterances in a small set of mathematics and science discussion transcripts. We started by completing group coding of one mathematics and one science discussion transcript and then moved on to individual coding with group reconciliation of one mathematics and one science discussion transcript. Based on this initial group and individual coding with reconciliation, we refined the annotation framework significantly by collapsing codes that were similar and removing codes that were not focused directly on engaging students in argumentation. The refined annotation framework included nine codes that captured the critical argumentation moves made by the PSTs and students to prompt and/or engage in mathematical or scientific argumentation during these discussions. Table 3 shows the list of the codes and their brief descriptions.

Before beginning the annotation coding for the 100 transcripts used in this study, we selected another small set of transcripts to use for training purposes across our team. In the training, we began by discussing each of the code names, descriptions, and examples. Then, each team member individually coded one science and one mathematics transcript, followed by group reconciliation of each one. After that, team members worked in pairs to code two additional transcripts (one per content area) individually followed by reconciliation with their partners. Once training was complete, five raters on our team completed the coding of the 100 transcripts across 15 weeks. Each week each rater coded two to three transcripts individually and reconciled one to two double-coded transcripts (from their previous week's coding) with a partner. In addition, our full team met on a bi-weekly basis to discuss coding challenges and make minor refinements, as needed, to the annotation framework to ensure that we maintained a shared understanding of how to consistently apply the codes to the discussion transcripts.

To ensure the quality of the coding process, weekly quality monitoring was implemented to evaluate the quality of the ratings by the coders. Key statistics that were calculated were percent exact agreement, Cohen's kappa, and the intraclass coefficient. These analyses were performed at the transcript level and the individual code level. While 18 transcripts were selected a priori for reconciliation between raters, nine others where the Cohen's kappa value fell below 0.50 were flagged for additional reconciliation. No adjustments were made to the agreement statistics after reconciliation because it was presumed the final coding would represent a perfect agreement between raters.

2.4 Statistical and NLP Models

In this study, we adopted a cross-validation method that has been widely adopted in ML and NLP areas to increase the sample size of testing data Bishop and Nasrabadi (2006). We partitioned the 100 transcripts into five groups of 20 testing transcripts. For the first group of the first 20 documents, we used the 80 remaining transcripts to estimate our prediction models. Then, we fit the models to the 20 transcripts of group 1 to obtain predicted scores. We repeated these steps five times, thus we obtained predicted scores for all 100 transcripts. Of note was that these predictions were not coming from the same models because they were estimated using slightly different training sets. The cross-validation process allowed us to increase the sample sizes of our testing set.

Code Name	Code	Description
Explicating Argumentation	**EXA**	Communicating the key features or characteristics of high-quality argumentation discussions (needs to be explicit)
Eliciting a Claim	**ELC**	Asking or encouraging others to share their claims related to the discussion's targeted student learning goal (with or without data or reasoning to support the claims).
Stating a Claim	**STC**	Sharing a claim related to the discussion's targeted student learning goal (with or without providing any data or reasoning).
Eliciting Data	**ELD**	Asking or encouraging others to provide data to support or refute a claim (with or without reasoning to support the claim).
Providing Data	**PVD**	Sharing data to support or refute a claim (with or without providing reasoning).
Eliciting Reasoning	**ELR**	Asking or encouraging others to share their reasoning to support or refute a claim (with or without data to support the claim).
Providing Reasoning	**PVR**	Sharing reasoning to support or refute a claim (with or without providing data).
Building Consensus	**BCS**	A focus on consensus and/or providing opportunities for building consensus among participants.
Evaluating	**EVL**	Providing an evaluation of an argument, or part of an argument (claim, data, and/or reasoning), or an evaluation of whether an argument is strong or weak.

Table 3: Final Coding Framework to Capture Argumentation Features of Transcripts.

To explore different predictors, we tried out two regression models including linear regression and ridge regression models; and two classification models including ridge classifier and support vector machine classifier. Based on our experiment, we found that the regression models performed better than the classification models. Among the two regression models, the performance of the ridge regression models was not significantly better than those of the linear regression counterparts. Thus, we only report the results of the linear regression methods and predictors in the remainder of this paper

2.5 Natural Language Processing Features

In this study, to encode each transcript, we explored a number of feature combinations including word features and human code features. We described them in the following sections.

2.5.1 Term Frequency-Inverse Document Frequency

Our Baseline model uses only word features, namely Term Frequency-Inverse Document Frequency (TF-IDF) (Salton & McGill, 1983). Term frequency is the number of times that a term (i.e. a word) appears in a conversation. The intuition is that a term that occurs more frequently may be descriptive of how PSTs engage students in classroom argumentation. Previous studies such as (Lugini and Litman, 2018) used TF-IDF among other features to detect argumentation aspects of classroom discussion. The weight for a term is computed based on the term frequency using the following equation:

$$w_{t,d} = \begin{cases} 1 + \log(TF_{t,d}) & TF_{t_d} > 0 \\ 0 & otherwise \end{cases}$$

otherwise, where TF_t is the term frequency of the term t in transcript d. This frequency weight tells us that a more frequent term is more important than the less frequent ones. Inverse Document Frequency (IDF) considers a term that frequently appears in many transcripts not to be a good discriminator. Hence, its weight must be lower than those that occur in fewer transcripts. Mathematically, given a document frequency n_t of a term t, the inverse document frequency is computed as:

$$IDF\,(t) = log\frac{N}{n_t},$$

where N is the number of documents we have. Overall, TF-IDF combines both the term frequency weight and the inverse document frequency weight by multiplication:

$$TFIDF(t) = [1 + \log(TF_{t,d})] \, log \frac{N}{n_t}$$

Following common practice in NLP, we removed stop words (e.g., a, an, the, etc.) Cui (2021), and words with extremely low corpus frequency (i.e., under 10 appearances across all 100 transcripts) Goldberg (2022). After the removal, we kept 617 words out of 3,174 unique words across our transcripts. This resulted in 617 TF-IDF features to the feature set.

2.5.2 Unigram Human Code Frequency

The second type of feature that we explored is the frequency of human codes assigned to each utterance within a transcript. Our intuition was that some codes such as Building Consensus (BCS) contributed more to the argumentation facilitation quality of a transcript than others (e.g., Eliciting Data [ELD]). Hence, the code frequency, which can be represented by code distribution, reflects the quality of engaging students in argumentation. In particular, the code weight is computed as follows:

$$CF_{unigram}(c_i) = \frac{n_{c_i}}{N_{unigram}}$$

where n_{c_i} is the number of appearances of code c_i and $N_{unigram}$ is the total number of codes used in the corresponding transcript. This added 9 features to the feature set.

2.5.3 Bigram Human Code Frequency

Code frequency can capture an aspect of the teaching quality. However, it does not consider how well a PST responded to the students, and the satisfaction of the students after a PST's response. As such, we proposed to use the human code bigram to capture these interactions between students and PSTs. Similar to the code frequency, instead of considering codes of a single utterance, we considered pairs of codes of two consecutive utterances, of which one is from a student and the other is from a PST. The bigram code frequency weight for a pair of codes (c_i, c_j) is computed as follows:

$$CF_{bigram}(c_i, c_j) = \frac{n_{ij}}{N_{bigram}}$$

where n_{ij} is the number of appearances of the pair of codes $(c_i, c_j)$and N_{bigram} is the total number of pairs that appear in the corresponding transcript. To explore the impact of interaction order, we further considered two categories of bigram code features. The first one reflects how a PST responds to a student (Student First), while the second one captures how a student responds to a PST (PST First). Since there are 9 code types, each of these categories inserted 81 new features into the feature set. Figure 1 shows some examples of how the unigram and bigram code frequencies are computed.

Utterance	Codes	Speaker
1	ELR	PST
2	PVD, PVR	Student
3	ELC, ELR	PST
4	PVD	Student
5	ELC	PST

Unigram Frequency	Bigram Frequency	
	Student First	**PST First**
ELR: 2	PVD-ELC: 2	ELR-PVD: 2
ELC: 2	PVD-ELR:1	ELR-PVR: 1
PVD: 2	PVR-ELC:1	ELC-PVD: 1
PVR: 1	PVR-ELR:1	

Figure 1: Examples of Unigram and Bigram Code Frequency. The left table shows 5 utterances in order with their corresponding codes and speakers. The right table shows the frequencies.

2.6 Regression Models to Predict Scores

The Python package scikit-learn was used to generate TF-IDF and code bigram features for all 100 transcripts. Then, for each of the five groups of 20 testing transcripts, we estimated seven regression models using the 80 remaining training transcripts and deployed these statistical models to predict scores for the 20 testing transcripts. In the baseline model, we only regressed human scores on the

Label	EXA	ELC	STC	ELD	PVD	ELR	PVR	BCS	EVL
Exact Agreement (%)	96.8	93.6	89.7	96.8	93.3	90.6	89.8	91.6	84.8
Cohen's Kappa	0.44	0.51	0.56	0.56	0.65	0.59	0.62	0.61	0.57

Table 4: Average Human-human Agreement for Nine Annotation Labels

TF-IDF feature of each training transcript. In model 1, we kept the variables in the baseline model and added the frequency of human codes assigned to each student utterance of the transcripts to the regression equation of the baseline model. In model 2, we replaced the frequency of codes assigned to students in model 1 with the frequency of codes of PST utterances. In model 3, code frequencies for all utterances (i.e., both students' and PSTs') were included in the regression equation on top of the TF-IDF feature. Models 4 through 6 followed the same logic as models 1 through 3. However, instead of using code frequencies, we took code bigram features of students as the first utterance and/or PSTs as the first utterance as independent variables in addition to the predictor of the baseline model.

2.7 Model Evaluation

To evaluate the performance of the automated scoring engines, we used two out of five statistics described in (Cui, 2021) which are (i) QWK and (ii) accuracy rate. QWK is a widely used measure of consistency between two sets of scores when agreement by chance and the contribution of each score point are considered. We relied on conventional criteria proposed for QWK by (Landis and Koch, 1977) to evaluate our model. Accuracy is the ratio of correct scoring over all the scores when human scores are treated as truth. The higher the accuracy the more desirable the model. In addition, we also took the percent of variance explained by each regression model (i.e., R-square) and computed mean square error (MSE) of the unrounded scores predicted by the models to add to the interpretation of our results.

3 Results

In this section, we report our results to provide evidence to answer our research questions. First, we show the level of inter-annotator consistency for our human annotation data. Second, we describe the performance of the models we developed to predict human scores. Third, we delve deeper into the level of accuracy at each human score point. Last, we discuss the value of adding human annotation information into the prediction models.

3.1 Human Annotation Consistency

One of our research questions was about the added value of human annotation to our prediction models. To answer the question, we treated human coding results as input on top of the TF-IDF feature. Table 4 presents the average percent exact agreement and Cohen's kappas across all 50 double-scored transcripts. One can see that the statistics for percent exact agreement varied from about 85% to 97%, and the kappa values were in the range of moderate agreement. Given the level of consistency across human annotators, the human annotation was deemed to be reliable enough to be used as an independent variable in our regression models.

3.2 Prediction Model Performances

Given that most of the kappas of human annotations were around 0.60, the annotation results appeared to be consistently coded enough to be fed into our regression models. Table 5 shows the QWK, accuracy rate, MSE, and R-square of the seven linear regression models we considered in this study. As we can see, the baseline model had the lowest QWK of 0.44, the lowest accuracy rate of 0.51, and the lowest R-square of 0.03. Its MSE of 0.49 was also the highest among all the models. The low R-square of the baseline model was of note because using TF-IDF alone can only help us explain 3% of the variance of the human scores. The highest QWK and R-square were observed for model 6 in which the code bigram of 81 pairs of labels for both students and PSTs was added to the list of independent variables along with the TF-IDF feature. For this highest-performing model, QWK was

Model	TFIDF	HC Frequency		HC Bigram		QWK	Acc	MSE	R^2
		Student	PST	Student First	PST First				
Baseline	YES	NO	NO	NO	NO	0.44	0.51	0.49	0.03
1	YES	YES	NO	NO	NO	0.49	0.54	0.43	0.14
2	YES	NO	YES	NO	NO	0.49	0.54	0.43	0.14
3	YES	YES	YES	NO	NO	0.49	0.54	0.43	0.14
4	YES	NO	NO	YES	NO	0.53	0.57	0.43	0.19
5	YES	NO	NO	NO	YES	0.52	0.54	0.39	0.19
6	YES	NO	NO	YES	YES	0.54	0.56	0.39	0.22

Table 5: Results of Linear Regression Models Using NLP and Human Coding Features

Model	Human Scores	Machine Scores			Row Total	Accuracy
		1	2	3		
Baseline	**1**	16	22	1	39	0.41
	2	13	25	2	40	0.63
	3	1	10	10	21	0.48
Model 6	**1**	17	22	0	39	0.44
	2	10	28	2	40	0.70
	3	0	10	11	21	0.52

Table 6: Confusion Matrices of The Baseline and The Best Model

0.10 higher than that of the baseline model. This increase indicated that adding the code bigram for both students and PSTs made the machine scores more consistent with human counterparts.

The most salient improvement by adding the code bigram for both students and PSTs was the increase of 0.19 of the R-square for model 6 from the baseline prediction. This finding suggested that the code bigram can be a valuable source of information contributing to the improvement of the predictive quality of the baseline regression model. The addition of the feature can help explain 19% more variance in human scores. Performance measures from models 2 through 5 were in between the baseline and the best model (i.e., model 6) with only one

Note: Rows of the baseline and the best-performing models are bolded. exception of the accuracy rate of model 4. Indeed, the QWK of these models varied from 0.49 to 0.53. Their accuracy rates were from 0.54 to 0.57. The R-square of these models ranged from 0.14 to 0.19. And, their MSEs were from 0.39 to 0.43. These results were indicative that adding human coding features did improve the prediction quality of the baseline regression model. In the next section, we looked closer into the prediction quality at each human score point (i.e., 1, 2, or 3).

3.3 Prediction Accuracy at Each Score Point

Information on how the models performed at each human score point helped us figure out ways to improve the machine scoring engine we are trying to develop. Table 6 presents the confusion matrices for the baseline and the best model (model 6). As one can see, the accuracy rate for the middle score point of 2 was highest across the models, followed by the score point of 3, and was lowest for a score point of 1. For the baseline model, the accuracy rates were 41%, 63%, and 48% for score points of 1, 2, and 3, respectively. When code bigrams for all speakers were added, these helped enhance these accuracy rates up to 44%, 70%, and 52%, respectively, for the score points. The accuracy improved the most by seven percentage points for the middle score of 2, by 4% for a score of 3, and 3% for a score of 1. Since human scores were whole numbers, a rounding step was needed to convert regression-based predictions into a score of 1, 2, or 3. To delve deeper into the added value of human annotation, we also investigated the difference between unrounded machine scores of the baseline and the best models.

3.4 Added Value of Human Annotations

Figure 1 shows the scatter plot of unrounded predicted scores of the baseline and the best models. Coded in circles were the machine scores for human scores of 1. Meanwhile, human scores of 2 and 3 were labeled by triangles and pluses, respectively. Looking at the plot, one pattern emerged the machine score ranges associated with each human score point of the best model were narrower than

the baseline model. Indeed, unrounded machine scores associated with a human score of 1 by the former varied from 0.80 to 2.39. Whereas, the range for the latter was from 0.50 to 2.67. The same pattern was seen for human score points of 2 and 3. The predicted scores for model 6 for the middle score point varied from 1.10 to 2.84. Meanwhile, the range for the baseline model was from 0.76 to 3.10. For the human score point of 3, model 6 predicted scores varied from 1.57 to 3.03, and the baseline score range was from 1.38 to 3.21. The results of our analyses provided supporting evidence to answer our research questions. In what follows, we conclude our paper and discuss the limitations as well as future directions and implications of our study. Figure 1. Predicted Scores of the Baseline and the Best Models.

4 Conclusions and Discussion

4.1 Answers to Research Questions and Connections to Existing Literature

In this study, we set out to investigate the degree to which machine scores using NLP features and human annotations agreed with human scores and how the addition of human coding information helped improve the agreement. As reported in the Results section, when both TF-IDF and human code bigram features were used, the machine scores were moderately consistent with human scores. The QWK of the highest-performing model was 0.54 which outperformed the level of agreement reported in some automated scoring engines using NLP features to build regression-based predictors for essays Attali and Burstein (2004); Chen et al. (2016); Lintean et al. (2012). This level of consistency between machine and human scores is also on par with existing automated scoring engines for short responses Nakamura et al. (2016), and for persuasive essays Farra et al. (2015). Some studies that adopted deep learning models to score written essays reported QWKs higher than our best QWK of 0.54 Kumar and Boulanger (2021); Ramanarayanan (2020). However, the scoring consistency of our best model is very close to the human-to-human QWK of 0.56 we reported earlier for human scores. Given the low stakes of the transcript scores and the ease with which our model can be explained and interpreted by our regression approach, we hold that model 6 can be used to automatically score teaching transcripts of PSTs so we can provide timely information to support their training. The addition of human annotations into the baseline model using only the TF-IDF feature improved the machine-human agreement level and R-square of the prediction model. The human code bigram feature made the machine scores more precise and helped increase the percentage of variance explained by the model by 19%. This result suggested that auxiliary information derived from the original transcripts by human annotators is likely to play an important role in building automated scoring engines.

4.2 Limitations and Future Directions

Our study has some limitations. First, given the exploratory nature of our investigation, we only used basic NLP features and traditional ML techniques in this first automated scoring study for the teaching transcripts. Commercial automated scoring engines such as e-rater® or IntelliMetricTM usually capitalize on more than one hundred NLP features to build their scoring models Shermis and Hamner (2013). To address this drawback, we plan to engineer more features such as sentence-level and/or sentiment features as well as n-gram features with $n \geq 2$ to examine more statistical or deep learning models. However, it is important to note that the difficulty of capturing robust data sets of certain types currently limits the field's ability to apply NLP to certain situations, like the type of classroom teaching represented in this data set, and learning to make progress with less data may not just improve methods but may open up entire areas of application of NLP currently out of reach due to the impracticality of generating large data sets Cui (2021).

Second, due to limited resources, we were able to annotate only 100 transcripts as input data for this study. Even if this sample size is similar to some other automated scoring studies Ghosh et al. (2016), we hold that more data will only do more good than harm. As follow-up steps, we plan to explore more advanced NLP and ML techniques such as automated annotation and deep learning which likely need more data than the ones we utilized in this study. We plan to annotate more transcripts, so we can move forward. As the first step, we will develop and test solutions to automatically annotate new transcripts at a level of agreement comparable to our inter-annotator consistency Devlin et al. (2019). In the second step, we will use the annotation results along with NLP features to build more regression and deep-learning models to predict scores.

Third, our model's success may be limited by imperfect overlap between the original rater scores and the annotations, as these processes were completed at different time points. While every effort was made to remain consistent with the construct definition as we developed the annotation codes, specificity itself in the newly generated codes might have led to slightly different high-level judgments than those made via the prior process. The relatively higher inter-rater reliability on our annotation coding relative to the original scoring suggests that agreement at the utterance level was easier to achieve than agreement on the holistic score. This suggests future study of how raters make such high-level judgments and even suggests a potentially useful blended model in which human raters might make use of ML-generated transcript coding and scoring results to inform their thinking in making such judgments.

References

Attali, Y. and Burstein, J. (2004). Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2004(2):i–21.

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

Benedict-Chambers, A. (2016). Using tools to promote novice teacher noticing of science teaching practices in post-rehearsal discussions. *Teaching and Teacher Education*, 59:28–44.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D., and Wolff, S. (1998). Computer analysis of essay content for automated score prediction: A prototype automated scoring system for gmat analytical writing assessment essays. *ETS Research Report Series*, 1998(1):i–67.

Chen, J., Fife, J. H., Bejar, I. I., and Rupp, A. A. (2016). Building e-rater® scoring models using machine learning methods. *ETS Research Report Series*, 2016(1):1–12.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Council, N. R. et al. (2013). Next generation science standards: For states, by states.

Cui, Z. (2021). Machine learning and small data. *Educational Measurement: Issues and Practice*, 40(4):8–12.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erduran, S., Simon, S., and Osborne, J. (2004). Tapping into argumentation: Developments in the application of toulmin's argument pattern for studying science discourse. *Science education*, 88(6):915–933.

Farra, N., Somasundaran, S., and Burstein, J. (2015). Scoring persuasive essays using opinions and their targets. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 64–74.

Ghosh, D., Khanam, A., Han, Y., and Muresan, S. (2016). Coarse-grained argumentation features for scoring persuasive essays. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.

Goldberg, Y. (2022). *Neural network methods for natural language processing*. Springer Nature.

Howell, H., Mikeska, J., Tierney, J., Baehr, B., and Lehman, P. (2021). Conceptualization and development of a performance task for assessing and building elementary preservice teachers' ability to facilitate argumentation focused discussions in mathematics: The ordering fractions task. *Research Memorandum No. HYPERLINK "https://www. ets. org/Media/Research/pdf/RM-21-10.pdf" RM-21-10). ETS*.

Initiative, C. C. S. S. et al. (2010). National governors association center for best practices and council of chief state school officers. *Retrieved December*, 11:2012.

Kumar, V. S. and Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, 31:538–584.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lintean, M., Rus, V., and Azevedo, R. (2012). Automatic detection of student mental models based on natural language student input during metacognitive skill training. *International Journal of Artificial Intelligence in Education*, 21(3):169–190.

Lugini, L. and Litman, D. (2017). Predicting specificity in classroom discussion. In Tetreault, J., Burstein, J., Leacock, C., and Yannakoudakis, H., editors, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–61, Copenhagen, Denmark. Association for Computational Linguistics.

Lugini, L. and Litman, D. (2018). Argument component classification for classroom discussions. In Slonim, N. and Aharonov, R., editors, *Proceedings of the 5th Workshop on Argument Mining*, pages 57–67, Brussels, Belgium. Association for Computational Linguistics.

Mikeska, J., Howell, H., Dieker, L., and Hynes, M. (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):781–812.

Mikeska, J. N. and Howell, H. (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):1356–1399.

Nakamura, C. M., Murphy, S. K., Christel, M. G., Stevens, S. M., and Zollman, D. A. (2016). Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics. *Physical Review Physics Education Research*, 12(1):010122.

Oyler, J. (2019). Exploring teacher contributions to student argumentation quality. *Studia paedagogica*, 24(4):173–198.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Educational*, pages 127–142.

Ramanarayanan, V. (2020). Design and development of a human-machine dialog corpus for the automated assessment of conversational english proficiency. In *INTERSPEECH*, pages 419–423.

Shermis, M. D. and Hamner, B. (2013). 19 contrasting state-of-the-art automated scoring of essays. *Handbook of automated essay evaluation: Current applications and new directions*, pages 313–346.

Surya, K., Gayakwad, E., and Nallakaruppan, M. (2019). Deep learning for short answer scoring. *Int. J. Recent. Technol. Eng.(IJRTE)*, 7(6).