Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild

Samuel L. Pugh[1], Shree Krishna Subburaj[1], Arjun Ramesh Rao[1], Angela E.B. Stewart[2], Jessica Andrews-Todd[3], Sidney K. D'Mello[1]
[1]University of Colorado Boulder; [2]Carnegie Mellon University; [3]Educational Testing Service
samuel.pugh@colorado.edu; jandrewstodd@ets.org; sidney.dmello@colorado.edu

ABSTRACT
We investigated the feasibility of using automatic speech recognition (ASR) and natural language processing (NLP) to classify collaborative problem solving (CPS) skills from recorded speech in noisy environments. We analyzed data from 44 dyads of middle and high school students who used videoconferencing to collaboratively solve physics and math problems (35 and 9 dyads in classroom and school environments, respectively). Trained coders identified seven cognitive and social CPS skills (e.g., sharing information) in 8,660 utterances. We used a state-of-the-art deep transfer learning approach for NLP, Bidirectional Encoder Representations from Transformers (BERT), with a special input representation enabling the model to analyze adjacent utterances for contextual cues. We achieved a micro-average AUROC score (across seven CPS skills) of .80 using ASR transcripts, compared to .91 for human transcripts, indicating a decrease in performance attributable to ASR error. We found that the noisy school setting introduced additional ASR error, which reduced model performance (micro-average AUROC of .78) compared to the lab (AUROC = .83). We discuss implications for real-time CPS assessment and support in schools.

Keywords
Collaborative problem solving; natural language processing; collaborative interfaces

1. INTRODUCTION
The modern world will increasingly require teams of heterogeneous individuals to coordinate their efforts, share skills and knowledge, and communicate effectively in order to solve complex and pressing problems like the global pandemic and climate change. Accordingly, collaborative problem solving (CPS) – defined as two or more people engaging in a coordinated attempt to construct and maintain a joint solution to a problem [57] – has been identified as a critical skill for the 21st century workforce [23, 27]. Despite its increasing importance, the most recent 2015 Programme for International Student Assessment (PISA) assessment revealed troubling deficiencies in CPS competency worldwide [49]. As a result, improving CPS proficiency has become a priority in educational research and policy [7, 8, 16, 37, 49].

Technology has fundamentally transformed both the modern workplace and classroom. Co-located teams in shared spaces are becoming less common, while distributed teams that work and collaborate remotely through virtual interfaces are on the rise [22, 36]. In 2020, the COVID-19 pandemic thrust this issue to the forefront of our attention, as workers and students across the globe were forced to adapt to a remote environment for extended periods of time. Accordingly, educational practitioners have emphasized the importance of providing students with the skills necessary to effectively collaborate in virtual settings [60].

The rise of videoconferencing in both workplace and learning environments brings with it the exciting opportunity to develop next-generation collaborative interfaces that can aid in teaching, assessing, and supporting CPS. Here we focus on the task of assessing CPS skills from spoken language with an eye for downstream applications including reflective feedback and dynamic interventions to improve CPS skills.

Like any latent construct (e.g., intelligence, knowledge), assessment of CPS skills entails identifying objective evidence for those constructs. Because collaboration inherently involves communication, one promising approach is to analyze communication between team members [58]. Indeed, the content of communication during CPS provides information about a team's cognitive and affective states, knowledge, information sharing, and coordination [27], and can serve as evidence of relevant CPS skills [3, 4].

However, analyzing the large amounts of data generated during open-ended collaboration is time consuming and costly, requiring trained human coders to review large corpus and hand code individual items for indicators of CPS. Previous work [24, 29, 58, 65] has attempted to automate this coding process using natural language processing (NLP) techniques. However, with the exception of [65], this has been limited to restricted forms of communication such as text chat, rather than open-ended verbal communication, which is characteristic of most real world CPS. As we elaborate below, the one study [65] that successfully analyzed spoken communications for evidence of CPS skills used data collected in a highly controlled lab environment, leaving open the question as to whether this approach will succeed in the wild, such as in noisy classroom environments.

In this work, we address the challenge of using speech recognition and NLP to automatically analyze open-ended student speech during videoconferencing-enabled collaborative problem solving in both real-world schools and in lab environments. Pursuing technologies capable of automatically capturing and analyzing spoken language during open-ended verbal CPS in authentic environments, whether face-to-face or via videoconferencing, is an important avenue of research. These technologies hold the potential for significantly improving real-

time assessment and support of CPS [58], whether by providing teachers with feedback on CPS in student groups or enabling just-in-time interventions to steer groups of problem solvers in the right direction.

1.1 Background and Related Work

We first present a brief discussion on theoretical frameworks of CPS to situate the CPS skills modeled in this study within the CPS literature. Then, we discuss prior work on computational models of CPS, specifically focusing on language-based models.

1.1.1 Frameworks of CPS

CPS has been defined as problem solving activities that involve interactions among a group of individuals [47]. One early attempt to conceptualize CPS was by Roschelle and Teasley [57] who proposed a joint problem space model that emphasized shared understanding of the task as a central aspect of CPS. More recently, the Assessment and Teaching of Twenty-First Century Skills (ATC21S) framework [28, 30] described CPS through a measurable and teachable set of social and cognitive skills based on interaction, self-evaluation and goal setting. Relatedly, the PISA 2015 [49] framework conceptualized CPS as a complex process involving three collaborative dimensions that overlap with four problem-solving processes resulting in 12 CPS skills. Building on these frameworks, Sun et al. [68] proposed a generalized competency framework for CPS skills based on interactions among triads, which defines a hierarchical CPS model involving three high-level facets of CPS, each composed of sub-facets and associated behavioral indicators. Another approach, and the framework adopted in this work, is the in-task assessment framework [34]. Informed by principles of evidence-centered design [41], this framework characterizes CPS through a hierarchical ontology [3], which lays out theoretically-grounded, generalizable CPS skills along with behavioral indicators of these skills.

1.1.2 Computational Models of CPS

The stream of interactions generated during problem solving is considered the richest source of information about a team's knowledge, skills, and abilities [27, 38]. Accordingly, prior research has used non-verbal behavioral signals like facial expressions to detect rapport loss in small groups during open-ended discussions [43]. Multimodal combinations of facial expressions, acoustics and prosody, eye gaze, and task context have been explored to predict CPS outcomes like task performance [42, 67]. Additionally, learning gains [32, 50], subjective performance [72] and CPS competence [13, 14] have been modelled using multimodal signals.

Focusing our review on studies that explored the use of language and speech based data, researchers have successfully used language to model CPS processes like idea sharing [24, 29], negotiation [65], and argumentation [58], as well as CPS outcomes such as task performance [10, 44, 51] and learning gains [55]. A common NLP approach involves quantifying the frequency of words and word phrases (n-grams) [24, 29, 44, 54, 58]. Further, some research has experimented with the use of additional lexical features like punctuation [24, 29, 58], part-of-speech tags [21, 44, 58], or emoticons [29]. In addition to using lexical features from language itself, researchers have derived features from conversational data which index team and conversational dynamics (e.g., turn taking). This approach has been used to provide feedback on collaboration [59], identify sociocognitive roles [20], and model intra- and interpersonal dynamics [19] during CPS.

Closely related to our work, Hao et al. [29] used pre-selected n-grams and emoticons to model four CPS facets of sharing ideas, negotiating, regulating problem-solving activities, and maintaining communication. Their study involved data collected from 1000 participants with at least one year of college experience randomly grouped into dyads. They used a linear chain conditional random field and extracted lexical features from sequential text chats between dyads. They found that sequential modeling achieved an average accuracy of 73.2%, which outperformed a majority-class baseline accuracy of 29%, and slightly outperformed standard classifiers (accuracies of 66.9% to 71.9%).

Whereas the Hao study analyzed text-chats among dyads, Stewart et al. [65] modeled the three CPS facets of construction of shared knowledge, negotiation and coordination, and maintaining team function from spoken trialogues (conversations among triads). The study involved 32 triads of undergraduate students from a medium-sized private university, engaged in a 20-minute computer programming task using video conferencing software in a lab setting. They used ASR to generate transcripts of the team's speech during problem solving, from which they derived n-gram features for modeling. They obtained area under the receiver operating characteristic curve (AUROC) scores of .85, .77 and .77 for the three CPS facets using random forest classifiers, exceeding chance baselines of 0.5. In a follow-up study [66], they investigated whether including additional modalities (facial expression, acoustic-prosodic features, task context) in addition to language improved classification accuracy. They found that a combination of language and task context yielded slight improvement over unimodal language models.

1.2 Current Study and Novelty

There are several novel aspects of this work. First, although recent work [65, 66] has successfully used ASR and NLP to automatically analyze speech during CPS in the lab, it is currently unknown whether this approach can be effective in the wild, for example in noisy real-world classrooms where CPS interactions would occur. Lab environments have the advantage of being free from ambient noises, distractions from other students, and various other complicating factors present in school environments.

Further, previous work has been limited to adults, namely undergraduate students. However, given the importance of CPS, it is imperative that technologies be developed that can help instruct and support CPS in middle and high school-aged students. Therefore, a second important question is whether this approach can be applied to children, who may have differing CPS abilities and communication styles. An accompanying question is whether ASR can provide sufficiently accurate transcripts of children's speech, as research has documented the degradation of ASR performance on children's speech due to ASR systems primarily trained on adult speech, and age-dependent spectral and temporal variability in speech signals [26, 45, 53].

We address these questions by recording audio of remote CPS among middle and high school students in both the lab and

computer-enabled classrooms with multiple teams interacting. We show for the first time that in noisy school environments, ASR can provide transcripts of sufficient accuracy to model CPS skills. Additionally, we quantify the decrease in predictive accuracy that can be attributed to ASR error (vs. NLP error) by comparing with models trained on human transcripts, and comparing lab- vs. classroom- environments.

Finally, an open question in this domain is which NLP algorithms should be used to automatically analyze CPS language. We explore the use of deep transfer learning for this NLP problem. Recent advances in state-of-the-art NLP have been attained by adapting attention-based language models [71], pre-trained on large amounts of unlabeled data, to specific NLP tasks (e.g., text classification) [31]. We demonstrate the efficacy of this approach, using the popular Bidirectional Encoder Representations from Transformers (BERT) model [18] for our NLP task, and compare results with a more traditional n-gram approach using random forest classifiers. We also investigate whether a sequential classifier, which considers adjacent (i.e. previous, subsequent) utterances for contextual cues, yields improved performance over single utterance classifiers. We present a method, similar to the approaches used in [12, 69], to capture adjacent utterances for context by constructing a special input representation for the BERT model, which improves classification accuracy.

2. METHOD

2.1 Data Collection

2.1.1 Contexts

Our primary data collection occurred in one United States east coast public middle school and one public high school from the same district. The study was run over two data collection periods. The first period included 61 students in the high school and 44 students in the middle school. Here, students participated in two 43 minute class periods. The second collection included 18 students from the same middle school. Because we did not have control over the acoustic environment in the school context, we also collected supplementary data from 18 students in the lab. In the second collection, students completed one 90 minute session. In both collections, students in the school environment completed the study from a computer lab in the school in which other students were also participating in the study. Data collection occurred prior to the COVID-19 pandemic, and as such classrooms were at normal capacity. Students in both environments were equipped with a personal headset and microphone (MPOW 071 USB Headset).

2.1.2 Participants

In all, 141 middle and high school students (age range: 12-15) completed some or all of the study. However, only a subset of 74 sessions (a session entails one dyad completing one of the tasks) were included in this analysis. Participants were excluded for the following reasons: we experienced technical challenges on the first day of data collection, either team member did not complete a consent form, one team member did not show up, or there were quality issues with the recorded audio stream. Our analyzed dataset consisted of 88 students (65% female; mean age = 13.6, SD = 0.90). The lab subset contained 18 students (50% female; mean age = 13.6, SD = 1.01) and the school subset contained 70 students (69% female; mean age = 13.6, SD = 0.87). The sample of 88 students was quite diverse with 26.1% self-reporting as Black/African American, 19.3% Hispanic/Latino, 15.9% Multiracial, 13.6% Asian/Asian American, 12.5% White, 2.3% American Indian/Alaska Native, 6.8% reported "Other", and 3.4% did not report ethnicity.

2.1.3 CPS Tasks

The study involved two separate CPS tasks. In one task on linear functions and argumentation (T-Shirt Math Task [1]), students worked together through a series of task items in which they sought to determine which of three t-shirt companies was the best choice for a student council to purchase t-shirts for classmates. They compared three companies with differing variable costs (price per shirt) and fixed costs (upfront fee) to determine which company should be chosen given the number of t-shirts to be purchased. Individual questions included populating the cost equation y = mx + b according to the costs of each company (see Figure 1B), identifying the correct graph for a given company's cost equation, and providing a recommendation as to which company was the best deal. During this task, only one student controlled the screen at a time (i.e. to enter responses to the questions), and the two students could alternate control as they chose.

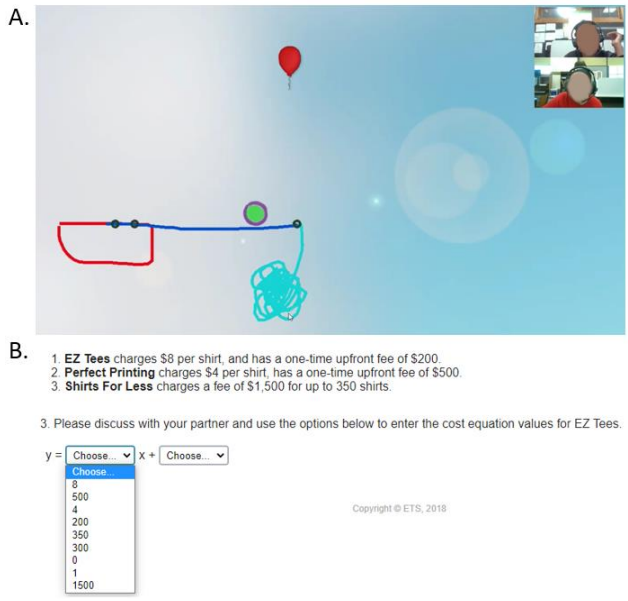

Figure 1. Screenshot examples of the videoconferencing setup and two CPS tasks. (A) Shows a level in Physics Playground, (B) shows a question from the T-Shirt Math Task (reproduced with permission from ETS).

The second task (Physics Playground [62]) was an educational physics game designed to help students learn concepts in Newtonian physics. In this task, students completed a series of six game levels in which they were tasked with drawing objects (e.g., lever, ramp, springboard) to guide a ball to hit a balloon target (see Figure 1A in which students are drawing a weight attached to the springboard to launch the ball towards the balloon). During this task, only one student controlled the game at a time. One student was selected to control first, and after

three levels had been completed (or half of the allotted time had elapsed), control was switched to the other student for the following three levels. Whereas the math task resembles more traditional school work and is more constrained by prior knowledge, the physics game provides more opportunities for creative exploration [35].

2.1.4 Procedure

Students were randomly assigned to pairs (27 mixed-gender, 17 same-gender pairs) and each student first individually completed a series of pre-surveys; details are not relevant here. Once both students in the pair completed the pre-surveys, a researcher enabled audio and video recording on each student's computer using Zoom video conferencing software (https://zoom.us) to record students' computer screens, faces, and voices. The student teams then worked together to complete the two CPS tasks, either on a different day or the same day (see above). The order of the tasks was counterbalanced so that half of the teams completed Physics Playground first and the other half completed the T-Shirt Math Task first. After completing each task students individually completed additional questionnaires not analyzed here.

2.2 CPS Ontology and CPS Skills

2.2.1 CPS Ontology (Framework)

We used a competency model represented as an ontology [3, 4] (similar to a concept map), which lays out the components of CPS and their relationships, along with indicators of CPS skills. The development of the ontology was based on discussions with subject matter experts as well as a literature review in relevant areas such as computer-supported collaborative learning, individual problem solving, communication, and linguistics [30, 39, 46, 48, 49, 64].

Our CPS ontology [3] includes nine high-level CPS skills across social and cognitive dimensions and sub-skills that correspond to each high-level skill. The social dimension includes four CPS skills: (1) *Maintaining communication* corresponds to content irrelevant social communications among teammates (e.g., greeting teammates or engaging in off-topic conversations); (2) *Sharing information* corresponds to task-relevant communication that is useful for solving the problem (e.g., sharing one's own knowledge, sharing the state of one's understanding); (3) *Establishing shared understanding* includes communication used to learn the perspectives of others and ensure that what has been said is understood by teammates (e.g., requesting information from teammates, providing responses that indicate comprehension); and (4) *Negotiating* corresponds to communication used to express agreement, express disagreement, or resolve conflicts that arise.

The cognitive dimension includes five CPS skills: (1) *Exploring and understanding* corresponds to communication and actions used to explore the environments in which teammates are working or understand the problem at hand (e.g., rereading problem prompts); (2) *Representing and formulating* includes communication used to build a mental representation of the problem and formulate hypotheses; (3) *Planning* corresponds to communication used to develop a plan for solving the problem (e.g., determining goals or establishing steps for carrying out a plan); (4) *Executing* corresponds to actions and communication used to carry out a plan (e.g., taking steps to carry out a plan, reporting to teammates what steps you are taking, or making suggestions to teammates about what steps they should take to carry out the plan); and (5) *Monitoring* includes communication used to monitor progress towards the goal or monitor teammates (e.g., checking the progress or status of teammates).

Table 1. The 7 CPS skills modeled, ordered from highest to lowest prevalence

CPS Skill	Base Rate	Dimension	Example Human Transcript	Corresponding ASR Transcript
Sharing Information	.26	Social	(Math) "Okay so first I think we should create like three equations to for each company"	"Okay Sir thank first we should create like three D creations for each arm company"
Establishing Shared Understanding	.25	Social	(Math) "Which one do you think is the best one"	"Twenty it's the best"
Negotiating	.16	Social	(Physics) "Umm no let's just do another idea I don't think it's gonna work anymore"	"Let's just do it another day I don't think it's going to work anymore"
Executing	.14	Cognitive	(Physics) "Okay and now put a weight down on that"	"Okay and now put a weight down on the"
Maintaining Communication	.07	Social	(Physics) "(laughs) Oh no this game is funny bro yeah I don't know what to do"	"This came funny I would like to do"
Monitoring	.06	Cognitive	(Physics) "That didn't work oh no"	"That didn't recall about"
Planning	.05	Cognitive	(Math) "Alright now we have to find a graph for this one now"	"Now we have to find a crusher this one now"

2.2.2 CPS Coding

Video recordings of student task sessions were segmented at the turn (or utterance) level and then coded by three trained raters using Dedoose qualitative analysis software [17]. For the coding, raters viewed each turn for each individual in a team and then labeled the turn as one of the CPS skills from the CPS ontology. To establish reliability, the three trained raters triple coded 20% of the videos. Intraclass correlations (ICCs) were used to estimate interrater reliability across rater judgments, as it can provide information about the consistency of the judgments among raters. The median ICC across the CPS skill ratings was .93, corresponding to excellent agreement [11].

Once reliability was established, the remaining videos were split among the three raters and coded independently. A total of 10,239 turns were coded across 80 CPS sessions with an average of 128 turns per session (SD = 70.5). Two CPS skills (exploring and understanding, and representing and formulating) occurred very infrequently (base rate < 1%) and were excluded from our analysis. The remaining seven CPS skills, with their base rate, cognitive/social dimension, and a sample utterance from the dataset, are shown in Table 1.

2.3 ASR and Human Transcript Generation

After segmenting and coding each utterance, we used the IBM Watson speech-to-text service [33] to generate ASR transcripts for each video. The service outputs transcripts with word-level start and stop times, as well as word-level confidence (between 0 and 1) for each word recognized. We constructed the transcript for each coded utterance by concatenating transcribed words within the utterance's human segmented time window. The confidence for each utterance was computed by taking the mean word confidence over all words in the utterance transcript. Utterances in which no words were recognized were assigned a confidence of 0. Because a single audio stream of each session was recorded (rather than individual audio streams from each student), the ASR transcripts can contain words from both speakers if there was overlap (elaborated below).

We also manually transcribed each utterance from the CPS videos. Human transcribers viewed the video segment (with audio) of each coded utterance and transcribed the words spoken by the indicated speaker (each utterance was coded for an individual student). Speech from the other student, if present in the segment, was not transcribed. Prior to transcription, guidelines were established among the human transcribers to ensure consistency in transcribing informal words or phrases (e.g., gonna, c'mon).

Because the segmented utterances sometimes contained speech from both speakers, we had alignment inconsistencies, as the ASR transcribed all words in a segment while the human transcripts only contained words spoken by the indicated student. To better assess ASR accuracy, we randomly sampled 10 utterances from each CPS session (8.5% of the data) and re-transcribed the utterances to include all words spoken in the segment, regardless of speaker. We refer to this as the Human Transcript Subset. We then computed a word error rate (WER) [9] for each utterance in this subset defined as (substitutions + insertions + deletions) / (words in human transcript), using the python package Jiwer [70].

2.4 Analyzed Dataset

Our dataset contains 74 CPS task sessions from 44 teams. This includes 30 teams with both the math and physics tasks in the dataset, nine teams with only the math task and five teams with only the physics task. 18 of the 74 sessions occurred in the lab, and the remaining 56 sessions occurred in school environments. The dataset consists of 8,660 utterances coded with CPS skills, and corresponding transcripts. Of these utterances, 2,751 (32%) were from lab sessions and the other 5,909 (68%) were from school sessions.

2.5 Machine Learning

We adopted a supervised classification approach to predict the ground truth CPS skill for each utterance. We first implemented a bag-of-n-grams approach using a Random Forest Classifier, as recent literature [65] has shown this method to be effective for the classification of CPS utterances. Next, we explored deep transfer learning as a means to improve upon this method. In particular, we leveraged pre-trained language models and employed the popular Bidirectional Encoder Representations from Transformers (BERT) model [18]. Additionally, we tested a method (BERT-seq) which takes a sequence of utterances as input (the utterance to classify plus the previous and subsequent utterances) to capture contextual information, in order to determine if including adjacent utterances improves classification accuracy. We trained separate models (RF, BERT, and BERT-seq) using the ASR transcripts and human transcripts as input.

2.5.1 Random Forest N-Grams

We first followed the approach outlined in [65] and trained Random Forest Classifiers to predict the CPS skill for each utterance using n-gram features. We used unigrams (words) and bigrams (two-word phrases) as the features for our Random Forest classifiers. Trigrams and beyond were not used since very few unique trigrams (only 6) occurred in >1% of utterances. We explored excluding n-grams that occurred at less than a minimum frequency in the training dataset, testing values of 0% (no filtering), 1% and 2% as hyperparameters. We used the scikit-learn [52] library's implementation of the Random Forest Classifier with 200 estimators.

2.5.2 BERT

We used a transfer learning approach and fine-tuned pre-trained BERT models to predict the CPS skill for each utterance. This entailed starting with a BERT model pre-trained on a large amount of unlabeled data, then fine-tuning it on our dataset of transcribed utterances and corresponding labels (CPS skills). We first processed the transcribed utterances using WordPiece tokenization [61]. This process entailed splitting an utterance into a sequence of words, or parts of words. Each unique word or word piece was then converted to an integer (called a token) according to BERT's pre-specified vocabulary. Finally, special tokens ([CLS] and [SEP]) were appended to the beginning and end of this sequence of integers and the sequence was provided as input to BERT (see Figure 2A). BERT mapped each input token to a 768-dimensional embedding, which serves as a semantic representation of the input token (the embedding of the special [CLS] and [SEP] tokens capture a semantic representation of the entire sequence of input tokens).

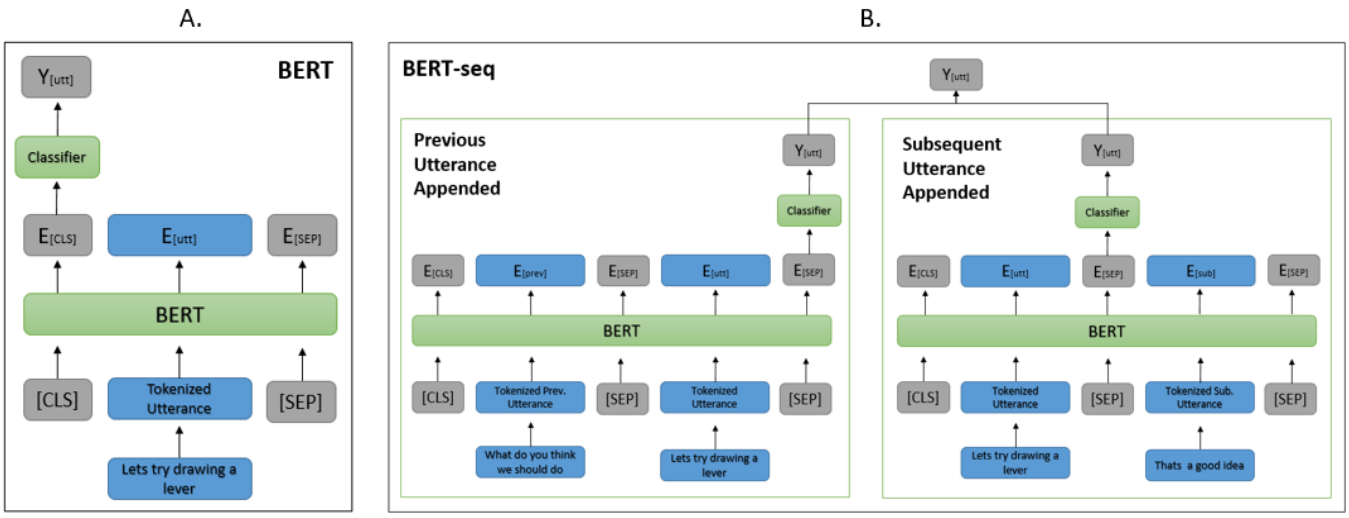

Figure 2. (A) The traditional BERT model used for text classification. (B) Our BERT-seq model which captures contextual information from the previous and subsequent utterances during classification.

For classification, the embedding of the [CLS] token was used as input to a fully connected layer (classifier), which output predicted probabilities for the seven CPS skills. We used multiclass learning, meaning that all seven CPS skills were predicted by one model.

2.5.3 BERT-seq

We propose a method to incorporate contextual utterances during classification by creating a special input representation, without augmenting the BERT architecture. This method takes a sequence of three utterances as input (the utterance to classify plus the previous and subsequent utterances), which are used to train two separate BERT models, each including either the previous or subsequent utterance in the BERT input (see Figure 2B). To add a pair of adjacent utterances to the input, we first processed each utterance individually using WordPiece tokenization as described above. The special [CLS] token was then added to the beginning of this sequence, and a [SEP] token was added to the end of both the first and second utterances. To classify the utterance, the embedding of the corresponding [SEP] token was used as input to a fully connected layer, which output predictions for the 7 CPS skills. Finally, the predicted probabilities of the previous and subsequent utterance models were averaged. This method of representing a sequence of utterances enables the self-attention layers of BERT to leverage contextual information from the previous and subsequent utterances, while still utilizing the pre-trained BERT weights.

For both BERT and BERT-seq we started with the transformers [73] library's implementation of the BertModel with the "bert-base-uncased" pre-trained weights, and used the BertTokenizer to process our utterances. We then fine-tuned the models for three epochs using a batch size of 16. We found that fine-tuning beyond three epochs did not substantially improve model performance.

2.5.4 Cross Validation

We used team-level 10-fold cross-validation to assess the accuracy of our classifiers. With our dataset of 44 teams, this entailed training a model with utterances from 90% of teams (39 or 40 teams), then evaluating the model's predictive accuracy on a test set containing utterances from the 10% of teams withheld during training (4 or 5 teams). This process was repeated ten times, such that every team appeared in the test set once. To compute accuracy metrics, predictions from all ten folds were aggregated and a single metric was computed on the full dataset. Team-level cross validation yields a better assessment of the method's generalizability to new teams because it ensures each model is never trained and evaluated on utterances from the same speaker. We used identical cross-validation folds for the RF, BERT and BERT-seq models as well as the human and ASR transcripts to ensure that differences in performance were not an artifact of the folds used. This experiment was repeated for 5 iterations, and different randomized cross-validation folds were used for each iteration.

3. RESULTS

3.1 ASR Accuracy

We compared WER in the lab and school subsets in order to quantify the speech recognition error that could be attributed to noisy school environments, as opposed to other factors such as difficulty recognizing children's speech, whispering or mumbling, audio quality, or inevitable ASR mistakes. We used the Human Transcript Subset as described in Section 2.3 for this comparison. The distributions of WER in the lab and school environments are shown in Figure 3. We found that WER was much lower in the lab environment than in schools (mean WER of .54 and .76, median WER of .50 and .91, respectively), indicating that significant ASR error is due to noisy school environments. We performed a non-parametric Kruskal-Wallis test [40] to statistically compare WER in the lab and school samples, and found that they differed significantly ($\chi^2(1) = 62.13$, $p < .001$).

As evident in Figure 3, a large proportion (47%) of the school utterances had a WER of 1 (compared to 19% for lab data), meaning no words were correctly recognized. However, WER was also high in the controlled lab environment, suggesting that speech recognition error may in part be attributable to factors beyond the complications of noisy school environments.

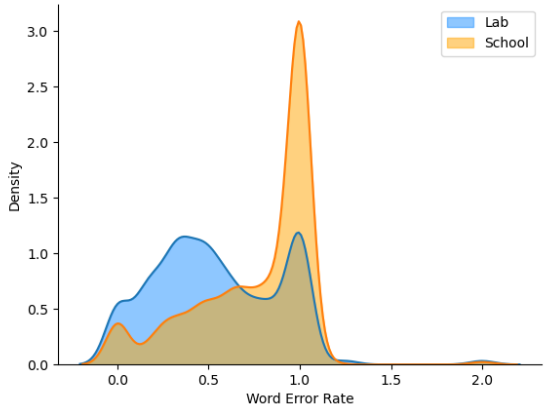

Figure 3. Gaussian kernel density estimates of the distribution of word error rates in the lab and school environments.

We also investigated the correlation between WER and ASR confidence to determine whether the confidence values produced by the ASR provided a good estimate of transcript accuracy. We found that WER and ASR confidence were significantly correlated (Spearman *rho* = -.74, $p < .001$).

3.2 Model Comparison

Next we compared the performance of our three NLP models (RF, BERT, BERT-seq). The models output a probability from 0 to 1 that an utterance is coded with each CPS skill. Accordingly, we report the area under the receiver operating characteristic curve (AUROC) for each skill, a common accuracy metric for model performance [6] which takes into account the true positive and false positive tradeoff across classification thresholds. Mean AUROC scores (over the five iterations) for the RF, BERT and BERT-seq models, using both human and ASR transcripts are reported in Table 2. We also report a chance baseline, created by randomly shuffling the labels within each CPS session and computing accuracy accordingly. Because shuffling is within sessions, the AUROCs for the shuffled models will slightly deviate from the 0.5 chance baseline. To determine if the three model's AUROC scores were significantly different for each CPS skill, we used a bootstrap method to statistically compare the AUROC values. Since five iterations of this experiment were conducted, we selected the model corresponding to the median AUROC value across the five iterations (for both human and ASR transcripts) on each CPS skill for statistical analysis. We performed this analysis in R using the pROC package [56] with 2,000 bootstrap permutations. Finally, we adjusted the resulting p-values using a false discovery rate (FDR) correction [5] to account for multiple testing across the seven CPS skills.

Without exception BERT-seq quantitatively yielded the highest AUROC scores for all seven CPS skills using both human and ASR transcripts, indicating that our method of incorporating adjacent utterances improves performance over single utterance classifiers. On average, BERT outperformed the RF model on both human and ASR transcripts, although there were some skills for which the RF AUROC scores were higher. From the statistical analysis described above, we found that with ASR transcripts BERT-seq had a significant advantage over the other two models for most skills (four of seven for BERT, five of seven for RF). We also found that there was no significant difference between BERT and RF for six of seven skills.

Table 2. Mean AUROC values (across 5 iterations) of the RF N-gram, BERT, and BERT-seq models on ASR and Human transcripts for all CPS skills.

CPS Skill	ASR Transcripts			Human Transcripts			
	RF	**BERT**	**BERT-seq**	**RF**	**BERT**	**BERT-seq**	**Shuffled**
Sharing Information	0.711	0.745 ^{R}	0.756 ^{R}	0.837	0.866 ^{R}	0.877 ^{R}	0.540
Establishing Shared Understanding	0.713	0.724	0.740 ^{RB}	0.872	0.894 ^{R}	0.907 ^{RB}	0.509
Negotiating	0.721	0.719	0.741 ^{B}	0.896	0.901	0.916 ^{RB}	0.510
Executing	0.745	0.767	0.784 ^{R}	0.897	0.914 ^{R}	0.926 ^{R}	0.574
Maintaining Communication	0.673	0.667	0.750 ^{RB}	0.849	0.853	0.901 ^{RB}	0.557
Monitoring	0.632	0.594	0.677 ^{RB}	0.812	0.792	0.843 ^{RB}	0.513
Planning	0.700	0.692	0.718	0.861 ^{B}	0.818	0.872 ^{B}	0.502
Micro Avg.	0.773	0.782	0.799	0.887	0.895	0.914	0.607

^{R} and ^{B} indicate the AUROC score was significantly higher than the RF and/or BERT models, respectively. Neither RF nor BERT ever outperformed BERT-seq.

We observed a similar pattern on the human transcripts, where BERT-seq significantly outperformed BERT on five of seven skills and RF on six of seven skills. Interestingly, on human transcripts the advantage of BERT over RF increased, with BERT having significantly higher scores on three skills, while RF was significantly better on only one. This finding suggests that with high quality transcripts which accurately capture the content of an utterance, BERT was the better model, whereas with noisy ASR transcripts there was no clear difference.

These results indicate that BERT-seq quantitatively outperformed both the traditional BERT and the RF n-gram approach for all seven CPS skills, using both the human and ASR transcripts. However, the statistical analysis revealed that for some CPS skills, this advantage was not statistically significant. As BERT-seq was the best model across CPS skills, we refer to these results in our comparison of human and ASR transcripts, and throughout the rest of this paper.

3.3 ASR vs. Human Transcripts

We found that using the ASR transcripts as input, our best model (BERT-seq) was able to accurately classify the seven CPS skills, yielding a micro-average AUROC score of .799. However, when the human transcripts were used, this average increased to .914 (see Table 2). We compared the human and ASR transcript results using the bootstrap method described above, and found that the human transcript AUROC scores were significantly (FDR corrected $p < .05$) higher than the ASR transcript scores for all seven CPS skills, an unsurprising result given the high word error rates in the ASR transcripts. However, we note that despite significant loss in performance due to speech recognition error, our model easily outperformed a shuffled baseline (micro-average AUROC of .607), supporting the hypothesis that CPS skills can be automatically predicted from ASR transcripts.

3.4 Classification Accuracy in Lab and School Environments

Next we compared classification accuracy in the lab and school environments in order to investigate the extent to which higher rates of ASR error in the school subset affected model performance. We report AUROC scores for the lab and school environments in Table 3. We found that on average, classification accuracy was substantially lower in the school subset compared to the lab subset (micro-average AUROC of .783 and .830, respectively). Further, for every individual skill, AUROC scores were quantitatively higher in the lab subset than in the school subset, with differences in AUROC values for individual skills ranging from .031 (Executing) to .102 (Negotiating). We again used the bootstrap method to statistically compare AUROC scores in the lab and school for each skill and found that scores were significantly higher in the lab subset for five out of seven CPS skills (see Table 3).

3.5 Classification Accuracy as a Function of ASR Confidence

Lastly, we examined the relationship between ASR confidence and classification accuracy. As discussed in section 3.1, the ASR confidence is a good proxy for word error rate, as the two values are significantly correlated. Therefore, we separated our 8,660 utterances into ten ASR confidence bins (0.0 – 0.1, etc.) and computed the micro-average AUROC score for each bin. The distribution of utterances and corresponding AUROC scores for each bin are shown in Figure 4A and 4B, respectively. Figure 4B also shows the human transcript AUROC score as a benchmark of the accuracy that would be expected under conditions of near-perfect speech recognition. The shuffled baseline is also shown to visualize improvement over chance.

Table 3. Mean AUROC scores (across 5 iterations) for each CPS skill in Lab and School environments. Results are from the BERT-seq model using ASR transcripts. Values marked with * were significantly higher in the Lab vs. School.

CPS Skill	Lab		School	
	AUC	**Base Rate**	**AUC**	**Base Rate**
Sharing Information	0.782*	.25	0.743	.27
Establishing Shared Understanding	0.786*	.26	0.716	.25
Negotiating	0.807*	.18	0.705	.15
Executing	0.804	.15	0.773	.13
Maintaining Communication	0.803*	.03	0.717	.08
Monitoring	0.701	.05	0.663	.07
Planning	0.760*	.06	0.688	.04
Micro Avg.	0.830		0.783	

We found that a large proportion of utterances (20%) fall in the [0.0 - 0.1) bin, indicating that the ASR had little to no confidence in their content. In fact, nearly all (97%) of the utterances in this bin have an empty ASR transcript, meaning no words were recognized during the utterance's segmented time window. In many cases, this occurred due to the students whispering or mumbling, which the ASR was unable to recognize. Excepting the significant zero inflation, the utterances appeared to be normally distributed around the [0.6 - 0.7) bin.

We observed a strong correlation between ASR confidence bin and classification accuracy (Spearman $rho = .94$, $p < .001$). Unsurprisingly, we found that for low confidence transcripts (< 0.3) a substantial gap exists between the ASR transcript AUROC score and the benchmark human transcript score (see Figure 4B). On these low confidence transcripts, model performance is near the shuffled chance baseline. Interestingly, despite many (77%) of these low confidence transcripts containing no words, the model was still able to outperform the chance baseline by learning the distribution of skills among empty transcripts in the training data. We found that accuracy increases steadily among the medium confidence transcripts (0.3 - 0.7). For high confidence transcripts (≥0.7), AUROC scores are near (though still lower than) the benchmark human transcript values. The

relationship between ASR confidence and classification accuracy indicates that it might be viable to filter out utterances with low confidence to improve reliability for downstream applications.

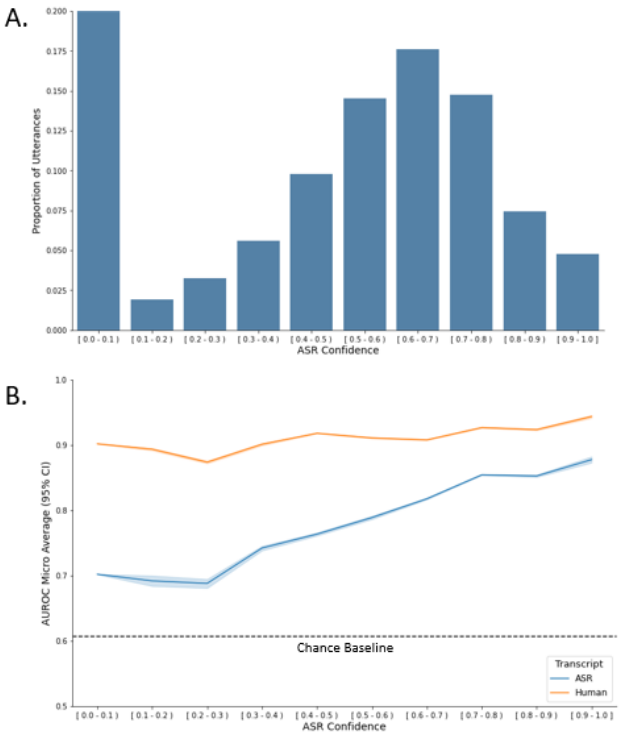

Figure 4. (A) Distribution of ASR confidence on all 8,660 utterances. (B) Model accuracy as a function of ASR confidence. Micro-average AUROC scores across the 7 CPS skills (with 95% CI across 5 iterations) are plotted for Human and ASR transcripts.

4. DISCUSSION

We investigated the feasibility of using automatic speech recognition and natural language processing to automatically classify student speech with CPS skills using data collected in both lab and real-world school environments. We compared performance using imperfect ASR transcripts with human transcripts, investigated differences between the lab and school environments, and explored three NLP approaches including bag-of-n-grams and deep transfer learning. In the rest of this section we discuss our main findings, applications of our models, as well as limitations and future directions of research.

4.1 Main Findings

We found that it is feasible to use ASR to transcribe middle and high school student's speech during CPS in both lab and school environments. However, we found that significant speech recognition error is introduced when speech is recorded in schools (mean WER of .76), likely as a result of noisy environments and distractions from other students. That said, speech recognition error was also high in the lab environment (mean WER of .54), suggesting that there may still be fundamental limitations associated with using ASR on children's speech in the context of remote CPS.

Despite imperfect speech recognition, we demonstrated that it is possible to automatically predict CPS skills from student speech in a real-world school environment. We built team-independent models that were able to predict CPS skills with reasonable accuracy (micro-average AUROC of .80) using ASR transcripts. Importantly, this result outperformed a shuffled baseline (micro-average AUROC of .61) by a significant margin. This finding is encouraging because it was previously unknown whether ASR could yield transcripts of sufficient quality to model CPS skills in noisy environments. Further, we demonstrated that by using high-fidelity human transcripts, this accuracy could be significantly improved (micro-average AUROC of .91). We demonstrated that in the absence of ASR error our NLP models were highly accurate, suggesting a useful upper bound of what can be achieved from spoken content alone.

We also improved upon NLP approaches previously used in CPS literature, demonstrating the advantage of deep transfer learning over standard classifiers for modeling CPS language. We found that on average, using both ASR and human transcripts, the deep transfer learning model (BERT) achieved slightly better accuracy than the Random Forest n-gram model (though the two were statistically tied for 3/7 CPS skills with human transcripts and 6/7 skills with ASR transcripts). This finding was unsurprising given that pre-trained language models have achieved state-of-the-art performance on many NLP benchmark tasks, including text classification.

Importantly, we found that we were able to further improve classification accuracy by constructing an input representation that enables BERT to capture information from adjacent utterances. This method showed significant improvement over the single utterance BERT and RF models, providing preliminary evidence of its viability. This finding suggests that in CPS, the context of an utterance (what was said before and after) may be important for accurate identification of particular CPS skills.

Finally, we examined the relationship between ASR confidence – a proxy for transcription quality – and classification accuracy. We found that the two were highly correlated, suggesting that downstream applications may be able to improve reliability of predictions by filtering out low confidence transcripts.

4.2 Applications

A key application of this work is the automatic assessment of CPS skills from open-ended speech in classrooms and beyond. As previously discussed, analyzing verbal communication for evidence of CPS skills is a costly and time-intensive process when trained human coders are used. Our findings suggest that automated methods using ASR and NLP may provide a viable alternative to the human-coding process. These automated methods hold great potential in improving the assessment and training of CPS skills, a priority of modern education [49]. However, given the imperfect accuracy of our models, and unanswered questions regarding how this approach may generalize to students with differing communication styles or cultural and linguistic backgrounds, this approach should be limited to formative assessment [63] focused on learning and improvement, rather than evaluation.

Our approach could advance this goal in several ways. For example, automatically generated reports could be sent to a teacher monitoring many groups of students engaged in CPS, informing the teacher of the extent to which each group is demonstrating CPS skills. Such a system could help the teacher

identify which groups need support and allocate their limited presence toward assisting those groups. Similarly, these reports could be used to identify individual student's strengths and weaknesses, and set appropriate goals for improvement. For instance, a student who frequently shares information yet seldom engages in negotiation or establishing shared understanding could be encouraged to listen to the ideas of their teammates and work to build on those ideas together.

In addition to passive assessment and off-line feedback, this approach could be leveraged by next-generation intelligent systems that actively monitor ongoing CPS and dynamically intervene in real time to yield improved CPS outcomes [15], or provide personalized on-line feedback to students. For example, a group frequently engaging in off-topic conversation could be prompted by the system to focus back on the problem-solving task, or a particular student within a group who hasn't shared information could be encouraged to share their ideas with the team. The specific intervention strategies, including when to intervene, how to present the intervention, and who the intervention should be targeted at (whole group vs. individual student) await design, testing, and refinement.

Importantly, a technology devised to assist in the training and assessment of CPS does little good if it is confined to the lab. Thus, the present results take a step towards the development of a system that can support CPS in real-world classrooms by monitoring open-ended verbal communication for CPS skills.

4.3 Limitations

There were some limitations of this work. First, although we used an automated approach for utterance transcription and CPS skill prediction, the sessions were segmented into utterances beforehand by human coders. This is a limitation because a fully automated pipeline would require the ASR to automatically detect and segment recorded speech into individual utterances, an already difficult task that may be further complicated by noisy school environments or the peculiarities of children's speech. Another related limitation is that due to the utterance segmentation and ASR transcription process we used, our ASR transcripts contain all speech that was recognized during an utterance's segmented time window. This means that some ASR transcripts contain words from both speakers, which introduces alignment inconsistencies between the ASR transcript and the coded CPS skill because utterances were coded at the individual student level. In particular, this introduces noise into the ASR transcripts when student's utterances overlap.

Another limitation of this work is that we considered only linguistic features to predict the coded CPS skills. We expect that model performance can be improved by modeling not only what students say (language), but considering how they say it (acoustic-prosodic information) and in the context of what they're doing (task-specific information). We hypothesize that the inclusion of these additional modalities may particularly improve performance for low confidence ASR transcripts, where the language transcribed by the speech recognizer is either missing altogether, or is a poor representation of what was actually said. Finally, although we demonstrated that our method for capturing contextual information from adjacent utterances improved accuracy, we did not compare this with other methods for incorporating contextual utterances such as conditional random fields or recurrent neural networks.

4.4 Future Work

The findings and limitations discussed in this section present several possibilities for improvement in future research. First, in order to develop a fully automated approach for modeling CPS skills, we plan to incorporate automatic utterance segmentation and speaker diarization into our ASR pipeline. Further, we plan to explore methods for incorporating information from other modalities in addition to language. For instance, including features such as acoustic-prosodic information, task context, facial expression, or body movement may enable more accurate prediction of CPS skills in cases where ASR fails to capture the content of an utterance.

Another direction of future research involves further exploration of how contextual utterances can be used to improve classification accuracy. We demonstrated a method for incorporating adjacent utterances in our model input, which improved performance over single utterance classifiers. In future work, we will explore methods for capturing contextual information beyond the previous and subsequent utterances (e.g., the five previous utterances). We also plan to investigate how the approach demonstrated in this paper, which leverages the model's attention mechanism to capture context, compares with other approaches (e.g., recurrent neural networks).

In addition to exploring methods for improving the accuracy of our models, we plan to investigate the utility of our CPS models. An open question is how accurate model predictions need to be to provide useful and actionable estimates for assessment, feedback, or intervention. Specifically, recent work [2, 25] has clustered students using the frequency of CPS skills to derive theoretically grounded profiles of collaborative problem solvers (e.g., active collaborators, social loafers). We plan to investigate whether model-derived estimates of CPS skill frequencies will yield high agreement to the clustering produced using human codes.

5. CONCLUSION

We combined automatic speech recognition and natural language processing to automatically predict CPS skills from student speech during problem solving in both lab and real-world school environments. Our findings suggest that despite significant speech recognition error in school environments, it is possible to predict expert-coded CPS skills using automatically generated transcripts. These findings open many possibilities for next-generation technologies that can further the goal of improved CPS training, assessment, and support in schools.

6. ACKNOWLEDGMENTS

This research was supported by the Institute of Educational Sciences (IES R305A170432), the NSF National AI Institute for Student-AI Teaming (iSAT) (DRL 2019805) and NSF DUE 1745442/1660877. The opinions expressed are those of the authors and do not represent views of the funding agencies.

7. REFERENCES

[1] Andrews-Todd, J. et al. 2019. Collaborative Problem Solving Assessment in an Online Mathematics Task. *ETS Research Report Series*. 2019, 1 (2019). DOI:https://doi.org/10.1002/ets2.12260.

[2] Andrews-Todd, J. et al. 2018. Identifying profiles of collaborative problem solvers in an online electronics environment. *Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018* (2018).

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

[4] Andrews-Todd, J. and Kerr, D. 2019. Application of Ontologies for Assessing Collaborative Problem Solving Skills. *International Journal of Testing*. 19, 2 (2019). DOI:https://doi.org/10.1080/15305058.2019.1573823.

[5] Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 57, 1 (1995). DOI:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[6] Bradley, A.P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*. 30, 7 (1997). DOI:https://doi.org/10.1016/S0031-3203(96)00142-2.

[7] C. Graesser, A. et al. 2018. Challenges of Assessing Collaborative Problem Solving. *Care, E., Griffin, P., Wilson, M. (Eds.), Assessment and teaching of 21st century skills: Research and applications*. 75–91.

[8] Care, E. et al. 2016. Assessment of Collaborative Problem Solving in Education Environments. *Applied Measurement in Education*. 29, 4 (2016). DOI:https://doi.org/10.1080/08957347.2016.1209204.

[9] Chen, S. et al. 1998. Evaluation metrics for language models. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. (1998).

[10] Chopade, P. et al. 2019. CPSX: Using AI-Machine Learning for Mapping Human-Human Interaction and Measurement of CPS Teamwork Skills. *2019 IEEE International Symposium on Technologies for Homeland Security, HST 2019* (2019).

[11] Cicchetti, D. V. 1994. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*. 6, 4 (1994). DOI:https://doi.org/10.1037/1040-3590.6.4.284.

[12] Cohan, A. et al. 2020. Pretrained language models for sequential sentence classification. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (2020).

[13] Cukurova, M. et al. 2020. Modelling collaborative problem-solving competence with transparent learning analytics: Is video data enough? *ACM International Conference Proceeding Series* (2020).

[14] Cukurova, M. et al. 2018. The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers and Education*. 116, (2018). DOI:https://doi.org/10.1016/j.compedu.2017.08.007.

[15] D'Mello, S. et al. 2019. Towards dynamic intelligent support for collaborative problem solving. *CEUR Workshop Proceedings* (2019).

[16] von Davier, A.A. et al. 2017. Interdisciplinary research agenda in support of assessment of collaborative problem solving: lessons learned from developing a Collaborative Science Assessment Prototype. *Computers in Human Behavior*. 76, (2017). DOI:https://doi.org/10.1016/j.chb.2017.04.059.

[17] Dedoose version 8.0.35 2018. Dedoose: Web applicaiton for managing, analyzing, and presenting qualitative and mixed method research data. *SocioCultural Research Consultants, LLC*.

[18] Devlin, J. et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (2019).

[19] Dowell, N.M.M. et al. 2020. Exploring the relationship between emergent sociocognitive roles, collaborative problem-solving skills, and outcomes: A group communication analysis. *Journal of Learning Analytics*. 7, 1 (2020). DOI:https://doi.org/10.18608/jla.2020.71.4.

[20] Dowell, N.M.M. et al. 2019. Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior Research Methods*. 51, 3 (2019). DOI:https://doi.org/10.3758/s13428-018-1102-z.

[21] Emara, M. et al. 2021. Examining Student Regulation of Collaborative, Computational, Problem-Solving Processes in Open-Ended Learning Environments. *Journal of Learning Analytics*. 8, 1 (2021). DOI:https://doi.org/10.18608/jla.2021.7230.

[22] Felstead, A. and Henseke, G. 2017. Assessing the growth of remote working and its consequences for effort, well-being and work-life balance. *New Technology, Work and Employment*. 32, 3 (2017). DOI:https://doi.org/10.1111/ntwe.12097.

[23] Fiore, S.M. et al. 2018. Collaborative problem-solving education for the twenty-first-century workforce. *Nature Human Behaviour*.

[24] Flor, M. et al. 2016. Automated classification of collaborative problem solving interactions in simulated science tasks. (2016).

[25] Forsyth, C. et al. 2020. Are You Really A Team Player ? Profiling of Collaborative Problem Solvers in an Online

Environment. *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020*. Edm (2020).

[26] Gerosa, M. et al. 2009. A review of ASR technologies for children's speech. *Proceedings of the 2nd Workshop on Child, Computer and Interaction, WOCCI '09* (2009).

[27] Graesser, A.C. et al. 2018. Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*. 19, 2 (2018), 59–92. DOI:https://doi.org/10.1177/1529100618808244.

[28] Griffin, P. et al. 2012. The changing role of education and schools. *Assessment and teaching of 21st century skills*.

[29] Hao, J. et al. 2017. CPS-Rater: Automated Sequential Annotation for Conversations in Collaborative Problem-Solving Activities. *ETS Research Report Series*. 2017, 1 (2017). DOI:https://doi.org/10.1002/ets2.12184.

[30] Hesse, F. et al. 2015. A Framework for Teachable Collaborative Problem Solving Skills. *Assessment and Teaching of 21st Century Skills*.

[31] Howard, J. and Ruder, S. 2018. Universal language model fine-tuning for text classification. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* (2018).

[32] Huang, K. et al. 2019. Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining* (2019).

[33] IBM Watson: *https://www.ibm.com/watson/services/speech-to-text/*. Accessed: 2021-03-02.

[34] Kerr, D. et al. 2016. The In-Task Assessment Framework for Behavioral Data. *The Handbook of Cognition and Assessment*.

[35] Kim, Y.J. and Shute, V.J. 2015. Opportunities and Challenges in Assessing and Supporting Creativity in Video Games. *Video Games and Creativity*.

[36] Kniffin, K.M. et al. 2020. COVID-19 and the Workplace: Implications, Issues, and Insights for Future Research and Action. *American Psychologist*. (2020). DOI:https://doi.org/10.1037/amp0000716.

[37] Koenig, J.A. 2011. *Assessing 21st Century Skills: Summary of a Workshop*.

[38] Lai, E. et al. 2017. *Skills for today: What We Know about Teaching and Assessing Collaboration*.

[39] Liu, L. et al. 2015. A tough nut to crack: Measuring collaborative problem solving. *Handbook of Research on Technology Tools for Real-World Skill Development*.

[40] McKight, P.E. and Najab, J. 2010. Kruskal-Wallis Test. *The Corsini Encyclopedia of Psychology*.

[41] Mislevy, R.J. et al. 2003. Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research & Perspective*. 1, 1 (2003). DOI:https://doi.org/10.1207/s15366359mea0101_02.

[42] Miura, G. and Okada, S. 2019. Task-independent multimodal prediction of group performance based on product dimensions. *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction* (2019).

[43] Müller, P. et al. 2018. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. *International Conference on Intelligent User Interfaces, Proceedings IUI* (2018).

[44] Murray, G. and Oertel, C. 2018. Predicting group performance in task-based interaction. *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction* (2018).

[45] Narayanan, S. and Potamianos, A. 2002. Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing*. 10, 2 (2002). DOI:https://doi.org/10.1109/89.985544.

[46] O'Neil, H.F.. C.G.K.W.K.. B.R.S. 1995. *Measurement of teamwork processes using computer simulation (CSE Tech. Rep. No. 399).*

[47] O'neil, H.F. et al. 2010. Computer-based feedback for computer-based collaborative problem solving. *Computer-Based Diagnostics and Systematic Analysis of Knowledge*.

[48] OECD 2013. *PISA 2012 Assessment and Analytical Framework: Mathematics, reading, science, problem solving and financial literacy*.

[49] OECD 2015. Pisa 2015 Collaborative Problem Solving Framework. (2015).

[50] Olsen, J.K. et al. 2020. Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology*. 51, 5 (2020). DOI:https://doi.org/10.1111/bjet.12982.

[51] Oviatt, S. and Cohen, A. 2013. Written and multimodal representations as predictors of expertise and problem-solving success in mathematics. *ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction* (2013).

[52] Pedregosa, F. et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 12, (2011).

[53] Potamianos, A. and Narayanan, S. 2003. Robust Recognition of Children's Speech. *IEEE Transactions on Speech and Audio Processing*. 11, 6 (2003). DOI:https://doi.org/10.1109/TSA.2003.818026.

[54] Prata, D.N. et al. 2009. Detecting and understanding the impact of cognitive and interpersonal conflict in computer supported collaborative learning environments. *EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining* (2009).

[55] Reilly, J.M. and Schneider, B. 2019. Predicting the

quality of collaborative problem solving through linguistic analysis of discourse. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining* (2019).

[56] Robin, X. et al. 2011. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 12, (2011). DOI:https://doi.org/10.1186/1471-2105-12-77.

[57] Roschelle, J. and Teasley, S.D. 1995. The Construction of Shared Knowledge in Collaborative Problem Solving. *Computer Supported Collaborative Learning*.

[58] Rosé, C. et al. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*. 3, 3 (2008). DOI:https://doi.org/10.1007/s11412-007-9034-0.

[59] Samrose, S. et al. 2018. CoCo: Collaboration Coach for Understanding Team Dynamics during Video Conferencing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 1, 4 (2018). DOI:https://doi.org/10.1145/3161186.

[60] Schulze, J. and Krumm, S. 2017. The "virtual team player": A review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. *Organizational Psychology Review*. 7, 1 (2017). DOI:https://doi.org/10.1177/2041386616675522.

[61] Schuster, M. and Nakajima, K. 2012. Japanese and Korean voice search. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2012).

[62] Shute, V.J. et al. 2013. Assessment and learning of qualitative physics in Newton's playground. *Journal of Educational Research*. 106, 6 (2013). DOI:https://doi.org/10.1080/00220671.2013.832970.

[63] Shute, V.J. 2008. Focus on formative feedback. *Review of Educational Research*. 78, 1 (2008). DOI:https://doi.org/10.3102/0034654307313795.

[64] Spada, H. et al. 2005. A new method to assess the quality of collaborative process in CSCL. *Computer Supported Collaborative Learning 2005: The Next 10 Years - Proceedings of the International Conference on Computer Supported Collaborative Learning 2005, CSCL 2005* (2005).

[65] Stewart, A.E.B. et al. 2019. I say, you say, we say: Using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving. *Proceedings of the ACM on Human-Computer Interaction*. 3, CSCW (2019). DOI:https://doi.org/10.1145/3359296.

[66] Stewart, A.E.B. et al. 2021. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction*. (2021). DOI:https://doi.org/10.1007/s11257-021-09290-y.

[67] Subburaj, S.K. et al. 2020. Multimodal, Multiparty Modeling of Collaborative Problem Solving Performance. *ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction* (2020).

[68] Sun, C. et al. 2020. Towards a generalized competency model of collaborative problem solving. *Computers and Education*. 143, (2020). DOI:https://doi.org/10.1016/j.compedu.2019.103672.

[69] Suresh, A. et al. 2021. Using Transformers to Provide Teachers with Personalized Feedback on their Classroom Discourse: The TalkMoves Application. *AAAI Spring Symposium Series 2021*.

[70] Vaessen N 2019. Word error rate for automatic speech recognition. https: //pypi.org/project/jiwer/.

[71] Vaswani, A. et al. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[72] Vrzakova, H. et al. 2020. Focused or stuck together: Multimodal patterns reveal triads' performance in collaborative problem solving. *ACM International Conference Proceeding Series* (2020).

[73] Wolf, T. et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv*.