Automatic Detection of Collaborative States in Small Groups Using Multimodal Features

No Author Given

No Institute Given

Abstract. Cultivating collaborative problem solving (CPS) skills in educational settings is critical in preparing students for the workforce. Monitoring and providing feedback to all groups is intractable for teachers in traditional classrooms but is potentially scalable with an AI agent who can see, listen, and interact with groups. For this to be feasible, CPS moves need to first be detected, a difficult task even in constrained environments. In this paper, we detect CPS facets in relatively unconstrained contexts: an in-person group task where students freely move, interact, and manipulate physical objects. We collected a novel audiovisual dataset of ten groups engaged in the task. Paralleling a real-world implementation, we automatically identified when individuals were speaking and segmented speech into utterances. Each utterance was labeled with a corresponding CPS facet. Then, multimodal machine learning models were trained to automatically identify the CPS facets using lingustic, visual, and prosodic features. We extracted linguistic features from Google automatic speech transcriptions of the utterances (as BERT embeddings), prosodic features using openSMILE, and visual features such as body pose. Our best multimodal model classified CPS states with an average weighted AUROC of .812 (SD = .030) across groups, thus establishing a state-of-the-art baseline for CPS detection on this dataset. This is the first work to classify CPS in an unconstrained *shared* physical environment using multimodal features. Further, this lays the groundwork for employing such a solution in a classroom context, and establishes a cornerstone technique that integrates with classroom agents for group work.

Keywords: Collaborative Problem Solving · Multimodal · Natural Language Processing · Computer Vision · Small Groups

1 Introduction

Working in teams is an essential skill in the workforce, which the education system needs to prepare students for. Such practices have been formalized into pedagogical techniques of *collaborative problem solving* (CPS) wherein students learn by working together to complete tasks, explore scenarios, and solve problems in them. With CPS, peers can develop practices of inquiry together and develop a "positive interdependence" [12, 15, 20], but doing so depends on having

an effective group dynamic that does not fall into dysfunction. With proper facilitation, this can be avoided, but this role usually falls to the teacher, and with a single teacher and many small groups, such facilitation becomes intractable.

For some group-facilitation tasks, such as aligning group goals and member responsibilities, an artificially intelligent agent can be a useful tool to help teachers manage groups. A prerequisite to agent interaction with a group is that the agent must be able to observe the group and detect its state. This requires consuming live input — which may include both verbal and nonverbal features — as well as some measurement of group performance. One way this has been approached is by detecting CPS skills in group interactions, such as in [21]. This work showed promising results in detection of CPS facets as defined by Sun et al. [22]. One limitation of this work is that the groups were virtual, with each person on a separate video and audio stream. In an in-person situation, there will be other challenges, such as overlapping audio [4], communication through gestures, and multi-user interaction with shared objects.

In this study, we approach the problem of detecting CPS using indicators defined by Sun et al. [22] in an in-person setting using a novel small group task. We collect a dataset of triads working together to solve a shared task, automatically extract verbal, prosodic, and visual features, and train multiple types of machine learning models over this data to detect CPS facets. We observe that CPS facets can be successfully detected over an in-person task using automatically extracted features. Both traditional machine learning methods (AdaBoost and random forest) and deep learning are successful. However, the traditional methods generally perform better without the visual features, while the deep learning approach sees improvement with the added visual features.

2 Related Work

There is a wealth of prior research studying the different dimensions of collaborative problem solving (CPS), from definitional frameworks, tasks that showcase and depend on successful CPS, evaluation, and individual empirical features correlated with CPS skills. In this section, we review multiple prior approaches to studying CPS.

Cukurova et al. [6] discuss the technical and social challenges inherent in modeling learning analytics using multimodal data and present a framework for modeling it using non-verbal behavior [7]. Andrews-Todd and Forsyth [1] defined aspects of CPS according to social indicators and cognitive indicators — these definitions accounted for both the task-specific components of group work, as well as the interrelational aspects of group work among members of the groups. However, Sun *et al.* [23] argued that separating these characteristics unintentionally obscures the fact that each component may occur simultaneously. For example, a group member asking a clarifying question is both *pursuing shared knowledge* and *clarifying the task*. They define an alternative framework composed of three main facets: *construction of shared knowledge*, *negotiation/coordination*, and *maintaining team function*, which are defined by *sub-facets* (e.g., "establishes common ground"), and in turn by *indicators* (e.g., "proposes specific solutions").

In this work, we adopt the Sun *et al.* framework from [22] and use it to annotate a novel collaborative problem solving task: the Weights Task (Section 3.1).

A variety of tasks have been proposed to study small group collaboration. One such task is the winter survival task, which has been used to explore group work and leadership emergence in groups [17]. Sanches-Cortes *et al.* used this task to automatically predict emergent group leaders in [19]. Another method of studying small groups is by having them to participate in virtual learning games, as in [21]. These methods of studying small groups show promising results. To extend on this work, we employ the Weights Task, which prompts group work in an in-person environment where participants must interact with physical objects.

Collaboration is a complex phenomenon, and can be evaluated through several different lenses. One such perspective is task success, where groups are evaluated based on the results of their collaboration. This approach was used by Sun *et al.* [23] and by Avci and Aran [2]. This is a strong approach, which evaluates based on the goal of the group itself rather than an extrinsic metric, but may neglect other aspects of group work, such as group communication and contributions that do not directly advance the group toward the stated goal. To alleviate this, another approach is tracking the presence of CPS, such as in [21]. This allows for a finer-grained representation of the group's interactions and considers a variety of measures. However, the virtual communication captured by [21] differs from the interactions of in-person collaboration. For example, having a shared space allows for communication through body language. Here, we evaluate in-person collaborative work at the speaker-turn level.

A number of works have extracted features from CPS tasks and used them to train machine learning models [2, 21]. These works consider CPS in a controlled, virtual environment and showed that CPS facets could be meaningfully detected using machine learning. We extend these findings, showing that CPS facets can still be automatically detected in in-person group work, despite the increased complexity. For example, prior work automatically had speech segmented by participant since participants were each participating remotely, whereas we perform speech diarization to identify which group member was speaking when.

Several works have explored what features are important for detecting in-person CPS states. Castillon *et al.* proposed a toolbox for feature extraction geared toward complex, in-person collaborative problem solving environments [5], but did not use them to train machine learning models. Works such as [8, 13] do study the use of visual features in modeling in-person CPS in physical situations using machine learning, and we took inspiration from the above works in the design of our study. Our work presents a novel dataset for a physical, in-person shared task and we utilized many of the aforementioned features, as well as additional features, to establish our state-of-the-art baseline for this dataset.

3 Methods

In this section we describe the collaborative problem solving task we performed our study on, and our methods of data collection, preprocessing, and model training.

3.1 Data Collection

Weights Task We collected an audiovisual dataset of small groups collaborating on an in-person, shared, physically-grounded problem-solving task, known as the Weights Task. An example still from the dataset can be seen in Figure 1. In this task, participants form triads to solve a small puzzle together while being audiovisually recorded. Participants are given five colored cubes of different weights (there is a predetermined pattern relating the weights of the different blocks), a balance scale, and a worksheet to track answers. We identify the weight of one block, and ask them to use the balance scale to identify the weights of the remaining blocks by exploring blocks or combinations of blocks that balance with each other. When participants have identified the weights of all five blocks, we remove the balance scale and provide a new block of unknown weight to participants. Participants must then try to identify the weight of the mystery block. To successfully do this, participants must have inferred the pattern in the block weights. They have two attempts. We then ask participants for the weight of a hypothetical next block in the sequence, according to the pattern. They again have two attempts. Recording ends at the end of this second attempt. Participants then complete a post-activity survey, which includes a demographics section and participant perspective on their group. A prior version of this task was described in **[Removed for double blind]** — our version extends those methods in several ways to elicit more collaborative moves.

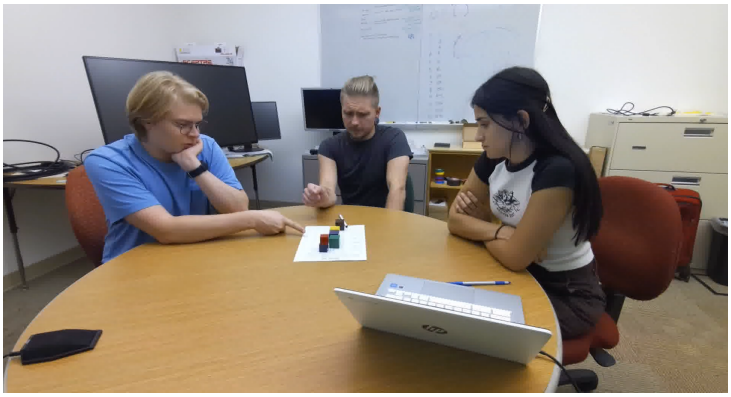

Fig. 1: The Weights Task

Participants Thirty participants were recruited for this study. All participants were over the age of 18 and spoke fluent English. All participants were students at a public research university in the western United States. Participants were 20% female and 80% male. Participants were between the ages of 19 and 35, with a mean age of 24.6 and a standard deviation of 4.6 years. Table 1 summarizes the demographic profile of participants. When asked to identify their ethnicity, 60% of participants identified as Caucasian, 10% identified as Hispanic or

Latino, and 30% identified as Asian. Participants indicated a range of native languages including English, Hindi, Assamese, Gujarati, Bengali, Telugu, Persian, Malayalam, Urdu, and Spanish.

Table 1: Demographics

Gender	Male	24
	Female	6
Native Language	English	18
	Non-English	12
Age	19-24	17
	25-35	13

Recording The full dataset consists of ten triads completing the Weights Task. The audiovisual recordings include audio and three angles of visual information including RGB and depth information. Recordings average 16 minutes. Audio recording used an MXL AC-404 Procon microphone as advised by findings from [**Removed for Double Blind**]. For video (RGBD) recording, we use three Azure Kinects due to their multichannel capabilities. The Azure Kinects are able to record with depth information, which provides richer information for visual features.

3.2 Audio Processing

Each audio recording was processed using Google's Voice Activity Detection (VAD) [16] to automatically segment audio files into utterances, with only one speaker per segment. This allows for utterance-level processing. Next, we transcribe the audio files using Google's automatic speech recognition (ASR). At this point, each group recording consisted of a collection of segmented audio files with transcripts for each segment. There were a total of 1,822 utterances, and the average utterance was 4.26 seconds long.

3.3 Annotations

Utterances from each group were annotated for collaborative problem solving (CPS). These were annotated according to the framework of CPS developed by Sun *et al.* [22]. Each automatically segmented utterance was labeled with the specific indicators of CPS, or as no indicator present. If the utterance had an indicator, the corresponding sub-facet and facet are inherited due to the nature of this framework. Coders were trained over utterances from one group, and then each utterance was annotated by two different coders. A single adjudicator finalized all of the utterance labels. Table 2 shows the average number of occurrences of each facet per group.

3.4 Verbal Features

Verbal features comprised features corresponding to the words per utterance spoken and transcribed by Google's ASR. Each group's utterance-level transcripts were preprocessed for formatting (including removing newlines and periods and surrounding the utterance with BERT's required `[CLS]` and `[SEP]` tokens), and

Table 2: Descriptive statistics of all 1822 utterances across all groups

	Average	SD	Min	Max
All utterances	182.20	80.51	90	380
# None	88.70	45.81	53	212
# Construction of shared knowledge	60.70	28.56	22	114
# Negotiation/Coordination	26.60	9.33	14	50
# Maintaining team function	6.20	4.35	0	16
Time (s)	4.26	2.84	0.84	23.64

then fed into the BERT Transformer model [9] to retrieve the sentence embedding for each utterance, providing a real-valued vector representation of the utterance in semantic space. BERT employs bidirectional pre-training to create language representations, where the sequence-level representation can be extracted from the `[CLS]` token prepended to the utterance. To expedite computation, we use the BERT-SMALL model first published in [24] and made available on the HuggingFace platform. Therefore the embedding size is 512 dimensions.

3.5 Prosodic Features

Prosodic features here refers to the non-linguistic features of speech. Each group's audio files were processed using openSMILE [11] to extract prosodic features of speech — e.g., features relating to frequency, amplitude, and balance. We used the extended feature set predefined by Eyben *et al.* [10]. This feature set is a basic standard set which aims to be minimalist while still effective. After processing, each utterance has an associated total of 88 prosodic features, such as loudness and spectral flux.

3.6 Visual Features

Visual features were extracted videos using RGBD information recorded by the Azure Kinects. We used a pipeline developed by Microsoft to extract skeletal joint information from all three participants. Each skeleton contains 32 joints, each with 3 position values and 4 orientation values, represented as numerical arrays. Each frame was expected to contain 3 bodies, and when fewer than 3 bodies were detected, the joint information of the missing bodies were set to 0s. When more than 3 bodies were detected, only the bodies most similar to the bodies detected in the prior frame were kept.

3.7 Model Training

We use three types of models for evaluation: a random forest, an AdaBoost classifier, and a neural network.

Figure 2 shows the architecture of the neural network. Visual features (e.g., points resulting from the joint tracking pipeline) were passed through a 3D convolutional neural network (CNN), then a 2D CNN, and a 1D CNN. Finally, a basis spline was used for dimensionality reduction and padding was used to ensure the resulting tensors held the same shape. The neural network classification head (blue box) consists of one hidden layer with ReLU activation followed by

the output later, which consists of 18[1] nodes with softmax activation for multinomial classification, or 1 node with sigmoid activation for binary classification.

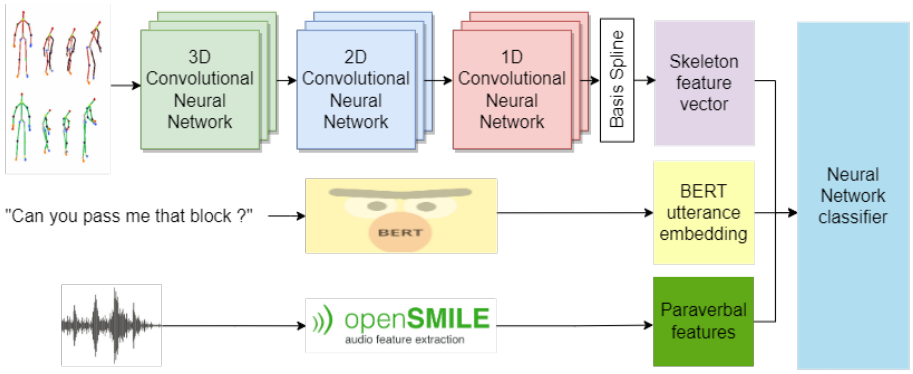

Fig. 2: Neural network classifier architecture

We perform hyperparameter search for random forest and AdaBoost classifiers using Hyperopt, a library which provides support for automatic distributed hyperparameter optimization [3]. In these cases, the visual features were reduced to joint position and rotation values from three frames evenly spread across each utterance. These values, along with the BERT embeddings and openSMILE features, were passed to Hyperopt using the `sklearn` AdaBoost and random forest classifiers [18]. Hyperparameter optimization for the neural network was performed using a grid search.

Evaluation Metrics We perform leave-one-group-out cross-validation using the `sklearn` library [18]. We evaluate our models using Area Under the Receiver Operating Characteristic Curve (AUROC) weighted by the number of instances of each class.

4 Results

Table 3 shows results of multinomial collaborative problem solving (CPS) facet classification, and respective standard deviations from the leave-one-group-out evaluation are given in Table 4. In general, the low standard deviations in Table 4 indicate CPS states were classified consistently with across groups. Our best performing multimodal model identified CPS with an average weighted AUROC of .812 (cross group SD of .03).

Table 5 shows the results of binary CPS facet classification with respective standard deviations given in Table 6. As discussed in our related works, binary

[1] 18 comes from the 17 CPS indicators from the Sun *et al.* framework, plus one class for no indicator. During evaluation, we use a dictionary to translate the predictions from indicator level to facet level. During training, the model still predicts over 18 classes, to optimize over a more fine grained level of prediction.

classification (presence or absence of a CPS facet) is important since some utterances may contain multiple collaborative components [23]. "None" in Table 3 is equivalent to the absence of all facets in Table 5. Area Under the Receiver Operating Characteristic Curve (AUROC) was computed using test results from every utterance.

Table 3: Weighted average AUROC for multinomial classification

	None			Construction of shared knowledge			Negotiation/ Coordination			Maintaining team function		
Modalities	RF	AB	NN	RF	AB	NN	RF	AB	NN	RF	AB	NN
Verbal	.867	.825	.828	.790	.747	.755	.725	.695	.479	.705	.687	.652
Prosodic	.827	.811	.519	.761	.734	.523	.671	.646	.491	.595	.672	.442
Visual	.562	.512	.499	.570	.515	.501	.514	.504	.500	.511	.631	.450
Verbal+Prosodic	.876	.839	.646	.801	.760	.603	.714	.694	.577	.682	.696	.523
Prosodic+Visual	.805	.794	.775	.738	.722	.718	.647	.642	.633	.577	.608	.546
Verbal+Visual	.859	.816	.878	.794	.738	.780	.718	.673	.729	.637	.633	.680
All Features	.858	.841	.876	.790	.757	.791	.722	.683	.711	.612	.658	.661

Table 4: Standard deviations of weighted average AUROC across groups for multinomial classification

	None			Construction of shared knowledge			Negotiation/ Coordination			Maintaining team function		
Modalities	RF	AB	NN	RF	AB	NN	RF	AB	NN	RF	AB	NN
Verbal	.042	.032	.032	.048	.036	.040	.063	.042	.059	.114	.095	.094
Prosodic	.067	.069	.013	.052	.055	.016	.034	.069	.012	.084	.140	.014
Visual	.055	.070	.007	.072	.065	.003	.058	.068	.008	.121	.085	.010
Verbal+Prosodic	.040	.051	.050	.044	.053	.047	.061	.054	.050	.151	.097	.144
Prosodic+Visual	.074	.078	.042	.063	.070	.030	.069	.063	.046	.111	.101	.111
Verbal+Visual	.038	.020	.034	.044	.106	.044	.061	.078	.041	.104	.217	.087
All Features	.049	.051	.039	.052	.057	.044	.073	.053	.057	.107	.095	.084

Table 5: Weighted Average AUROC for binary classification

	Construction of shared knowledge			Negotiation/ Coordination			Maintaining team function		
Modalities	RF	AB	NN	RF	AB	NN	RF	AB	NN
Verbal	.793	.786	.572	.712	.709	.555	.704	.705	.501
Prosodic	.756	.733	.529	.665	.661	.513	.622	.718	.533
Visual	.545	.515	.501	.505	.530	.500	.415	.656	.454
Verbal + Prosodic	.801	.779	.589	.742	.728	.525	.706	.730	.487
Prosodic + Visual	.739	.723	.695	.642	.659	.621	.525	.667	.499
Verbal + Visual	.787	.782	.790	.733	.719	.730	.626	.670	.676
All Features	.794	.770	.745	.721	.712	.724	.664	.677	.477

5 Discussion

In many cases, we achieve results comparable to or even exceeding those reported in [21], even though our shared environment and task are noisier and our data size is smaller (30 participants compared to 111).

Table 6: Standard Deviations of Weighted Average AUROC across all 10 groups for Binary Classification

	Construction of shared knowledge			Negotiation/ Coordination			Maintaining team function		
Modalities	RF	AB	NN	RF	AB	NN	RF	AB	NN
Verbal	.056	.045	.043	.045	.044	.039	.111	.121	.269
Prosodic	.048	.065	.135	.034	.058	.073	.081	.100	.199
Visual	.068	.040	.001	.063	.028	.003	.128	.080	.159
Verbal + Prosodic	.049	.046	.133	.057	.040	.048	.128	.080	.262
Prosodic + Visual	.050	.068	.093	.067	.061	.058	.118	.069	.177
Verbal + Visual	.048	.044	.042	.049	.047	.048	.071	.116	.283
All Features	.051	.045	.102	.042	.038	.047	.128	.108	.248

We often observe that performance with feature combinations does not significantly exceed that with verbal (linguistic) features alone. In these cases, the utterances or numerical representations thereof usually carry sufficient information to classify or detect CPS facets most of the time. Adding prosodic or visual features provides only a slight boost to performance.

The neural network, on the other hand, performs at just around chance using visual features alone, but performance improves when features are combined. This is particularly the case for the "Negotiation/Coordination" facet, suggesting then when multiple channels are available, the neural network model can effectively learn correlations between them. This raises the question of data availability, which is a known hurdle for neural networks. In this regard, "Maintaining team function" is a difficult facet for the neural network to handle, since it is sparsely occurring in our data overall (Table 2).

5.1 Example Analysis

A look into specific classifications can reveal when a model or set of features is failing to generate correct outputs. For example, an utterance with the transcription "I still think the brown one is heavier than this" can be correctly classified as construction of shared knowledge when all features are given; however, the neural network fails to classify the utterance when given only visual features. The neural network is given features extracted from joint location and rotation, but there is not annotation of the semantics of any given movement. In the accompanying video (still shown in Figure 3), there is a lot of movement from the group, but nothing distinctly meaningful such as a gesture. In this example, the verbal features are crucial in making a correct classification.

6 Conclusion

In this study, we have used several tools for automatic feature extraction and trained multiple machine learning classifier models to detect collaborative problem solving (CPS) in small groups. We achieve promising results on multimodal detection of CPS using a novel combination of features, in a challenging in-person setting, in a task that requires real-time interaction with physical objects. This

Fig. 3: Still from video segment containing utterance "I still think the brown one is heavier than this."

demonstrates the technical feasibility of tracking small group interactions between individuals and objects in the environment for CPS detection, and is the first step in developing an agent that can successfully track groups in real time. This may be applied in real-life settings, such as in classrooms, which would help educators monitor the state of small groups and may indicate which groups need extra assistance or facilitation to maintain successful group dynamics. Eventually, an agent itself may also be able to perform some interventions with groups, which could help reduce some of the lower-level facilitation tasks currently taken on by teachers, or provide it where it would otherwise have been unavailable. This ultimately can improve small group collaboration in the classroom.

6.1 Future Work and Limitations

Despite the promising results we exhibit, we should note that our data size is still relatively small (Stewert *et al.* had 111 participants [21]) and our facet-level annotations are coarse-grained. Further, while our participants exhibit a range of ethnic, national, and linguistic backgrounds, all participants are students at a mountain state US university and they still tend to satisfy most conditions of the WEIRD paradigm [14]. Future work includes scaling up to more data, include iterations of the Weights Task performed in different environments, and different tasks entirely, as well as performing an analysis of classification at the sub-facet and indicator levels, and including annotations that capture other channels, like the aforementioned non-verbal behavior or gesture. Finally, our dataset is collected outside of classrooms — classroom environments will inevitably be subject to additional noise, which future work will need to address.

We expected visual features to improve CPS detection in in-person settings; however, our models that only utilized visual features identified CPS with chance performance. This may be because our visual features focused only on the skeletal data from the Kinects — considering how impactful BERT features were for verbal classification, future work should focus on state-of-the-art techniques for visual feature extraction.

References

1. Andrews-Todd, J., Forsyth, C.M.: Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. Computers in Human Behavior **104**, 105759 (Mar 2020). https://doi.org/10.1016/j.chb.2018.10.025, https://www.sciencedirect.com/science/article/pii/S0747563218305156
2. Avci, U., Aran, O.: Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. IEEE Transactions on Multimedia **18**(4), 643–658 (Apr 2016). https://doi.org/10.1109/TMM.2016.2521348
3. Bergstra, J., Yamins, D., Cox, D.D.: Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In: Proc. of the 30th International Conference on Machine Learning (2013)
4. Bradford, M., Hansen, P., Beveridge, J.R., Krishnaswamy, N., Blanchard, N.: A deep dive into microphone hardware for recording collaborative group work. In: Proceedings of the International Conference on Educational Data Mining (2022)
5. Castillon, I., VanderHoeven, H., Bradford, M., Venkatesha, V., Krishnaswamy, N., Blanchard, N.: Multimodal Features for Group Dynamic-Aware Agents. In: Interdisciplinary Approaches to Getting AI Experts and Education Stakeholders Talking Workshop at AIEd. International AIEd Society (2022)
6. Cukurova, M., Giannakos, M., Martinez-Maldonado, R.: The promise and challenges of multimodal learning analytics. British Journal of Educational Technology **51**(5), 1441–1449 (2020), publisher: WILEY
7. Cukurova, M., Luckin, R., Millán, E., Mavrikis, M.: The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. Computers Education **116**, 93–109 (2018), publisher: Elsevier
8. Cukurova, M., Zhou, Q., Spikol, D., Landolfi, L.: Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough? In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 270–275 (2020)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (May 2019). https://doi.org/10.48550/arXiv.1810.04805, http://arxiv.org/abs/1810.04805
10. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing **7**(2), 190–202 (Apr 2016). https://doi.org/10.1109/TAFFC.2015.2457417
11. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1459–1462. MM '10, Association for Computing Machinery, New York, NY, USA (Oct 2010). https://doi.org/10.1145/1873951.1874246, https://doi.org/10.1145/1873951.1874246
12. Graesser, A.C., Fiore, S.M., Greiff, S., Andrews-Todd, J., Foltz, P.W., Hesse, F.W.: Advancing the Science of Collaborative Problem Solving. Psychological Science in the Public Interest **19**(2), 59–92 (Nov 2018). https://doi.org/10.1177/1529100618808244, https://doi.org/10.1177/1529100618808244

13. Grover, S., Bienkowski, M., Tamrakar, A., Siddiquie, B., Salter, D., Divakaran, A.: Multimodal analytics to study collaborative problem solving in pair programming. In: Proceedings of the Sixth International Conference on Learning Analytics Knowledge. pp. 516–517 (2016)
14. Henrich, J., Heine, S.J., Norenzayan, A.: The weirdest people in the world? Behavioral and brain sciences **33**(2-3), 61–83 (2010)
15. Howe, C., Tolmie, A., Greer, K., Mackenzie, M.: Peer collaboration and conceptual growth in physics: Task influences on children's understanding of heating and cooling. Cognition and instruction **13**(4), 483–503 (1995)
16. Karrer, R.: Google WebRTC Voice Activity Detection (VAD) module (2022), https://www.mathworks.com/matlabcentral/fileexchange/78895-google-webrtc-voice-activity-detection-vad-module
17. Kickul, J., Neuman, G.: Emergent leadership behaviors: The function of personality and cognitive ability in determining teamwork performance and KSAS. Journal of Business and Psychology **15**, 27–51 (2000). https://doi.org/10.1023/A:1007714801558, place: Germany Publisher: Springer
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
19. Sanchez-Cortes, D., Aran, O., Mast, M., Gatica-Perez, D.: A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. IEEE Transactions on Multimedia **14**, 816–832 (Jun 2012). https://doi.org/10.1109/TMM.2011.2181941
20. Slavin, R.: Research on Cooperative Learning: Consensus and Controversy. Educational Leadership **47** (Jan 1990)
21. Stewart, A.E.B., Keirn, Z., D'Mello, S.K.: Multimodal modeling of collaborative problem-solving facets in triads. User Modeling and User-Adapted Interaction **31**(4), 713–751 (Sep 2021). https://doi.org/10.1007/s11257-021-09290-y, https://doi.org/10.1007/s11257-021-09290-y
22. Sun, C., Shute, V.J., Stewart, A., Yonehiro, J., Duran, N., D'Mello, S.: Towards a generalized competency model of collaborative problem solving. Computers & Education **143**, 103672 (2020), https://www.sciencedirect.com/science/article/pii/S0360131519302258
23. Sun, C., Shute, V.J., Stewart, A.E.B., Beck-White, Q., Reinhardt, C.R., Zhou, G., Duran, N., D'Mello, S.K.: The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. Computers in Human Behavior **128**, 107120 (Mar 2022). https://doi.org/10.1016/j.chb.2021.107120, https://www.sciencedirect.com/science/article/pii/S074756322100443X
24. Turc, I., Chang, M.W., Lee, K., Toutanova, K.: Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962 (2019)