CPSX: Using AI-Machine Learning for Mapping Human-Human Interaction and Measurement of CPS Teamwork Skills

Pravin Chopade, David Edwards, Saad M. Khan, Alejandro Andrade and Scott Pu
Artificial Intelligence and Machine Learning
ACTNext, ACT Inc., Iowa City, IA, USA

Abstract— **The objective of this work is to present a machine learning (ML) -based framework to identify evidence about collaborative problem solving (CPS) cognitive (teamwork) and social-emotional learning (SEL) skills from the dyadic (human-human-HH) interactions. This work extends our previous work (Chopade et al. IEEE HST 2018, LAK2019) [1], [2]. Explicitly, we are interested in how teamwork skills and team dynamics are demonstrated as verbal and nonverbal behaviors, and how these behaviors can be captured and analyzed via passive data collection. For this work we use a two-player cooperative CPS game, Crisis in Space (CIS) from LRNG (Previously GlassLab Inc). During the summer of 2018, we implemented this CIS game for interns as a group study. A total of 34 participants played the game and provided study and survey data. During the study, we collected participants' game play data, such as audio, video and eye tracking data streams. This research involves analyzing CIS multimodal game data, and developing skill models, and machine learning techniques for CPS skills measurement. In this paper, we present our ML framework for the analysis of audio data along with preliminary results from a pilot study. The analysis of audio data uses natural language processing (NLP) techniques, such as bag-of-words and sentence embedding. Our preliminary results show that various NLP features can be used to describe successful and unsuccessful CPS performances. The ML based framework supports the development of evidence centered design for teamwork skills-mapping and aims to help teams operate effectively in a complex situation. Potential applications of this work include support for the Department of Homeland Security (DHS), and the US Army for the development of learner and team centric training, cohort, and team behavioral skill-mapping.**

Keywords-Collaborative Problem Solving (CPS), Human-Human Interactions, Teamwork Skills, Congitive Skills, Social-Emotional Learning (SEL) Skills, Behaviors, Performance Measurement, Artificial Intelligence, Machine Learning, Natural Language Processing (NLP).

I. INTRODUCTION

Collaborative problem solving (CPS) constructs, defined as a combination of 21st century skills and behaviors, which are progressively more important in today's multifaceted and interconnected world. CPS is widely considered to be a core competency in the workforce [3] and is a key component of several standards for primary and secondary education [4], [5], and [6]. Additionally, collaboration has been used in high-stakes assessments of traditional educational outcomes in several state-wide assessments [7], [8], [9], and [10], and is a component of the recently developed Common Core assessments [11] and other international assessment programs [12], [13] and [14]. This clearly indicates a growing awareness of the impact of collaboration on academic outcomes, and therefore, there is a growing interest in the teaching and assessment of users' (students') CPS skills [15]. However, little work has focused on the assessment design, scoring, and reporting of CPS skills, and specifically in the context of formative, multimodal measurements of these skills. To date, educational assessments, both summative and formative, have focused on the more concrete and accessible aspects of these skillsets, i.e., individual cognitive skills [14], [15], [16]. Due to the omission or diminished representation of these hard to assess but critical skillsets, current assessments offer limited insights about users' (students') proficiency of CPS skills.

In addition, there is a lack of clarity on the definition of the CPS construct and there is still not a valid and reliable system for the assessment of CPS skills [17]. In principle, CPS consists of multiple skills, such as resource and knowledge sharing, which must be coordinated on a moment-by-moment basis for successful outcomes [18]. These skills, which are the primary evidence of collaboration, can be difficult to assess in any context, much less during in-vivo collaboration. Consequently, efforts to assess CPS skills often involve the use of simulations, games, and other team-based classroom activities. Game-based (Artificial Intelligence-AI and ML) skill measurement as well as simulations may provide an environment that more accurately represents how a team would perform CPS. Computer-based educational environments are transforming how tests are delivered by providing more realistic and authentic assessment scenarios. A game-based learning and assessment system allows users (students) to demonstrate their mastery at solving complex challenges, making assessment engaging and rewarding. Thus, simulations and games offer an exciting new paradigm for the assessment of knowledge and skills that are difficult to assess with more familiar item types such as multiple-choice questions and self-reports.

Recent advances in the fields of AI-machine learning, educational data science, computer vision, and affective computing have allowed researchers to analyze and score a wider range of response data. These technological advances expand the types of activities that can be used to characterize the knowledge, skills, and attributes of test takers [19]. This paper

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

focuses on implementing various machine learning models for identifying teamwork skills during the play of a collaborative game called "Crisis in Space" (CIS) from LRNG (Previously Glass Lab Inc) for a two-person (Human-Human) jigsaw assessment task. We present results from a study that utilized a game-based prototype to measure CPS teamwork skills evidence.

In order to create a chain of CPS skills evidence from real Human-Human (HH) interaction, this paper is organized as follows. The next section, Section II, describes some relevant work in the area of CPS and teamwork-based assessment. In Section III, we discuss CPS teamwork skills with micro-analysis of CPS deep structure and task sequence and its alignment with the multimodal analytics framework. In Section IV, we present the game Crisis in Space (CIS), which allows CPS real HH interaction, along with our experimental design and multimodal data collection. In Section V, we present our ML framework and in Section VI we outline our preliminary results from an early-stage analysis of mapping CPS skills evidence using data from a summer 2018 pilot study. Finally, sections VII and VIII conclude our paper with important key findings, potential applications and guidelines for the next phases of this work.

II. Related Work

In the past few years, several theoretical frameworks have attempted to define CPS skills. For instance, [14], [15], [20], [21], [22], [23] define CPS as a process that depends on the coordination of cognitive and social skills. Hesse et al. [24] provided a CPS theoretical framework that is supported by findings from diverse fields of research, and other researchers have also published similar research related to CPS based assessments [25], [26], [27], [28], [29]. For example, Hao et al. [25] presented a method for analyzing process data from game-based tasks. Their results show a strong correlation between collaborative skills and game scores obtained from a pump repair task scenario [25]. In related work, Graesser's et al. [29] tackled the problem of integrating problem-solving measurement and collaboration measurement with computer agents. Graesser et al. [21] created AutoTutor, an intelligent tutoring system (ITS) that demonstrates CPS assessment via natural language student-agent interactions. While interacting with computer agents, AutoTutor assesses students' cognitive and emotional states. A few other related works have been produced around assessing CPS performance. For instance, Rosen and Foltz [30] illustrated a method for analyzing students' CPS performance in human-to-agent (H-A), and human-to-human (H-H) standardized assessment situations. Their method provides thorough, large-scale H-A and H-H assessments targeting collaboration and social competencies. Care et al. [31] outlined a method for the identification of CPS sub-skills and a method for skill assessment. Based on CPS sub-skills, their method aims to identify levels of ability among students and provide formative feedback to teachers for student improvement.

Additionally, Von Davier et al. [32], [33], [34], [35] have led recent efforts towards developing robust and innovative methods for CPS assessments. In her recent publications [32], [33], [34], [35], von Davier highlights the need to incorporate data mining tools and innovative methods into CPS assessments. Von Davier and Halpin [35] sketched out a number of issues facing models that are used to derive interaction quality from existing data. These issues stem primarily from the complexity and reflexivity of the tasks. These authors indicated that statistical models will have to be tailored to the level of collaboration involved in each task, as well as the collaboration structure.

Researchers at ACTNext, ACT, Inc. are extensively engaged in CPS research and have developed a framework for CPS constructs and teamwork evidence analysis [1], [2], [23], [36], [37], [38], [39]. In our most recent work [1], [2], we presented a ML framework for efficient measurement and prediction of human performance in a teamwork and collaborative learning related task. This work introduced a multimodal system architecture to encourage understanding of team dynamics, interactions, and behaviors in a collaborative learning environment (CLE). In an effort to further explore CPS sub-skills and meaningful information from log-data, our work models CPS sub-skills and related two-person human-human interactions using the 'CIS' game.

The next section describes the CPS cognitive, teamwork, SEL skills, and interaction conditions those are central for the broader research framework.

III. CPS Skills and Multimodal Analytics Framework

The purpose of our work is to measure CPS skills based on the CPS theoretical construct. According to recent research on CPS [14], [21], [22], [23] the collaborative problem-solving construct is segmented into two components: a) Cognitive Team Task skills, and b) Socio-Emotional Team skills. These two components are key to any effective collaboration within a group and are further broken down into CPS attributes, as shown in Fig. 1. Cognitive skills such as: individual knowledge building; team knowledge building; decision making and execution; monitoring and social-emotional learning (SEL) skills: leadership; cooperation; relationships (rapport); resilience [23]. Following the CIS game play session, participants (students) will be asked to respond to a brief survey about their gameplay experience, a background questionnaire, a CPS Survey, and an ACT Tessera Assessment [40].

As discussed earlier, CPS constructs are not readily amenable to traditional item-response-like analysis. In fact, the measurement of these constructs entails the recognition and extraction of patterns of behaviors and activities from time-series data that constitutes the interaction between individuals and between individuals and the simulation task. This data is multimodal by nature i.e., it encompasses multiple modalities like mouse clicks, keystrokes, as well as audio, video and eye-tracking, which enable real-time capture and recognition of the user's verbal, non-verbal, and paralinguistic behaviors. These different modalities can be integrated into a measurement model to describe the user's actions and behaviors indicative of cognitive and SEL processes, enabling the scoring of complex CPS competencies. We attempt to build ML models from these multimodal data to measure various CPS skills. First, we provide a description of our CIS game and the design of multimodal data collection, discussed in section IV. Second, we discuss our ML framework and feature the engineering process. In section V, we

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

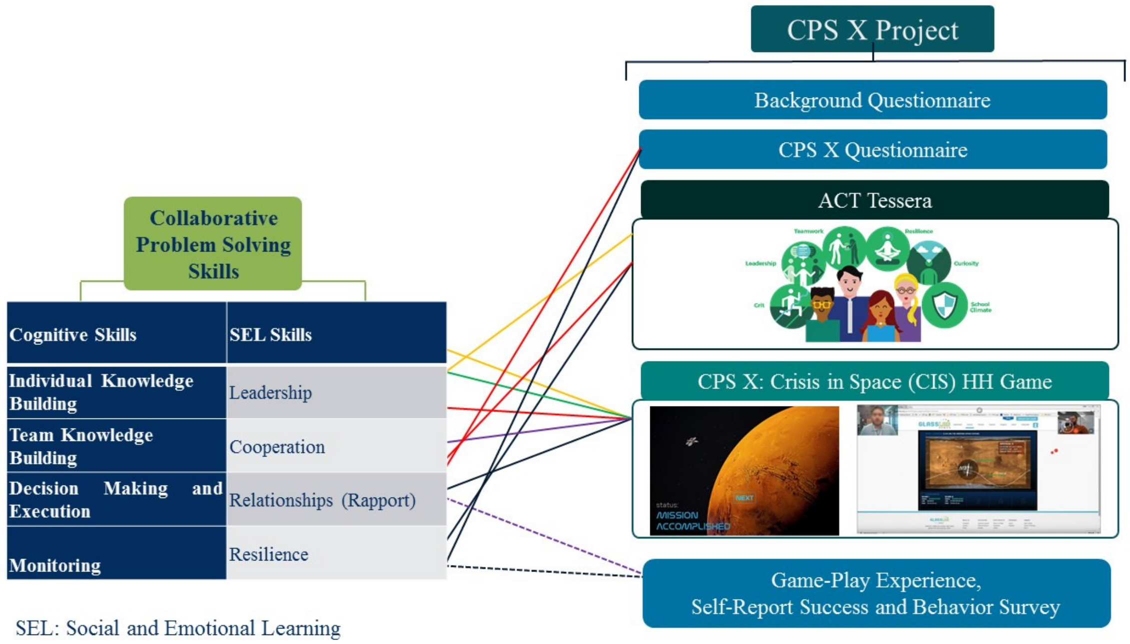

Fig. 1. CPS skills and multi-modal analytics framework. Copyrights ACTNext, ACT, Inc., USA.

focus on presenting an ML model developed for the analysis of audio data to provide evidence of the participants' communication attribute (Cognitive Team Knowledge Building CPS subskill). Then, preliminary results are presented for the mission success rate based on an automated analysis of the dyads' conversations.

In the next section we provide a description of our CPS CIS game design for collecting multimodal data and the mapping of CPS teamwork skill to behavioral evidence.

IV. CIS Game Design-Task and Multimodal Data Collection

During August-September of 2018 a first iteration of this research was implemented, and data from a group of summer interns and a few team members (N=34, Female: 15, Male: 19) was collected. Participants played the CPS game called Crisis in Space, from LRNG (Previously GlassLab Inc.), which is a two-player (Human-Human) jigsaw game where a series of puzzles must be solved by sharing information [2]. The game is played by two people at a time, who alternate the roles of Astronauts (A) and Engineers (E) on a mission to Mars. Each dyad goes through a series of five (5) missions, which includes sixteen (16) sub-tasks in total (See Figs 2 and 3). Users must cooperate to deal with potential dangers and threats in order to ensure that the mission is a success. The road is difficult as users must decode encrypted messages, avoid asteroid impacts, and put exploratory satellites in orbit around Mars. While not all participants succeed in their missions, all participants report a high level of engagement with the game [2].

Data collection for the summer pilot study proceeded as follows. Each participant sat in a room with a laptop, a camera, a microphone, a speakerphone, and some scratch paper. Participants connected to each other via a Skype call, as shown in Fig. 2. This setup enabled easy syncing between data sources, as well as simple isolation of voices (phone quality). In total, this game provided five separate forms of data, including: 1) Game Logs, 2) Chat Logs 3) Eye Tracking with screen capture, 4) Portrait Videos, and 5) Audio files. For the purpose of this study, we analyzed the audio stream, which was also submitted to Amazon Transcribe web-service to generate text transcriptions. This phase involved the development of a machine learning model, which is presented in the next section.

Fig. 2. CIS Game and experimental data collection setup.

V. Machine Learning Model of Communication Skills

This section presents our ML framework, shown in Fig. 3, for participants' audio analysis and how we plan to map them to CPS skills. The ML model of audio streams consists of the integration and analysis of two data modules: a) acoustic frames, and b) text features. In this paper, however, we focus on the analysis of textual data using various natural language processing (NLP) technique.

In what follows, we outline our preliminary work on

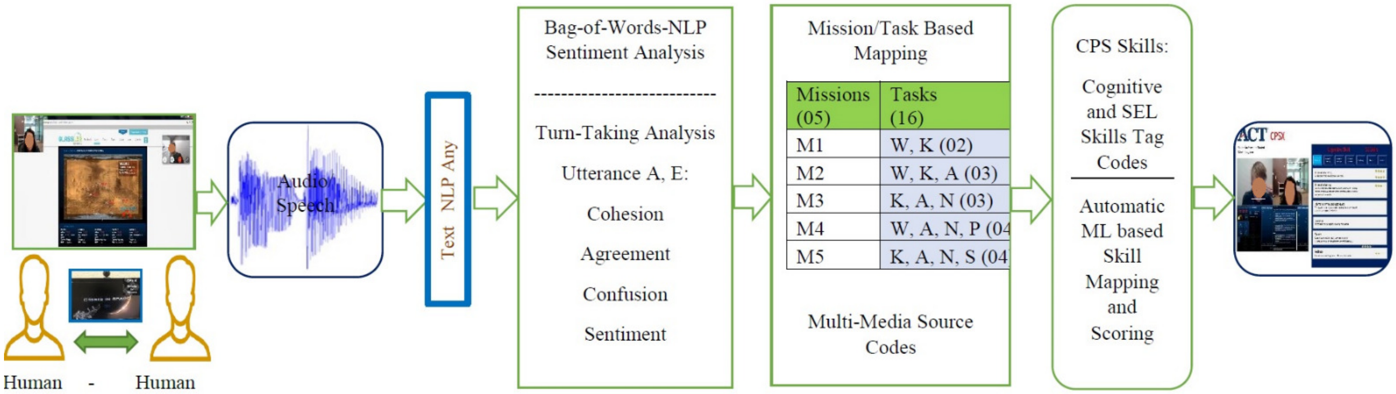

Fig. 3. ML-based audio and conversation analysis framework. Copyrights ACTNext, ACT, Inc., USA.

developing NLP features to automatically extract meaningful evidence of a dyad's communication skills. The goal was to develop quantitative features to represent various aspects of the conversation that might be mapped onto CPS skills. In particular, we used a two-pronged approach: a) a bag-of-words technique to represent general aspects of the conversation; and b) an analysis of the conversational turns using sentence embedding to represent finer-grained, moment-by-moment information.

A. *Bag-of-Words Analysis*

First, using Amazon Web Services (AWS) Transcribe service we converted audio conversation files to a JSON format and then to a text file format. A MATLAB [41] script for NLP Sentiment analysis was used for text feature extraction. This analysis consisted of a bag-of-words approach where raw text data was converted to lowercase, punctuation erased, stop words removed, and lemmatized, to finally create a vector of tokens per each audio file. Then, we used latent dirichlet allocation (LDA) and latent semantic analysis (LSA) [41], which discover relationships between documents and the words that they contain, to analyze the preprocessed text. For the LDA analysis, we extracted features such as topic concentration and perplexity. For the LSA analysis, component weights and document scores were calculated. In order to detect positive and negative words, we created a list of 10 common positive and negative words to tally the number of positive and negative words in each dyadic CIS game HH interactions.

B. *Turn-by-Turn Analysis*

A turn-by-turn transcript was developed from the Amazon Transcribe web-service raw data. A turn-by-turn transcript contains all the words uttered by a speaker during each conversational turn. A conversational turn is the period a speaker uses to communicate an idea and usually finishes when a second speaker starts producing an utterance. Text was spellchecked and punctuation and spacing was normalized using a Natural Language Toolkit (NLTK) [42] preprocessing module. For the turn-taking analysis, four features were developed: sentiment, cohesion, agreement, and confusion (See Fig. 3). These values were calculated using pre-trained sentence embedding from the Universal Sentence Encoder [43] and Flair's pre-trained sentiment analysis model [44].

Sentiment, cohesion, agreement, and confusion features were extracted from each interactional turn (i.e., a speaker's utterance) using the Universal Sentence Encoder. The Universal Sentence Encoder retrieves a 512-dimensional vector for each utterance that encodes the meaning of the text in a geometrical space. Then, these embedding vectors were compared to anchor vectors. Cohesion values were extracted from cosine similarities between two adjacent utterances. Agreement values were extracted from cosine similarities between a given utterance and a set of agreement anchors (e.g., "you're right", "sounds good"). Confusion values were extracted from cosine similarities between a given utterance and a set of confusion anchors (e.g., "I don't understand what we are supposed to do here."). Finally, a fourth feature was computed using Flair's pre-trained sentiment model to produce sentiment values for each utterance. Sentiment values range between 1 (i.e., positive) and -1 (i.e., negative).

To provide some validity evidence, these four features were used to predict the likelihood of a dyad's success at a given mission. In total, there are 115 missions for the 23 dyads. In the next section we discuss some preliminary results and further steps in our CPS teamwork skill analysis.

VI. Preliminary Results

In this section, we present and discuss some of the preliminary results based on audio analysis from the pilot study participants.

A. *Results from the Bag-of-Words Analysis*

Table I shows the results obtained for the bag-of-words approach. For all groups, we obtained a histogram plot for unique word distribution to explore the communication differences between successful and unsuccessful groups. Table I also shows some differences in LDA and LSA features, as well as NLP Sentiment analysis values. Fig. 4 shows an application of audio text NLP analysis for Group 13 (a successful dyad) and Group 21 (an unsuccessful dyad) for the roles Astronauts (A) and Engineers (E). These preliminary results indicate patterns of spoken keywords and information sharing required for completing CIS game tasks. We will keep developing this approach to map CPS teamwork skills evidence from dyadic CIS game play interactions.

Table I. Audio Text NLP Sentiment Analysis.

Group No	Role	Total Mission	RawBag	CleanBag	LDA Model		LSA Model			Num of	Num of
		Success out	NumWords	NumWords	Training	Topic	Component	Doc	Feature Strength	P+ words	N- words
		of 05	Vocabulary	Vocabulary	perplexity	Concentration	Weights	Score	Exponent		
SG13	A	4	419	67	124.4	0.25	89910	299.85	2	2	1
	E	4	431	47	123.2	0.25	86029	293.307	2	2	0
SG17	A	4	425	84	110.2	0.25	154827	393.4806	2	3	1
	E	4	377	59	107.4	0.25	121400	348.4293	2	3	1
SG18	A	3	525	92	139.5	0.25	164980	406.181	2	1	1
	E	3	456	65	128.3	0.25	106140	325.7959	2	2	0
SG19	A	2	475	68	141.9	0.25	92204	303.6511	2	1	1
	E	2	397	52	105.8	0.25	143900	379.3376	2	1	0
SG16	A	1	444	75	121.2	0.25	147150	383.5987	2	1	1
	E	1	477	56	119.2	0.25	141341	375.9535	2	1	1
SG21	A	0	488	76	142.3	0.25	148430	385.2661	2	3	1
	E	0	374	46	116.8	0.25	68725	262.1545	2	0	0
	A: Astronaut; E: Engineer										

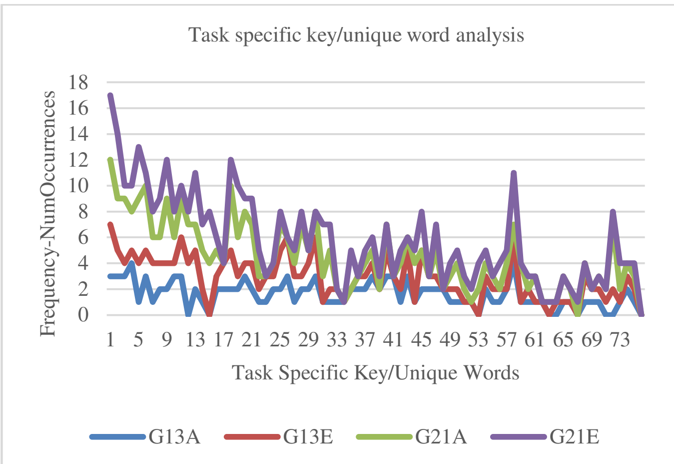

Fig. 4. Pattern of task specific key/unique words distribution for Group 13 (a successful dyad) and 21 (an unsuccessful dyad).

B. *Turn-Taking Analysis Results and Mission Success Rate*

An example of NLP feature distribution from a successful versus an unsuccessful mission can be seen in Fig. 5. A logistic regression was fit to the mission success values (1 = successful, 0 = unsuccessful) using the mission feature percentiles as predictors. Results show that these features explain about 17% of the success rate (R-squared = 0.1702) and, as anticipated, successful dyads have higher Sentiment values, higher Cohesion values, lower Confusion values and, interestingly, lower Agreement values. We are conducting additional explanatory analyses on these features to keep testing their relevance for the CPS skill mapping.

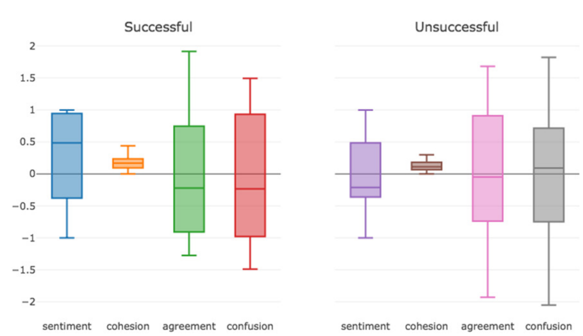

Fig. 5. Distribution of feature values for Group 13 (successful) and Group 09 (unsuccessful) during Mission 1.

VII. Discussion

In this paper, an approach to identify and subsequently score CPS teamwork skills has been presented. Poorly performing groups could be identified from information available early on in the assigned task. These groups could either be reconstituted to mitigate individual conflicts, or remediation could be provided in the specific collaborative skills lacking in the group. Such interventions would not only increase teams (students') collaborative problem-solving skills but could potentially increase educational outcomes in the content area of the collaborative task at hand (scientific inquiry and team communication, in the tasks in this study). Additionally, the development of such a multimodal framework (CPSX) would provide the ability to examine the impact of CPS on educational outcomes such as the scientific inquiry skills and argumentation skills measured in the computerized educational environments used in this study. Collaborative problem-solving skills have also proven to be malleable.

Responses to the questions posed in this paper will aid us better understanding how cognitive and SEL skills can be measured via non-invasive methods. The results from this CPSX project are designed to provide individuals and organizations for whom the assessment of cognitive and SEL skills is important with valid methods by which to measure these skills in relevant contexts.

VIII. Applications And Future Work

The ML based framework supports the development of evidence centered design for teamwork skills mapping and helps teams to operate in a complex situation. Potential applications of this work include support for Department of Homeland Security (DHS), US Army for the development of learner and team centric training, cohort, team cognitive, behavioral, SEL skills mapping. CIS game like modules can be developed for different real-time situational teamwork skill mapping and analysis.

In our future work, after low-level feature extraction, we plan to use Bayesian network models (BNM) and convolutional neural networks (CNNs) to perform classification and regression analyses on the audio features extracted. This work will require Subject Mater Experts (SMEs) to manually code various aspects of the collected data with relevant CPS subskill tags.

Acknowledgment

The authors would like to thank Andrew Cantine-Communications and Publications Manager, ACTNext for editing this work. We are also thankful to ACT, Inc. for their ongoing support as this paper took shape. The authors are thankful and acknowledge the support from our consultant Dr. Bryan Maddox, Assessment MicroAnalytics, University of East Anglia, Norwich, UK and LRNG, (Previously Glass Lab) for this work.

References

[1] P. Chopade, S. M. Khan, D. Edwards, and A. von Davier, "Machine learning for efficient assessment and prediction of human performance in collaborative learning environments, *IEEE International Symposium Technology f*or Homeland Security (IEEE-HST), October 23 - 24, 2018, Woburn, MA USA, Publisher IEEE Xplore, pp. 1-6. doi:10.1109/THS.2018.8574203

[2] P. Chopade, D. Edwards, and S. M. Khan, "Designing a digital jigsaw game based measurement of collaborative problem-solving skills, at the *9th International Conference on Learning Analytics and Knowledge (LAK19),* ASU Tempe, Arizona, USA, March 4-8 2019, pp. 26-31, 2019. URL: http://bit.ly/lak19-companion-proceedings

[3] H. F. O'Neil, S. Chuang, and G. K. W. K. Chung, "Issues in the computer-based assessment of collaborative problem solving" in *Assessment in Education: Principles, Policy & Practice*, 10, pp. 361-374, 2003.

[4] Common Core State Standards Initiative.: Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects. Washington, DC: National Governors Association and the Council of Chief State School Officers, 2010.

[5] M. Binkley, O. Erstad, J. Herman, S. Raizen, M. Ripley, M. Miller-Ricci, and M. Rumble, "Defining twenty-first century skills", in: P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills*, pp. 17-66, 2012. New York, NY: Springer.

[6] O. Astrachan, T. Barnes, D. Garcia, J. Paul, B. Simon, and L. Snyder, "CS principles: Piloting a new course at national scale", in: T. J. Cortina, E. L. Walker, L. S. King, & D. R. Musicant (Eds.). *In: Proceedings of the 42nd ACM technical symposium on computer science education*, pp. 397-398, 2011. New York, NY: ACM.

[7] Maryland State Department of Education.: Maryland School Performance Assessment Program: Public release tasks. Baltimore, MD: Author, 1994.

[8] Connecticut State Board of Education.: Connecticut Academic Performance Test (CAPT) interdisciplinary assessment. Hartford, CT: Author 1996.

[9] M. Saner, K. Stecher, and R. Bell, "The effects of working in pairs in science performance assessments", *Educational Assessment, 2,* pp. 325-338, 1994.

[10] M. Pomplun, "Cooperative groups: Alternative assessment for students with disabilities? *The Journal of Special Education*, 30, pp. 1-17, 1996.

[11] Smarter Balanced Assessment Consortium.: Thermometer crickets: Grade 11 performance task, 2012.

[12] National Research Council: Learning science through computer games and simulations. Washington, DC: The National Academies Press 2011.

[13] National Research Council: Assessing 21st century skills. Washington, DC: National Academies Press, 2011.

[14] OECD. The PISA 2015 *collaborative problem solving framework*. OECD Publications, 2013.

[15] OECD. *PISA 2015 Results in Focus*. Paris: France: OECD. Retrieved from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf, pp.1-16, 2016.

[16] S. Fiore, A. Graesser, S. Greiff, P. Griffin, B. Gong, P. Kyllonen, . . . A. von Davier, "*Collaborative problem solving: Considerations for the National Assessment of Educational Progress",* In: IES, National Center for Education Statistics (NCES), April 2017.

[17] Y. Rosen and M. Tager, "Computer-based assessment of collaborative problem-solving skills: Human-to-agent versus human-to-human approach", Research & Innovation Network, Pearson Education, 2013.

[18] N. Miyake and P. Kirschner, "The social and interactive dimensions of collaborative learning", in *The Cambridge handbook of the learning sciences*, (Cambridge Handbooks in Psychology). Cambridge: Cambridge University Press, pp. 418-438, 2014.

[19] R. Baker, S. D'Mello, M. Rodrigo, and A. Graesser, "Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments", in: *International Journal of Human-Computer Studies*, 68 (4), pp. 223-241, 2010.

[20] W. Camara, et al., "Beyond academics: A Holistic Framework for enhancing education and workplace success", Research Report Series, (4) 2015.

[21] A. Graesser, B.-C. Kuo, and C.-H. Liao, "Complex problem solving in assessments of collaborative problem solving", In: *J. Intell*., 5(2), 10, 2017. doi:10.3390/jintelligence5020010

[22] A. Graesser, S. Fiore, S. Greiff, J. Andrews-Todd, P. Foltz, and F. Hesse, "Advancing the science of collaborative problem solving. psychological science in the public interest, 19(2), pp. 59-92, 2018.

[23] A. Andrade, B. Maddox, D. Edwards, P. Chopade, and S. M. Khan, "Quantitative multimodal interaction analysis for the assessment of problem-solving skills in a collaborative online game", *ICQE19, International Conference on Quantitative Ethnography 2019*, October 20-22, 2019 Madison, Wisconsin.

[24] F. Hesse, E. Care, J. Buder, K. Sassenberg, and P. Griffin, "A framework for teachable collaborative problem solving skills", in: Griffin P., Care E. (eds) Assessment and Teaching of 21st Century Skills. *Educational Assessment in an Information Age*. Springer, Dordrecht, pp. 37-56, 2015.

[25] J. Hao, Z. Shu, and A. von Davier, "Analyzing process data from game/scenario-based tasks: An edit distance approach", In: *Journal of Educational Data Mining*, 7(1), pp. 33-50, 2015.

[26] J. Hao, L. Smith, R. Mislevy, A. von Davier, and M. Bauer, "Taming log files from game and simulation based assessment: Data model and data analysis tool", in: ETS Research Report Series, ETS RR-16–10. Princeton, NJ: Educational Testing Service, 2016.

[27] P. Kyllonen, "Measurement of 21st century skills within the common core state standards", Educational Testing Center, 2012.

[28] J. Voogt and N. Roblin, "A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies", in: *Journal of Curriculum Studies, 44(3),* pp. 299–321, 2012.

[29] A. Graesser, L. Haiying and F. Carol, "Learning by communicating in natural language with conversational agents", in: *Current Directions in Psychological Science,* 23.5, pp. 374-380, 2014.

[30] Y. Rosen and P. Foltz, "Assessing collaborative problem solving through automated technologies", Research and Practice, in: Technology Enhanced Learning, 9(3), pp. 389-410, 2014.

[31] E. Care, C. Scoular, and P. Griffin, "Assessment of collaborative problem solving in education environments", in: *Applied Measurement in Education*, vol. 29, no. 4, pp. 250-264, 2016.

[32] A. von Davier, M. Zhu, and P. Kyllonen, P. (Eds.), *Innovative assessment of collaboration.* Springer International Publishing, 2017.

[33] A. von Davier, "Computational Psychometrics in support of collaborative educational assessment", in: *Journal of Educational Measurement, 54*, pp. 3-11, 2017.

[34] A. von Davier (Ed.): Measurement issues in collaborative learning and assessment. [Special Issue], in: *Journal of Educational Measurement*. 54(1), 2017.

[35] A. von Davier and P. Halpin, "Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations", in: ETS Research Report Series, ETS RR-13-41, 2013.

[36] P. Chopade, K. Stoeffler, S. M. Khan, Y. Rosen, S. Swartz, and A. von Davier, "Human-Agent assessment: Interaction and sub-skills scoring for collaborative problem solving", In: Penstein Rosé C. et al. (eds) *Artificial Intelligence in Education. AIED 2018. Lecture Notes in Computer Science*, vol 10948, pp 52-57, 2018. Springer, Cham, DOI https://doi.org/10.1007/978-3-319-93846-2_10

[37] P. Chopade, S. M. Khan, K. Stoeffler, D. Edward, Y. Rosen, and A. von Davier, "Framework for effective teamwork assessment in collaborative learning and problem solving", in *19th International Conference on Artificial Intelligence in Education (AIED),* LAS, The Festival of learning, ARL workshop "Assessment and Intervention during Team Tutoring" AIED-2018, vol-2153, pp-48-59, 2018. http://ceur-ws.org/Vol-2153/paper6.pdf

[38] P. Chopade, M. Yudelson, B. Deonovic, and A. von Davier, "Modeling dynamic team interactions for intelligent tutoring", in Joan Johnston, Robert Sottilare, Anne M. Sinatra, C. Shawn Burke (ed.) *Building Intelligent Tutoring Systems for Teams (Research on Managing Groups and Teams, vol. 19)* Emerald Publishing Limited, pp.131 - 151, ISBN: 978-1-78754-474-1, September 2018.

[39] S. Khan, "Multimodal behavioral analytics in intelligent learning and assessment systems," in I*nnovative Assessment of Collaboration. Methodology of Educational Measurement and Assessment*, A. von Davier, Z. M. and K. P., Eds., Springer, Cham, pp. 173-184, 2017.

[40] ACT Tessera® Technical Bulletin. Iowa City, IA: ACT, 2018.

[41] Mathworks Inc. MATLAB- Statistics and Machine Learning, Signal and Audio Processing Toolbox, Deep Learning Toolbox, 2018.

[42] S. Bird, E. Klein, and E. Loper, Natural language processing with Python: Analyzing text with the natural language toolkit, in: O'Reilly Media, Inc., 2009.

[43] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. John, N. Constant et al., "Universal sentence encoder", In: *arXiv preprint arXiv*: 1803.11175 2018.

[44] A. Akbik, B. Duncan, and R. Vollgraf, "Contextual string embeddings for sequence labeling", in: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638-1649, 2018.