User Modeling and User-Adapted Interaction
https://doi.org/10.1007/s11257-021-09290-y

Multimodal modeling of collaborative problem-solving facets in triads

Angela E. B. Stewart[1] · **Zachary Keirn**[1] · **Sidney K. D'Mello**[1]

Received: 9 December 2019 / Accepted in revised form: 7 January 2021

Abstract

Collaborative problem-solving (CPS) is ubiquitous in everyday life, including work, family, leisure activities, etc. With collaborations increasingly occurring remotely, next-generation collaborative interfaces could enhance CPS processes and outcomes with dynamic interventions or by generating feedback for after-action reviews. Automatic modeling of CPS processes (called facets here) is a precursor to this goal. Accordingly, we build automated detectors of three critical CPS facets—construction of shared knowledge, negotiation and coordination, and maintaining team function—derived from a validated CPS framework. We used data of 32 triads who collaborated via a commercial videoconferencing software, to solve challenging problems in a visual programming task. We generated transcripts of 11,163 utterances using automatic speech recognition, which were then coded by trained humans for evidence of the three CPS facets. We used both standard and deep sequential learning classifiers to model the human-coded facets from linguistic, task context, facial expressions, and acoustic–prosodic features in a team-independent fashion. We found that models relying on nonverbal signals yielded above-chance accuracies (area under the receiver operating characteristic curve, AUROC) ranging from .53 to .83, with increases in model accuracy when language information was included (AUROCS from .72 to .86). There were no advantages of deep sequential learning methods over standard classifiers. Overall, Random Forest classifiers using language and task context features performed best, achieving AUROC scores of .86, .78, and .79 for construction of shared knowledge, negotiation/coordination, and maintaining team function, respectively. We discuss application of our work to real-time systems that assess CPS and intervene to improve CPS outcomes.

Keywords Collaborative problem solving · Multimodal modelling · Language models · Educational technology

✉ Angela E. B. Stewart
angelast@andrew.cmu.edu

[1] University of Colorado Boulder, Boulder, USA

Published online: 02 February 2021

1 Introduction

Problem-solving involves identifying a sequence of operations to convert a given state to a goal state (Newell and Simon 1972). Collaborative problem-solving (CPS) involves two or more people working together to find a solution to a problem (OECD 2016). Be it school, work, service, leisure, or family environments, individuals are constantly engaging in CPS. In fact, CPS is considered a critical twenty-first century skill for productivity in an increasingly global workforce (Fiore et al. 2018; Graesser et al. 2018; OECD 2016). In addition, education practitioners have particularly emphasized the need for building skills for remote collaborations (OECD 2016; Schulze and Krumm 2017) as teams become distributed and schooling or working from home becomes the norm. Thus, CPS has become a pervasive part of our lives, whether we are face-to-face or not.

Given the ubiquity and importance of CPS, modern education has focused on teaching relevant skills, such as goal setting (Lai et al. 2017), idea sharing (Hao et al. 2017), and shared construction of solutions (Roschelle and Teasley 1995), often through short or long-term group projects (Graesser et al. 2018; OECD 2016). However, teams often fail to productively engage in these complex CPS skills, resulting in *process loss*, where teams do not live up to expectations (Kerr and Tindale 2004). In contrast, *process gain*, where team performance is greater than any combination of individual performance—or the whole is greater than the sum of its parts—is rare (Kerr and Tindale 2004).

Productive CPS proves even more difficult when interactions become virtual (as opposed to co-located). The rich social signals available in face-to-face interaction are muted when collaborating remotely (Alterman and Harsch 2017; Schulze and Krumm 2017). Poor audio quality, low video resolution, and lagging audio-visual signals dampen communication of social signals, like eye gaze and breathing patterns, which cue turn taking and other social interactions (Kendon 1967; Schulze and Krumm 2017). This can lead to difficulty coordinating action and maintaining engagement, reduced team cohesion, rapport, and consequently performance (Schulze and Krumm 2017).

There have been efforts to support CPS in computer-mediated environments by encouraging effective collaborative behaviors. For example, some systems give individuals feedback on their verbal participation (Calacci et al. 2016; Faucett et al. 2017; Samrose et al. 2018), interruption of teammates (Calacci et al. 2016; Faucett et al. 2017; Samrose et al. 2018), and attentional focus (Faucett et al. 2017; Gutwin et al. 2017; Schlösser et al. 2018). However, merely receiving feedback on these low-level behaviors might not provide individuals with sufficient insight into the underlying socio-cognitive processes necessary for successful CPS.

Systems that go beyond providing feedback on behavioral signals by dynamically responding to the unfolding collaborative process could help improve the outcomes of computer-mediated collaboration. Accordingly, recent work has focused on modeling socio-cognitive constructs related to effective CPS, such as rapport loss (Müller et al. 2018), empathy-skill (Ishii et al. 2018), team cohesion

(Hung and Gatica-Perez 2010), and argumentation (Lu et al. 2011; Prata et al. 2009; Rosé et al. 2008). Although these socio-cognitive constructs are crucial to successful collaborative interactions, they are not direct measures of CPS, which involves sharing ideas, negotiation among competing ideas, monitoring execution of a solution, and keeping the team motivated (Hao et al. 2017), where research is much more limited.

Some recent work has focused on detecting specific CPS behaviors, such as providing relevant information for the task or asking for clarification from language-based features (Flor et al. 2016; Hao et al. 2017). However, these studies rely on text-based interactions between collaborators, where the content of the chat is a direct representation of what was communicated. There is a dearth of work exploring CPS modeling in face-to-face or videoconferencing environments where individuals can see and hear each other, and communicate through language, gesture, voice tone, and body posture and gesture. These signals provide valuable insight into non-verbal communication in addition to the verbal content. For example, emotion-rich modalities, such as facial expressions (Littlewort et al. 2011) or acoustic–prosodic information (Eyben et al. 2013) could be useful for modeling CPS.

Accordingly, our work examines multimodal models of CPS, based on an empirically validated theoretical framework of three core CPS facets (Sun et al. 2020): construction of shared knowledge, negotiation/coordination, and maintaining team function. We use multimodal data including language, acoustic–prosodic features, facial features, body movement, and task information, collected from 32 triads engaged in a computer-mediated challenging visual programming task. As we review below, this is the first such study of its kind.

2 Background and related work

We first review frameworks of CPS, followed by models of collaborative behaviors, processes, and facets.

2.1 What is collaborative problem-solving?

The teamwork literature is vast and has been studied extensively. Broadly, teamwork refers to two or more people working together toward a shared goal. The five important components of effective teams include team leadership, mutual performance monitoring, backup behavior, adaptability, and team orientation (Salas et al. 2005). CPS is a specific form of teamwork where two or more people coordinate to solve a problem (OECD 2016; Roschelle and Teasley 1995). CPS skill has been defined in terms of core competencies (Hesse et al. 2015; OECD 2016) and effective actions (Andrews-Todd and Forsyth 2018; Cukurova et al. 2018; Nelson 1999) to enable progress toward the problem-solving goal and establish a positive collaborative environment. An early conception of CPS can be derived from Roschelle and Teasley (1995), who noted that CPS is fundamentally about building and maintaining a joint problem space. According to this model, language and actions during the

collaboration should enhance the mental model of the shared problem. Teammates should continuously monitor the collaboration for breakdowns in the shared conception and take actions to repair the joint problem space.

Two recent frameworks define specific facets of CPS. First, the Assessment and Teaching of Twenty-first Century Skills (ATC21S) framework (Griffin et al. 2012; Hesse et al. 2015) outlines a measurable and teachable set of cognitive and social skills pertaining to CPS. Social skills focus on interaction with teammates and persevering to complete the task (i.e., the collaboration part of CPS). This involves engaging in perspective-taking where a teammate considers the problem from another teammate's viewpoint and engaging in effective social regulation processes, where teammates negotiate and compromise as well as harness individual team members' strengths. Conversely, cognitive skills focus on managing the task itself (i.e., the problem-solving part of CPS). In order to effectively problem solve, teammates must engage in task regulation where they analyze the problem, make plans, execute them, revise plans, and move the collaboration forward. Finally, there should be learning and knowledge-building as a result of the collaboration.

Similarly, to the ATC21S framework, the Programme for International Student Assessment (PISA) framework (OECD 2016) defines three collaborative competencies that interact with four problem-solving processes, resulting in 12 levels of CPS skills (Graesser et al. 2018; Webb and Gibson 2015). The first collaborative competency involves establishing common ground where teammates should communicate their knowledge and ideas proactively while working to understand others' ideas and establishing shared meaning. The second is taking appropriate action where teammates should provide reasons to support their solution proposals and negotiate with others to achieve a consensual solution plan. The third involves maintaining a functioning team, which involves each teammate understanding their role in their team, monitoring for communication breakdowns, and adapting when a breakdown occurs. These CPS competencies interact with the following four problem-solving processes: (1) exploring and understanding the problem; (2) organizing and integrating information with personal knowledge; (3) planning and executing a solution, (4) monitoring the plan and reflecting on how to improve it.

2.2 Modeling collaborative behaviors, processes, and facets

Researchers interested in modeling team interaction have focused on low-level behaviors (e.g., speech rate), traits of the individual and team (e.g., empathy and cohesion), and CPS facets (e.g., sharing ideas). We discuss pertinent research in these three areas below.

2.2.1 Models of low-level behaviors

There is extensive work in modeling low-level behaviors pertaining to nonverbal communication (Latif et al. 2014). For example, visual focus of attention (Otsuka et al. 2018), which is key to managing turn taking, has been modeled using camera-based estimates of head pose and gaze position (Otsuka et al. 2018). Further,

end-of-turn and turn-taking prediction (de Kok and Heylen 2009; Dielmann et al. 2010; Jokinen et al. 2013), a signal describing how the conversational dynamics progress, has been modeled from acoustic–prosodic (de Kok and Heylen 2009; Dielmann et al. 2010), head pose (de Kok and Heylen 2009; Dielmann et al. 2010), gaze position (de Kok and Heylen 2009), and dialogue act features (de Kok and Heylen 2009; Dielmann et al. 2010).

Low-level signals in conversational scenarios have long been investigated in terms of coordination and interpersonal synergy, where behaviors like facial expressions and eye gaze become linked over time (Fusaroli et al. 2014; Krafft et al. 2016; Richardson et al. 2007). Behavioral coordination predicts maintaining common ground, establishing social bonding, and improving social interactions (Delaherche et al. 2012; Grafsgaard et al. 2018), all of which are key to CPS. Research has focused on eye gaze (Richardson et al. 2007), head and body movements (Amon et al. 2019; Duran and Fusaroli 2017; Grafsgaard et al. 2018), speech rate (Amon et al. 2019; Duran and Fusaroli 2017; Stewart et al. 2018), physiology (Palumbo et al. 2017), and facial expressions (Grafsgaard et al. 2018). For example, in one study involving CPS, deep neural networks were used to prospectively predict the speech rate of one team member from verbal and nonverbal behaviors of two other team members, up to 6 s in advance (Stewart et al. 2018).

2.2.2 Models of collaborative traits, socio-cognitive constructs, processes, and outcomes

Researchers have focused on using low-level signals to model stable traits related to collaboration. For example, traits like empathy-skill relate to the quality of collaborative interactions (Kelly and Barsade 2001) and have been modeled from multimodal traces, such as gaze and turn-taking dynamics (Ishii et al. 2018). Dominance, which can affect perceptions and motives in social interactions (Hall et al. 2005), has been modeled from a multimodal combination of turn-taking dynamics and visual activity (motion) (Aran and Gatica-Perez 2010). Related, individual leadership, which can influence team decisions, has been extensively modeled from pose-based estimates of visual focus of attention (Beyan et al. 2016a), dialog patterns (Sanchez-Cortes et al. 2010), or multimodal combinations of visual focus of attention, head, and body activity (Beyan et al. 2016b).

Similar to models of stable traits, socio-cognitive constructs in group scenarios have also been modeled from measurable behaviors. For example, comprehension of computer programs was modeled from gaze (Jermann and Sharma 2018), and social regulation from tabletop computer interaction patterns (Evans et al. 2016). Many researchers have taken a multimodal approach to modeling socio-cognitive constructs. For example, team cohesion has been modeled from dialog dynamics and visual activity (Hung and Gatica-Perez 2010), speaker influence from speech and head movement features (Nihei et al. 2014), and rapport from turn-taking, prosody, facial expression, and motion (Müller et al. 2018). Researchers have specifically chosen these socio-cognitive constructs because they are key to successful social interactions. For example, social regulation is key to managing collaborative learning outcomes (Evans et al. 2016), team cohesion is important to a sense of belonging

(Hung and Gatica-Perez 2010), speaker influence is related to discussion flow (Nihei et al. 2014), and rapport is key to relationship-building (Sinha and Cassell 2015).

In educational contexts, content produced by teams has been analyzed for collaborative learning processes and academically productive talk (Dyke et al. 2012; Rosé et al. 2008; Tegos et al. 2015, 2016). For example, research has focused on understanding group epistemic activity (Rosé et al. 2008; Yoo and Kim 2014) and argumentation (Lu et al. 2011; Prata et al. 2009; Rosé et al. 2008). Researchers have also successfully shown that academically productive talk can be supported in teammates through real-time interventions which monitor the unfolding conversation and use conversational agents to guide the discussion accordingly (Dyke et al. 2012; Tegos et al. 2015, 2016). For example, conversational agents have been used to encourage students to explicitly provide reasoning to support a solution and build upon their teammates' ideas (Tegos et al. 2015).

In addition to assessing collaborative learning processes, there have been efforts to model objective outcomes (Chopade et al. 2019; Murray and Oertel 2018; Subburaj et al. 2020; Vrzakova et al. 2020; Yoo and Kim 2014). For example, post-test scores (Stewart and D'Mello 2018), can be used as an objective measure of learning. Researchers often adopt a multimodal learning analytic approach and posit that utilizing behavioral traces from a variety of signals will outperform unimodal signals. For example, Yoo and Kim (2014) used multimodal behavioral patterns (language and interaction features) from online discussion groups to predict team-level grades. They found that acting as an information giver (as opposed to receiver) was positively correlated with project grades. Additionally, the use of positive emotion words and early discussion (as opposed to procrastinated discussion) positively correlated with grades. In face-to-face interactions (as opposed to online groups), hand, body, and face tracking have been used to infer proximity and position of group members as well as direction of attention to predict group grades in a project-based learning task (Spikol et al. 2018). Also in face-to-face collaborative learning, rule-based models on video, writing, and speech data have been used to predict solution correctness and domain expertise with 96% and 100% accuracy, respectively, with multimodal analytics outperforming unimodal signals (Oviatt and Cohen 2013).

Exemplary work on modeling CPS task performance comes from Murray and Oertel (2018), who predicted objective expert-rated task performance on a discussion-based CPS task. They trained a Random Forest classifier on acoustic–prosodic and linguistic features to predict task performance and achieved a mean-squared error of 64.4 (compared to a mean-prediction baseline of 79.3). Related, Chopade et al. (2019) regressed task success (binary successful or not) onto language features like cohesion and agreement. Their regression models explained about 17% of the variance in task success (Chopade et al. 2019). Finally, Subburaj et al. (2020) predicted binary task success from a multimodal combination of face, eye gaze, acoustic–prosodic, and task context information. They found that a multimodal combination of face, eye gaze, and task context features outperformed unimodal models and other multimodal combinations. Additionally, models that equally weighted behavioral signals from all teams outperformed, or performed equivalently to models that weighted teammates based on individual difference measures (e.g., personality), role on the team, or behaviors (e.g., verbosity).

The work of Vrzakova et al. (2020) also demonstrates how multimodal behavioral signals can be more informative in concert than alone. They correlated unimodal primitives (interaction with environment, speech, and body movement), and their bimodal and multimodal combinations with objective task performance and subjective team perceptions of the collaboration. They found that when unimodal primitives were correlated with outcomes, adding multimodal information significantly improved explanatory power.

Taken together, researchers have computationally modeled a large set of socio-cognitive constructs, collaborative processes, and collaborative outcomes from behavioral measures, especially using multimodal signals.

2.2.3 Models of collaborative problem-solving facets

Taking a step toward modeling CPS itself, prior work has focused on CPS skill assessment. However, researchers have chosen to use simplified interaction environments, presumably for precision of measurement. For example, many environments only allow communication through pre-defined responses (Chopade et al. 2018; Polyak et al. 2017; Rosen 2015; Stoeffler et al. 2018), which correspond to levels of particular CPS skills based on theoretical models of CPS, such as the ACT Holistic Framework (Camara et al. 2015) or the PISA framework (OECD 2016). While reliable and precise assessment are important goals, the ecological validity of these systems is low as real-world collaboration relies on open-ended communication with an effectively limitless set of possible responses.

CPS has been modeled in somewhat less restrictive environments where individuals used text chat to communicate. In such environments, simple language-based features that quantify the frequency of words and word phrases (n-grams), emoticons, and punctuation has been used to model CPS skills (Flor et al. 2016; Hao et al. 2017). Two studies relevant to our work used text chats in a STEM CPS task. In one study, researchers trained computer models on human annotations of four CPS facets (sharing ideas, negotiating ideas, regulating problem-solving activities, maintaining communication) (Hao et al. 2017). They pre-selected theoretically informative n-grams and emoticons to model the CPS facets using linear-chain conditional random fields on sequential text chats. They found that sequential modeling achieved an average accuracy of 73.2%, which outperformed a randomly shuffled baseline (accuracy of 29%) and slightly outperformed standard classifiers (accuracies of 66.9–71.9%). Rather than modeling the high-level facets, Flor et al. (2016) modeled 31 behavioral indicators of CPS, such as expressing agreement or disagreement with teammates. They used n-gram and punctuation frequencies, as well as automatically tagged dialog acts, achieving an accuracy of 60.3%, which beat the majority class baseline of 24.9%. Taken together, these studies demonstrate the feasibility of using language-based approaches to monitor CPS, at least with dyads engaged in text chats.

Stewart et al. (2019) went a step further by building fully automated models of three CPS facets in a videoconferencing environment, where participants could see and hear each other. They used data from the same challenging computer programming task analyzed here to predict expert codes for three CPS facets: construction

of shared knowledge, negotiation/coordination, and maintaining team function. Language-based Random Forest models were built using frequency counts of words and two-word phrases (bag of n-grams) from automatically transcribed speech, as well as features from the Linguistic Inquiry Word Count dictionary (Tausczik and Pennebaker 2010). Their models achieved accuracies, quantified as area under the receiver operating characteristic curve (AUROC), of .77 to .85 using n-grams, while word dictionary models achieved similar values (.73 to .82), both beating chance (AUROC = .50).

Taken together, language-based models have been successful at predicting CPS and thus serve as a starting point for our work. Beyond language, there is some work on multimodal modeling of CPS competencies in face-to-face interactions (Cukurova et al. 2020; Grover et al. 2016). Specifically, computer vision techniques have been used to estimate body pose (Cukurova et al. 2020; Grover et al. 2016), as well as where teammates are looking (Cukurova et al. 2020) to predict level of CPS competency (low, medium, or high). Although these works rely on face-to-face scenarios, they demonstrate that nonverbal features can too be useful in modeling CPS.

3 Current study: contribution, novelty, and research questions

We use spoken language, task context, facial expressions, body movement, and acoustic–prosodic features to automatically model three key CPS facets (construction of shared knowledge, negotiation/coordination, and maintaining team function) derived from a theoretical and empirically validated CPS framework (Sun et al. 2020) similar to those discussed above. We train our models on data collected from 32 triads engaging in a challenging visual programming task using a standard videoconferencing environment with audio and visual signals, as well as screen sharing. We compare verbal and nonverbal models, as well as combine the two. Additionally, we compare standard classifiers with deep sequential learning approaches.

3.1 Novelty and contribution

Our study is novel in several respects. For one, previous studies that considered remote, computer-supported interaction have restricted communication to text chats amongst dyads (Flor et al. 2016; Hao et al. 2017). In our study, triads collaborated on a CPS task using a standard videoconferencing interface. Collaborators could choose any communication medium, including language, gesture, facial expression, verbal tone, or mouse movements, yielding rich and unrestricted behaviors, etc. Further, in contrast to prior studies, we model CPS in triadic collaborations, which further complicates the interaction because a teammate must coordinate expertise, ideas, and skills with two other teammates rather than just one. Thus, we model CPS in triads collaborating on an open-ended task in an environment that supports multiple communication signals. This yields a complicated, genuine social interaction with an effectively limitless set of verbal and nonverbal behaviors.

We also go beyond unimodal language-based models (Flor et al. 2016; Hao et al. 2017; Stewart et al. 2019) by incorporating nonverbal signals. Multimodal approaches have been applied phenomena related to CPS (Sect. 2.2.2) or to levels of CPS competency (Sect. 2.2.3), but to our knowledge, this is the first work using multimodal feature sets to predict CPS facets, and in computer-mediated environments (as opposed to face-to-face). We specifically include task context, facial expression, and acoustic–prosodic features. These features might provide important emotional and social context to the collaboration that might not be detected from language alone. In other words, in addition to modeling what was said (language), we model how it was said (facial expression and acoustic–prosodic features), and what was occurring at the time (task context).

Finally, we contrast standard classifiers and deep sequential learning approaches. It should be noted that this has been done for unimodal language models (Hao et al. 2017), but not with multimodal models of CPS facets. Deep sequential models provide the potential benefit of capturing the temporal nature of this complex multimodal data in a way that standard classifiers cannot.

3.2 Research questions

We address three research questions.

RQ1: To what extent can behavioral signals be used to automatically model CPS facets? We must first establish feasibility of the unimodal behavioral signals to model CPS facets, before examining trade-offs of different approaches. Theoretically, language features will be predictive of CPS facets, due to the inherently verbal nature of CPS in a videoconferencing environment. However, features that index task context and nonverbal communication might also be predictive.

RQ2: Do deep sequential learning approaches improve prediction accuracy of CPS facets compared to standard machine learning classifiers? Our modeling task is quite complex. Individuals bring unique skills and attitudes to the team and produce behaviors that are influenced by and influence others. Therefore, perhaps deep sequential models that can capture temporal dependencies are needed to sufficiently represent our data. Conversely, it could be the case that deep sequential learning models yield no additional performance gains beyond static machine learning approaches like Random Forest classifiers, which can model interactivity and nonlinearity but not temporal dynamics.

RQ3: Do multimodal features improve modeling compared to unimodal feature sets? We hypothesize that multimodal feature sets index more rich communication patterns and thus will provide a boost in model performance. Thus, we explore how much, if at all, multimodal feature sets improve modeling accuracy above and beyond unimodal models. We also identify precisely what combinations of feature sets yield the best performance.

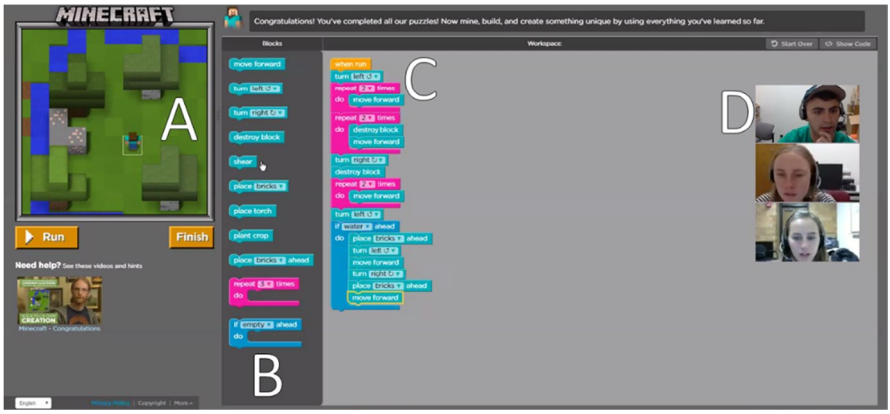

Fig. 1 Minecraft-themed Hour of Code from Code.org. Participants could visualize the results of running their code (A), a code bank of possible blocks to use (B), the code they generated (C) and their team's faces (D)

4 Data collection

4.1 Participants

Participants were 111 (63.1% female, average age = 19.4 years) undergraduate students from a medium-sized private, highly-selective, Midwestern U.S. university. Participants self-reported a variety of majors and were not specific to any department. Participants were 74.8% Caucasian, 9.9% Hispanic/Latino, 8.1% Asian, 2.7% Other, 0.9% Black, 0.9% American Indian/Native Alaskan; 2.7% did not report ethnicity. Participants were assigned to teams of three based on scheduling constraints, resulting in 37 teams. Nineteen participants from ten teams (27%) indicated they knew at least one person from their team prior to participation. No participant reported having prior programming experience. Participants were compensated with course credit.

4.2 Learning environment

Teams collaborated in a block-based programming environment, which is increasingly used to teach relevant computer science skills to students in formal and informal learning settings (Weintrop 2015; Weintrop and Wilensky 2016). We specifically used code.org's Minecraft-themed Hour of Code (Fig. 1) (Code Studio n.d.), which is an online game-based resource to learn basic computer programming principles in an hour. It uses Blockly (Fraser 2015), a visual programming language that represents lines of code (such as if statements) as interlocking blocks. Blocks only interlock with other syntactically correct blocks, allowing participants to focus on the coding logic and computing principles without considering syntax errors.

4.3 Procedure

4.3.1 Initial setup

Participants were each randomly assigned to one of three computer-enabled rooms. Each computer was connected to the internet via an Ethernet cable, rather than WiFi, for a better-quality signal. Each computer had a webcam and microphone for videoconferencing, and screen sharing capabilities through Zoom.[1] During the collaboration, one participant's screen was shared so that everyone viewed the same content. Separate audio tracks were recorded for each participant at 16,000 Hz. Videos of each participants' face and upper body and screen content were recorded at 25 Hz. Due to limitations with zoom, the videos of participants were quite small as evident in Fig. 1.

Each participant individually filled out demographic data including gender, age, major, and self-reported standardized test score (ACT and/or SAT). Participants also completed the validated 10-item version of the Big Five Inventory (BFI) (Gosling et al. 2003) to assess personality in five dimensions: extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience. These individual difference measures are not analyzed here, as they are not relevant to our Research Questions.

4.3.2 Introductory phase

After individually completing the surveys, teams completed five levels and viewed three accompanying videos that taught them how to use the programming environment as well as basic computer programming principles, such as loops and if statements. In these five levels, teams were required to build structures within the game and navigate around obstacles. One randomly assigned participant was tasked with controlling the team's actions in the environment. The other two participants were tasked with contributing to the collaboration. Participants were specifically instructed to collaborate as a team to complete the levels within 20 min.

After 20 min had elapsed or the team indicated they finished the five levels and viewed the three videos, screen sharing was disabled. Participants individually rated their satisfaction with their team's performance, communication, cooperation, and agreeableness as follows: "I am satisfied with my team's *performance* at completing the lessons," "I am satisfied with how we *communicated* with each other," "I am satisfied with how we mutually *cooperated* to complete the lessons," and "I am satisfied with how *agreeable* my teammates are." They indicated whether they were very dissatisfied (1), somewhat dissatisfied, slightly dissatisfied, slightly satisfied, somewhat satisfied, or very satisfied (6) on a six-point scale for all items.

[1] https://zoom.us.

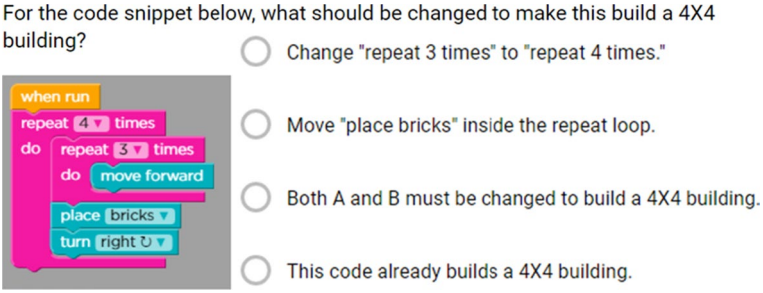

Fig. 2 Example post-test question

4.3.3 Challenging collaborative problem phase (main task)

After individually completing the ratings, screen sharing was enabled, and teams were tasked with collaboratively completing a challenging programming task in the same Hour of Code environment. The same team member who controlled interaction with the environment during the lessons also controlled the interaction during the coding challenge. In the challenge, teams were given 20 min to build a 4×4 brick building using at least one if statement and one repeat loop. At least three of the bricks must be built over water, and the code must total 15 code blocks or less. After 20 min had elapsed or the team sufficiently completed the coding challenge, screen sharing was disabled. Participants individually completed the same subjective measures of their team's performance, communication, cooperation, and agreeableness with the wording adapted for the challenge level.

4.3.4 Post-test and debriefing

Finally, participants individually completed a ten-item researcher-created multiple-choice test to assess their conceptual knowledge of the coding concepts (such as repeat loops and if statements). Each post-test item had a single correct answer out of four possible answers, and the possible range of scores was 0–100%. Figure 2 shows an example post-test question.

4.4 Data exclusion

We analyze the coding challenge only since the introductory lessons were primarily intended to familiarize participants with their teammates, the Hour of Code environment, and basic programming principles. Due to technical errors, four teams were removed because at least one participant in the team was missing an audio recording, and one team was removed because they were missing a screen recording (both audio and screen recordings are used in modeling). In all, 32 teams were analyzed.

4.5 Automated speech recognition (ASR)

Each participant's individual audio file was automatically transcribed using the IBM Watson Speech to Text service.[2] The service generates a transcript, start time, stop time, and transcription confidence for each utterance (the ASR also does utterance segmentation). Within a team, we interleaved transcripts (using the utterance start time) to produce one team-level transcript. Sometimes the ASR improperly segments utterances by splitting a single utterance (as identified by humans) into multiple segments. To remedy this, sequential utterances were combined into a single utterance if they belonged to the same speaker and there were less than 2 s (we also experimented with 1.5 s, 2 s, and 3 s thresholds) between the end of one utterance and the start of the next. In total, there were 11,163 utterances across the 32 teams for the 20-min challenging task.

To assess accuracy of the automatic transcription, we had a human transcribe a randomly selected 10% of the utterances, sampled from all participants. We computed word error rate (Hunt 1990), as: (substitutions + insertions + deletions)/(words in human transcript), and set word error rate to zero if the automated transcription indicated speech when there was none (6% of the selected utterances). The average word error rate was 45% (SD = 0.54), indicating considerable imperfections in the transcription, which increases challenge for automated modeling as discussed in Sect. 5.5.

4.6 Expert coding of CPS facets

We annotated teammates' language (utterances) using a theoretically grounded and empirically validated CPS framework (Sun et al. 2020). The framework defines three CPS facets: (1) construction of shared knowledge, (2) negotiation/coordination, and ((3) maintaining team function. Each facet has three observable verbal indicators that form the basis of the expert coding and make it an ideal choice for coding our data.

Construction of shared knowledge involves sharing ideas and expertise with other teammates and establishing shared understanding amongst the team. Verbal indicators include "proposes specific solutions," "talks about givens and constraints of the task," and "confirms understanding by asking questions/paraphrasing." Negotiation and coordination is an iterative process for developing and executing a team solution and revising the solution as necessary. It can be captured with the following verbal indicators: "provides reasons to support a potential solution," "responds to others questions/ideas," and "talks about results." Maintaining team function reflects a positive team dynamic where collaborators are conscious about being part of a team and proactively contribute to its success. Verbal indicators include the following: "asks if others have suggestions," "compliments or encourages others," and "gives instructions." In total, there were nine verbal indicators (three per facet). An

[2] https://www.ibm.com/watson/services/speech-to-text/.

Table 1 Example continuous dialog (human and automatic transcriptions) from a team interaction with the coded indicators and CPS facets. Speaker A is controlling the interaction with the Hour of Code environment

Speaker	Human transcription	Automated transcription	Coded indicators
A	Do you want to see if this works at all? Then we can figure out something if it doesn't	Do you want to see if this works at all then we can figure out something else it doesn't	Asks for suggestions (Maintain.)
C	Yeah	If	Responds to questions/ideas (Neg./Coord.)
B	We could uh…yeah, just try it. What we do in the beginning is we can like, we can turn right and then place the water blocks and he can turn around again and then just do what we were doing earlier	We can %HESITATION yeah I just try what we do in the beginning as we can light make him turn right place the fix that water block and turn around again and just do what we're doing earlier	Proposes specific solutions (Constr.)
A	We can just add another one	We can add another one	
A	Right	Right	
C	Like right there?	Like right there	Asking questions/paraphrasing (Constr.)
A	Uh-huh	Who	
B	Yeah, cause he's standing on that thing, and he comes down and he can't do anything from there	Yeah I can see standing and I think it sums down and he can't do anything from there	Provides reasons (Neg./Coord.); Talks about results (Neg./Coord.)
C	Yeah	Yeah	
A	Uh-huh	He	
B	So maybe in the beginning we can have him fill in that water block, so we don't have, we don't need that water statement then?	So maybe in the beginning we can have them fill in that water block so we don't and we don't need that that water statement and	Proposes specific solutions (Constr.); Provides reasons (Neg./Coord.)
A	So here?	So here	Asking questions/ paraphrasing (Constr.)
B	Yes. Instead of turning right, he can just turn…see where he's standing	Yes it instead of turning right you can just turn all CC standing	Responds to questions/ideas (Neg./Coord.); Gives instructions (Maintain.)

Constr. construction of shared knowledge, *Neg./Coord.* negotiation/coordination, *Maintain.* maintaining team function

Table 2 Percentage of instances that had 0, 1, 2, or all 3 indicators for each facet

No. indicators in utterance	Construction of shared knowledge (%)	Negotiation/coordination (%)	Maintaining team function (%)
0	66.96	84.72	90.14
1	31.31	14.69	9.791
2	1.711	.5912	.0717
All	.0179	.0000	.0000

example collaborative interaction with high-quality automated transcription and the expert-coded indicators is shown in Table 1.

Two experts were initially trained and viewed ten 90-s video clips from ten randomly selected teams. They counted the number of times each indicator occurred per clip. After reaching adequate reliability, coders were trained to code the individual automatically transcribed utterances for the presence of each indicator. Coders watched video recordings side-by-side with the transcripts and counted the number of times each indicator occurred in an utterance. They reached an agreement of .98 [Gwet's AC1 metric (Gwet 2014)] on two 5-min video samples consisting of 254 utterances. The 32 videos were then randomly assigned to the coders, who individually coded their videos.

The majority of the counts within an utterance were either 0 or 1 (average of 99.82% across all indicators), so we converted the resultant indicator counts to binary variables. For each facet, if all of the indicator counts were 0, then that utterance was coded as a 0. Otherwise, if at least one of the indicators occurred, it was coded as a 1. Distributions of the number of indicators present in each utterance are shown in Table 2. In total, 33%, 15%, and 10% of the utterances exhibited evidence of constructing shared knowledge, negotiation/coordination, and maintaining team function, respectively. In what follows, we investigated whether we can learn models to reproduce these utterance-level human codes in a team-generalizable fashion. We begin with language-based models (Sect. 5), compare them to include nonverbal channels (Sect. 6), followed by a detailed analysis of the most accurate models (Sect. 7).

5 Language models

We chose language as the initial modality for modeling the high-level CPS facets because language indexes the content of the collaboration, and the expert codes were based on verbal contributions (Sect. 4.6). We compared both standard classifiers using n-grams with recurrent neural networks with word embeddings.

5.1 Supervised classification with Random Forest classifiers

We chose to use a Random Forest classifier as it achieves equivalent or better accuracies than other standard classifiers, including Naïve Bayes, Logistic Regression, Support Vector Machine, and Adaboost.[3] We used a bag-of-n-grams approach, where occurrence counts of words and word phrases in the automatically transcribed utterance served as input features. Utterances were tokenized using the nltk (Bird and Loper 2004) tokenizer. We experimented with whether to perform word stemming on the n-grams using the nltk Snowball Stemmer (Porter 2001), and whether to remove stop words using the Glasgow Information Retrieval Group stop word list (Lo et al. 2005). We considered language models both with and without stemming, as well as with and without stop word removal. Neither stemming, nor removing stop words improved performance of our models, so we did not do either in our final models. In addition to n-grams, we also computed speech rate (words per second) as the total number of words in an utterance by the elapsed time (in seconds) of that utterance. Finally, we computed the transcription confidence (between 0 and 100%) provided by the IBM Watson ASR. In the case of merged utterances (Sect. 4.5), we took the average confidence value of the utterances involved.

We used a random under-sampling implementation from the imbalanced-learn library (Lemaitre et al. 2017) to account for class imbalance in the dataset. We also tested random oversampling and the synthetic minority oversampling technique (SMOTE), but found no improvement in prediction accuracy. Under sampling was only performed on the training set, and class distributions for the validation and testing sets were left unchanged.

5.2 Recurrent neural networks with word embeddings

We also trained long short-term memory (LSTM) recurrent neural networks, which is a special type of recurrent neural network that can learn long-term dependencies (Hochreiter and Schmidhuber 1997) by selectively retaining and forgetting information across input sequences. Each utterance was represented as a fixed-length sequence of word embedding vectors. We chose 14-word sequences as 90% of the utterances had 14 words or less. Shorter sequences were padded at the end. We used Global Vectors for Word Representations (gloVe) embeddings (Pennington et al. 2014) (100 dimensions) to represent each word and set the embedding weights to be trainable, such that the net would update them over time.

To reduce computation time, we fixed several of the parameters of the network and focused on training the weights instead (cross-validation procedure described

[3] We utilized the same cross-validation procedure described in Sect. 5.3 for these models. We specifically tuned the following hyperparameters for each classifier: Naïve Bayes—alpha, class prior distributions; Logistic Regression—regularization norm, regularization strength; Support Vector Machines—kernel, regularization strength; Adaboost—number of estimators, learning rate.

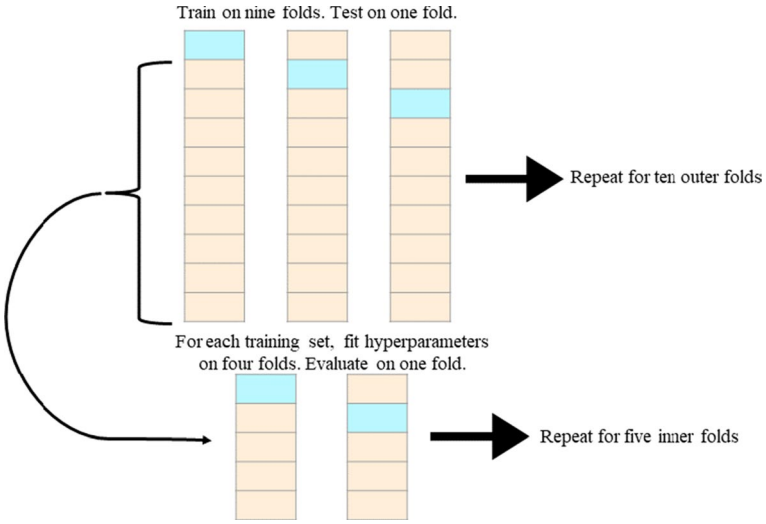

Fig. 3 Graphical representation of our nested cross-validation procedure

in Sect. 5.3). We used one hidden layer, containing 100 units to correspond with the embedding dimension of the vectors. Additionally, we used the Adamax optimizer, sigmoid activation function, a batch size of 128, and batch normalization. LSTMs were trained to 15 epochs. We performed random oversampling on the training set only.

5.3 Cross-validation and hyperparameter tuning

We used team-level 10-fold nested cross-validation (Fig. 3). By team-level, we mean that all the utterances for a given team were in the training set or testing set, but never both, which is important for team-level generalizability. On each of the ten test set iterations, a different fold was held out as the test set, and the other nine folds were used as the training set. This ensures that data used for training were not used for testing the model.

Within each of the ten iterations, the training set only was again split into five folds for hyperparameter tuning. A model was fit and scored using every combination of hyperparameters (see below) via a grid search for each of the five validation folds. The scores for each parameter combination across each validation fold were averaged, and the hyperparameters that resulted in the highest average area under the receiver operating characteristic curve (AUROC), which served as our accuracy metric, were preserved. A model was then fit on the full training set using these hyperparameters, and predictions were generated on the test fold. These predictions were pooled over the ten test folds before accuracy metrics were computed for that model. Note that the test set was not used to fit the model, but only to compute accuracy.

Table 3 AUROC for the Random Forest n-gram and LSTM word embedding models

Model	Construction of shared knowledge	Negotiation/ coordination	Maintaining team function
N-Grams	.86	.77	.77
Word embeddings	.86	.75	.75

For the Random Forest classifier, we tuned five hyperparameters. First, we varied the range of n-grams to include unigrams or bigrams. We chose not to test beyond bigrams because trigrams (and beyond) occurred in less than 1% of the utterances. Bigrams were filtered using pointwise mutual information (PMI) (Church and Hanks 1990; Lin 1998) to ensure that meaningful bigrams (such as "repeat loop") are preserved rather than bigrams that were merely the result of frequent words occurring next to one another (such as "next the"). We tested a low PMI of 2 and a high PMI of 4. We excluded n-grams that occurred in less than 0%, 1%, or 2% of the training utterances with the specific percentage included as a hyperparameter. This minimum frequency cutoff is important to ensure that n-grams that are specific to a single team are filtered. We also included the number of trees in the forest (100, 500, or 1000) and maximum depth of the trees (no maximum depth, 10, or 20) as hyperparameters. For the LSTM word embedding models, we fixed the parameters of the model (Sect. 5.2) and used the described procedure to train the weights.

5.4 Results

We selected the area under the receiver operating characteristic curve (AUROC) as our accuracy metric. AUROC assesses the true-positive and false-positive trade-off across prediction threshold values (Hanley and McNeil 1982). An AUROC of 0.5 reflects chance performance. Results are shown in Table 3.

For construction of shared knowledge, the Random Forest and LSTM with word embeddings models yield similar AUROC values, which reflect a 72% improvement over chance. The Random Forest models yield only slightly better predictions than the LSTM with word embeddings model for negotiation/coordination and maintaining team function. For both, the Random Forest models yield a 54% improvement over chance, while the word embeddings model yields a 50% improvement over chance. Since the LSTM with word embeddings do not improve performance over a simpler bag-of-n-grams approach, we use the latter in subsequent analyses.

5.5 Effect of transcription errors

We investigated the effect of transcription errors (Sect. 4.5, distribution of word error rates in Fig. 4) on prediction accuracy. Using the method described in Sect. 5.1, we trained a Random Forest classifier on the 1114 human-transcribed utterances and compared it to one trained on corresponding automated transcriptions. Results are shown in Table 4. The human and automated transcriptions yielded similar AUROC

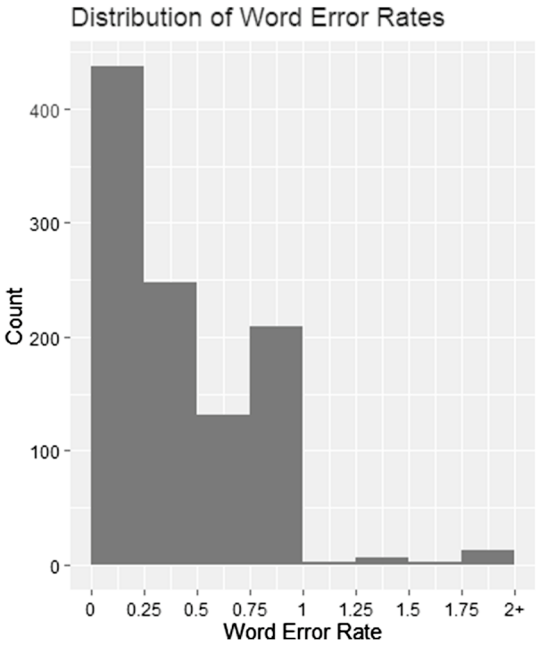

Fig. 4 Histogram of word error rates per utterance

Table 4 AUROC for the human and automated transcriptions of 10% of the utterances

Facet	Human	Automatic
Construction of shared knowledge	.84	.82
Negotiation/coordination	.69	.68
Maintaining team function	.75	.69

values for construction of shared knowledge and negotiation/coordination (i.e., 2.4% and 1.5%, boost in accuracy with using human transcriptions). However, for maintaining team function, the human transcriptions were 8.7% more accurate than the automated transcriptions. Therefore, automated transcription inaccuracies have a measurable but not excessive (average of 4.2%) effect on detection accuracies.

5.6 Discussion

We investigated the extent to which high-level collaborative problem-solving (CPS) facets could be automatically modeled from language. We compared two approaches: Random Forest models that adopted a bag of n-grams approach and LSTMs that used sequences of words in an utterance and word embeddings. Importantly, all models outperformed chance, which is notable given the low base rates for all three facets (33% for construction of shared knowledge, 15% for negotiation/coordination, and 10% for maintaining team function). The Random Forest classifier performed equivalently, or slightly better than the LSTMs, so it is preferable for

parsimony. That said, a deep sequential learning approach might be more useful on a larger dataset. We also demonstrate that automatic speech recognition errors only slightly affected classification performance, suggesting robustness of the approach.

Although this initial exploration demonstrates the feasibility of modeling CPS facets in a team generalizable way, these results should be interpreted in light of some self-imposed constraints. Specifically, we restricted our input to language alone, which is an incomplete representation of any social interaction. As such, the present results can be considered a useful starting point, with improvement possible as additional sources of information that index other aspects of the collaboration are added. For example, contextual information might provide insight into the team's actions in the environment and facial expressions or acoustic–prosodic features could indicate emotional states. These multimodal models are explored next.

6 Multimodal models

We extend our work to include multimodal data. Previous work detecting CPS facets has relied on language features (Flor et al. 2016; Hao et al. 2017), thus prediction accuracy might be limited to only what language can tell us. Accordingly, we focused on three additional modalities, which should increase model performance as they measure unique aspects of the collaboration. We use task context features to get a sense of team behaviors in the environment. We also included face and acoustic–prosodic features, which index emotional states (Eyben et al. 2013; Littlewort et al. 2011), turn-taking dynamics (Levitan et al. 2012), and coordination (Latif et al. 2014; Levitan et al. 2012). As before, we also compared standard classifiers with deep sequential learning methods, specifically Random Forest with LSTMs.

6.1 Feature engineering

We did not have direct access to log files since the Hour of Code environment is hosted by a third party. As an alternative, we used the screen recording (25 Hz) to extract high-level task context features that measure the teams' actions within the environment. We used a validated motion estimation algorithm (Westlund et al. 2015) to compute the proportion of pixels that change from a continuously updated background image, comprised of the previous four frames. This algorithm was applied to two areas of interest of the screen videos: the code runtime environment (A in Fig. 1) and the code bank and workspace (B and C in Fig. 1). Changes in the code bank/workspace indicate edits to a solution, while changes in the code runtime area indicate an attempt to test code. A lack of change in each signals deliberation or negotiation as teams discusses their next steps.

We used the videos of participant's faces (25 Hz) to extract facial and motion features. Facial features were extracted using Emotient (Littlewort et al. 2011), which is a commercialized version of the Computer Expression Recognition Toolbox

(CERT) computer vision software. Emotient provides likelihood estimates of the presence of 20 action units [specifically 1, 2, 4, 5, 6,7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, 26, 28, and 43 (Ekman 1997)]. Face motion was computed with the same motion estimate algorithm used on the screen videos. Action unit and motion estimates were computed for each frame.

For each participant, we used the openSMILE toolkit (Eyben et al. 2013) on the individual audio files to extract acoustic–prosodic features over 10 ms windows (i.e., 100 Hz). We extracted the following features: fundamental frequency, loudness, center frequency of the first through third formants, first through third formant amplitudes, harmonics to noise ratio, jitter, and shimmer.

6.2 Data aggregation and standardization

We aggregated features for each utterance because modeling was done at the utterance level. We created three feature sets. First, we created a task context feature set, which included the average value of the screen motion in each area of interest over the utterance duration. We additionally computed the proportion of the total collaborative session completed at each utterance (i.e., the start time of the utterance divided by the total collaboration time) to indicate progress in the session.

A facial feature set was formed from the action unit and face motion estimates. For motion estimates, we simply took the mean value of movement over the duration of the utterance. Motion estimates were z-score standardized across the 20-min session, per participant to account for individual differences. Because modeling is done at the team level, for a given utterance, we then took the mean motion estimates across the three team members, to get a team-level value. For each action unit, we similarly computed the mean value over the duration of the utterance, excluding values where the face could not be tracked due to occlusions, camera-positioning, or quick movements. If the face could not be tracked for the entirety of the utterance, the action unit value was marked as missing for that utterance. Action units were z-score standardized per participant, to account for individual differences. Missing values were then replaced with a zero (i.e., the mean for the participant since we are working with z-scored values). We then computed the mean of each action unit within the team to yield a single team-level estimate for each action unit. We also computed a facial feature validity score (ranging from 0 to 3) by summing the number of participants in the team where the face was tracked at any point during the utterance.

Finally, we generated an acoustic–prosodic feature set by computing the means of the 11 openSMILE features for the duration of the utterance. Similar to the motion and action unit estimates, we z-scored acoustic–prosodic features for each participant. These features were only computed for the current speaker of an utterance, rather than all team members, because they are not meaningful when a person is not speaking. In all, we had three task context features, 22 facial features, and 11 acoustic–prosodic features.

6.3 Random Forest classifier

We used a Random Forest classifier to model the three CPS facets at the utterance level using the same modeling procedure described in Sect. 5.3. We added the task context, facial, and acoustic–prosodic features to the utterance-level n-gram counts, where n-grams were generated using the same procedure described in Sect. 5.1. We focus on the Random Forest classifier because it outperformed or performed equivalently to other standard classifiers including Naïve Bayes, Logistic Regression, Support Vector Machine, and Adaboost.

6.4 Deep neural networks

In addition to using a standard Random Forest classifier, we tested two deep neural network models. The first was a feed-forward neural network (FFNN) with a single fully connected layer. The second was a long short-term memory network (LSTM). We chose to use FFNNs and LSTMs because they have been applied to similar data and modalities (Fan et al. 2015; Mao et al. 2015; Pham et al. 2017).

The LSTM was trained on sequences of inputs from the utterance-level data. A sequence of inputs is formed by using data from sequential utterances. We experimented with sequence lengths of two through five utterances. Additionally, the LSTM network used a tanh activation function (Keras, n.d.) with a softmax output layer (Bishop 2006). Our final LSTM models had one hidden layer with 32 units. We settled on one hidden layer after testing models with one, two, and three hidden layers, which achieved similar performance. Similarly, we settled on 32 units after comparing validation loss (mean square error) across models with 8, 32, and 128 units. The FFNN was trained on individual inputs from utterance. It employed a single hidden layer and used a leaky rectified linear unit activation function (Maas et al. 2013) between the hidden and softmax output layer. The LSTM model performed better than or equivalent to the FFNN, thus we focus on it in this paper.

We used team-level 10-fold cross-validation to train and test our models. Within each fold, we further split the data into 60% training, 30% validation, and 10% testing. Further, within each fold, we z-scored and normalized all features to a range of -1 to 1. We also tested normalizing with a -3 to 3 range, but found that it yielded similar performance to a -1 to 1 normalization scheme. We only used the training data to compute the statistics needed for the z-scoring and normalization (e.g., mean, standard deviation), which were subsequently applied to the validation and testing sets. Missing values were replaced with a value of five, which was chosen to be outside the normalized range. Further, a binary mask was used to indicate if data were missing for each modality as this was shown to be useful for training deep models with missing data (Lipton et al. 2016).

Neural networks use gradient descent and backpropagation to update the weights during each pass of the training (referred to as a training epoch). At each epoch, a loss function (mean-squared error) was computed and the weights were updated. We used a Nesterov Adam (Dozat 2016) adaptive learning rate algorithm, to tune the

Table 5 AUROC for the Random Forest (RF) n-gram and LSTM models

Modalities	Construction of shared knowledge		Negotiation/coordination		Maintaining team function	
	RF	LSTM	RF	LSTM	RF	LSTM
Language Based						
Language	.86	.80	.77	.75	.77	.73
Language + Task	.86	.80	.78	.75	.79	.73
Language + Task + Face + Acoustic–Prosodic	.85	.81	.76	.75	.75	.72
Nonverbal Only						
Task	.60	.56	.55	.53	.62	.58
Face	.60	.60	.55	.56	.54	.57
Acoustic–Prosodic	.74	.72	.64	.62	.61	.58
Task + Face + Acoustic–Prosodic	.75	.72	.65	.62	.65	.59

learning rate. We fixed the number of training epochs to 50 since the models converged within 50 epochs.

We experimented with batch normalization (Ioffe and Szegedy 2015), *l2* weight regularization (Ng 2004), and dropout (Srivastava et al. 2014) to prevent overfitting. Dropout had no discernible impact when combined with the other two methods, so we did not use dropout. Additionally, the default batch normalization and kernel regularization from Keras were adequate for our data.

6.5 Results

6.5.1 Main results

We used the language models as a starting point and incrementally added modalities. Using feature-level fusion, we first added task context features to assess what the team was doing in addition to what they were saying. Then, we included face and acoustic–prosodic features as measures of emotional content, turn-taking dynamics, and coordination, again using feature-level fusion. Results of these models are shown in Table 5.

We found that the multimodal Random Forest models tied or outperformed the LSTM models in almost all cases, suggesting that the additional computational expense of deep sequential learning models does not improve accuracy in our case. Random Forest models achieved an average AUROC of .75, .67, and .68 for construction of shared knowledge, negotiation/coordination, and maintaining team function, respectively, whereas LSTM models yielded average scores of 72, .65, and .64. Therefore, we focus on the Random Forest models for subsequent comparisons.

For the Random Forest models of negotiation/coordination and maintaining team function, the addition of task context features to language provided a marginal boost

Table 6 AUROC for the nonverbal models with additional feature sets

Modalities	Additional features included?	Construction of shared knowledge	Negotiation/ coordination	Maintaining team function	Improvement over restricted feature set
Face	No	.60	.55	.54	Yes
	Yes	.77	.65	.64	
Acoustic–Prosodic	No	.74	.64	.61	Yes
	Yes	.83	.70	.70	
Task + Face + Acoustic–Prosodic	No	.75	.65	.65	Yes
	Yes	.83	.70	.70	
Language + Task + Face + Acoustic–Prosodic	No	.85	.76	.75	No
	Yes	.84	.73	.73	

(1.3% for negotiation/coordination and 2.6% for maintaining team function) in classification accuracy. Thus, adding the context of what the team was doing to language data is marginally helpful to prediction accuracy for these two facets. However, there was no boost in classification accuracy for construction of shared knowledge, demonstrating that context is not as integral to detection of this facet.

Interestingly, adding nonverbal (face and acoustic–prosodic) information to language and task features slightly (average decrease of 2.9%) inhibited performance for all three facets, suggesting that these features might be adding noise or conflict with the language features. Taken altogether, we conclude that language provides the best foundation for modeling CPS facets, but task context features can marginally increase model performance.

Do the nonverbal behaviors contain any viable signals? We found that the language-free Random Forest models yielded at least 8% above-chance accuracies (i.e., all AUROC values at least .54), suggesting that they do contain some pertinent information. Models constructed from acoustic–prosodic features were the most accurate of the unimodal models, yielding an average 12.7% and 17.6% improvement over the task context and facial feature unimodal models. The multimodal model combining all nonverbal behavioral features yielded a slight boost in accuracy (average 3.2%) over the acoustic–prosodic features.

Given the above two sets of results, we trained an additional model that combined features from the best language-based (language + task) and unimodal nonverbal model (acoustic–prosodic). We found that this model yielded no improvement over the combined language and task models, with AUROC scores of .85, .78, and .77 for construction of shared knowledge, negotiation/coordination, and maintaining team function, respectively. Note, we also evaluated whether decision-level fusion improved results by averaging prediction probabilities of the unimodal models. This did not improve model accuracy irrespective of the modalities combined.

6.5.2 Supplemental analyses

We initially utilized a relatively small subset of face and acoustic–prosodic feature, so we expanded them to understand if more comprehensive features yielded model improvements. In particular, we considered additional aggregation strategies rather than simply averaging over the utterance duration. For each utterance, teammate, and facial feature (Sect. 6.1, 20 action units + 1 motion feature), we computed the following statistics over the duration of the utterance: standard deviation, minimum, maximum, skew, and kurtosis. We again, took the mean of each of these features across the three teammates, to get a single team-level value. This yielded 127 face features ([21 features × 6 aggregation strategies] + 1 validity feature).

We expanded the acoustic–prosodic feature set to include the additional openSMILE features, including those described in Sect. 6.1 and the following additions: ratio of energy of the first F0 harmonic to the energy of the second F0 harmonic, ratio of energy of the first F0 harmonic to the energy of the highest harmonic in the third formant, Hammarberg Index, first formant bandwidth, Mel Frequency Cepstral Coefficients 1 to 4, ratio of the summed energy from 50–1000 Hz and 1000–5000 Hz, spectral slope 0–500 and 500–1500, and spectral flux. For each of these features, we computed the mean over the utterance duration for the speaker only (same as Sect. 6.2). We additionally computed the additional aggregation statistics (standard deviation, minimum, maximum, skew, and kurtosis), resulting in 138 acoustic–prosodic features (23 features × 6 aggregation strategies).

We adopted the same modeling procedure described in Sect. 6.3 (Random Forest only) including the additional face and acoustic–prosodic features along with the language and task context features. Results are shown in Table 6. The additional features did not improve model performance when also using language information, likely due to the model overfitting in a given fold because of a relatively large number of features. However, when language is not included, the additional features do improve model accuracy. This suggests these additional features serve to provide complementary information in the absence of language.

6.6 Discussion

We examined the trade-offs amongst various multimodal models. We found that all of our models outperformed chance, demonstrating that nonverbal signals can be indeed be useful for this inherently verbal collaboration. Further, Random Forest models consistently performed as well as or better than LSTM models, which is particularly important to a real-time system, where computational resources might be limited.

Unsurprisingly, language-based models outperformed nonverbal models, since the basis of the coding scheme was verbal and language was the primary mode of communication in this task. However, we demonstrated that the nonverbal modalities do still contain viable signals, more so for acoustic–prosodic features. This is again unsurprising, given the emphasis on verbal communication. The acoustic–prosodic features might also generalize better than language features because they are

Fig. 5 Precision and recall across thresholds for the three CPS facets. The threshold where precision and recall match and F_1 is maximized is shown. *Constr. of Shared Knowledge* construction of shared knowledge, *Neg./Coord.* negotiation/coordination, *Maintain.* maintaining team function

not as specific to the particular task compared to the n-grams. However, this hypothesis needs further validation. A multimodal combination of the nonverbal feature sets yielded better prediction accuracy than unimodal models, with additional accuracy boosts when a more comprehensive feature set was included for face and acoustic–prosodic features.

Our best-performing models used a combination of language and task context features. The addition of task context features marginally boosted prediction accuracy for the negotiation/coordination and maintaining team function facets, above using language alone. This small boost in accuracy could suggest that the model was picking up on behaviors integral to interaction patterns. For example, a key indicator of negotiation/coordination is talking about results. In order for results to be discussed, the code must be run (i.e., motion in the runtime area of interest). Further, providing instructions to teammates is another indicator of maintaining team function. There should be reflected as a change in the code bank and workspace area according to the instruction being given. We did not obtain a comparative boost for shared knowledge construction, presumably because the verbal content was sufficient for this facet, and accuracy was already quite high.

Taken together, we demonstrate that all modality combinations perform better than chance, although language is the most powerful modality, with small boosts obtained from task context features. Therefore, we further examine the language and task context models for inclusion in real-time systems that support CPS.

7 Deeper analysis of language and task context model analyses

The Random Forest model combining language and task context feature yielded the best accuracy results. In this section, we aim to get a fuller picture of the conditions under which the model should be used, with an eye toward using them to drive real-time interventions (future work). Specifically, we investigate trade-offs between false-positive and false-negative rates and how well the models discriminate between the three facets.

7.1 Precision and recall

Real-time systems must balance the false-positive versus false-negative trade-off. Accordingly, we analyzed precision and recall by first computing binary predictions from the continuous probabilities output by the Random Forest model. Figure 5 shows how precision and recall vary as a function of the threshold. We analyzed two prediction thresholds. The first threshold is where precision and recall of the positive class (i.e., presence of each of the facets) are equal. The second threshold is where F_1, the harmonic mean of precision and recall, is maximized.

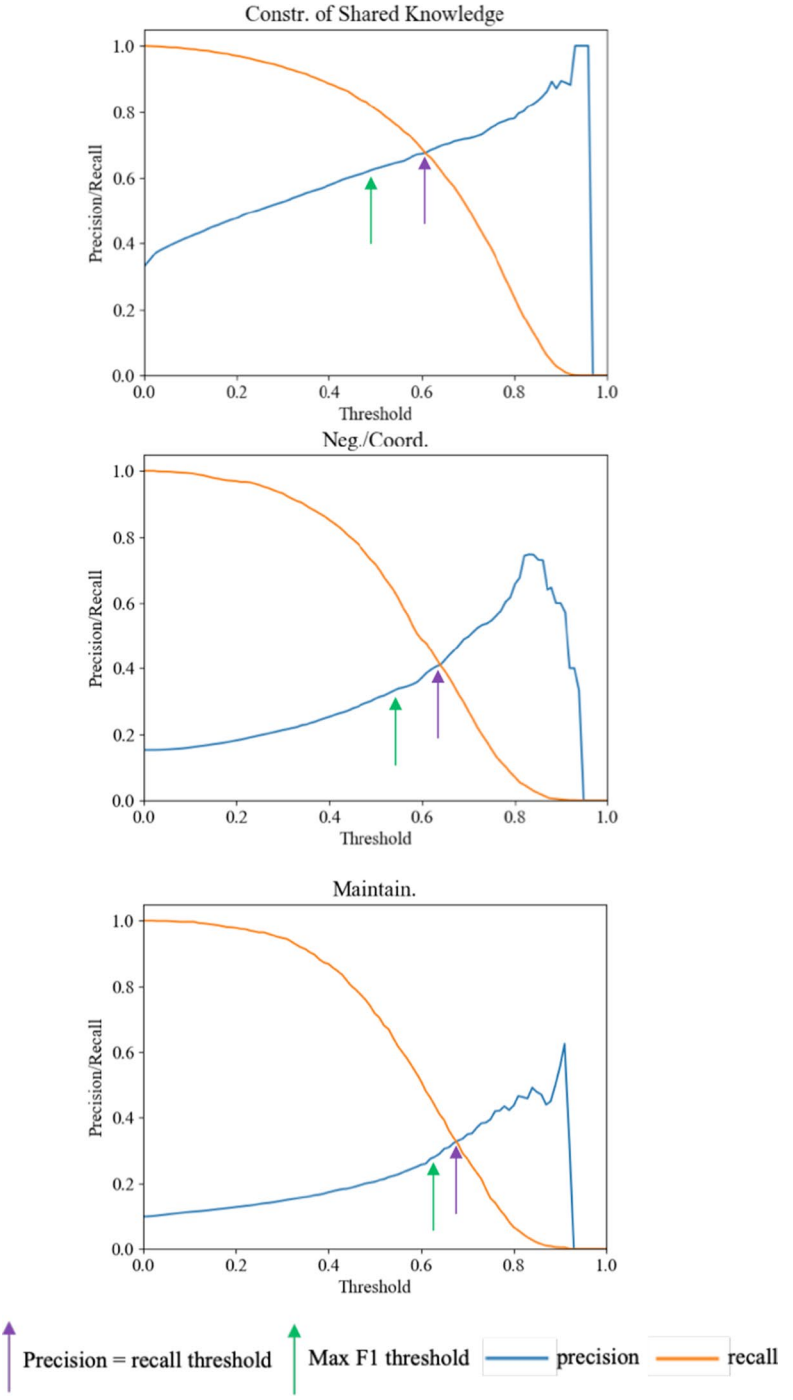
Constr. of Shared Knowledge
Precision/Recall
Threshold
Neg./Coord.
Precision/Recall
Threshold
Maintain.
Precision/Recall
Threshold
Precision = recall threshold
Max F1 threshold
precision
recall

Table 7 Precision, recall, and predicted rate at the threshold where precision and recall match and where F_1 is highest

	Threshold type	Threshold value	Precision	Recall	Pred. rate	Base rate
Constr. of shared knowledge	Precision = Recall	.61	.68	.67	.33	.33
	Max F_1	.48	.62	.83	.44	
Neg./Coord	Precision = Recall	.64	.41	.41	.15	.15
	Max F_1	.54	.33	.64	.29	
Maintain	Precision = Recall	.68	.33	.32	.10	.10
	Max F_1	.62	.28	.46	.17	
Neg./Coord. + Maintain	Precision = Recall	.61	.49	.48	.24	.24
	Max F_1	.52	.43	.67	.38	

Constr. of Shared Knowledge construction of shared knowledge, *Neg./Coord.* negotiation/coordination, *Maintain.* maintaining team function

Accuracy metrics are shown in Table 7 and confusion matrices in Fig. 6. As expected, there is clear evidence of a precision–recall trade-off. Specifically, recall is higher when using the threshold that maximizes F_1, but this comes at the cost of over predicting the positive class by an average of 66%, thus decreasing precision. That said, when comparing the matched precision and recall threshold to the maximized F_1 threshold, the increase in recall (average = 41%) is greater than the decrease in precision (average = 14%). The choice of model therefore depends on the specific application.

That said, the precision for the negotiation/coordination and maintaining team function models are still quite low, ostensibly due to the considerable class imbalance. To address this, we also trained a model after combining these two facets. We calculated the binary combined value as one if either negotiation/coordination or maintaining team function was one (note, both facets could be coded as a one). Otherwise, the binary combined value was zero. This resulted in an increased base rate 24%. This approach yielded an AUROC of .76 and increased the precision and recall quite substantially for the model that equated the two (see Table 7). Thus, it might be prudent to combine these two facets for applications that prefer higher accuracy at the cost of discriminability.

7.2 Discriminability

Discriminability pertains to the extent to which items that are supposed to be unrelated are actually unrelated. We computed correlations between the model predictions of each of the three facets to assess discriminability of our models. To do this, we first computed individual-level scores of the CPS facets as the average (across all utterances for a participant) of the binary human codes and model predictions based on the two aforementioned thresholding methods. We then computed the Pearson correlation among the facets, which are shown in Table 8. A lower correlation value corresponds to higher discriminability.

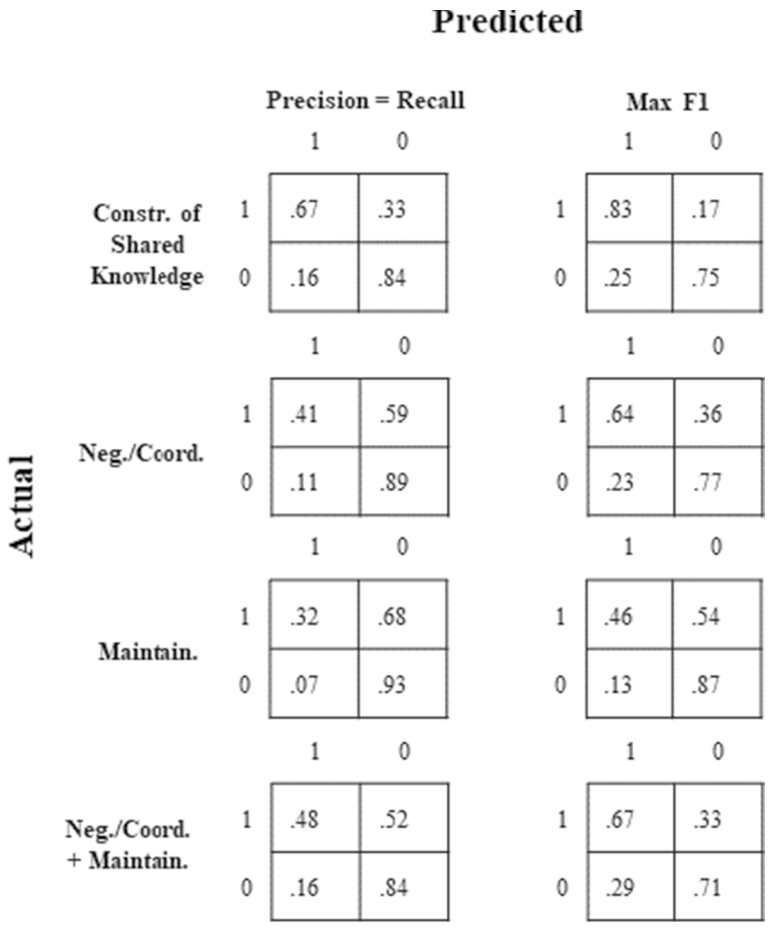

Fig. 6 Confusion matrices for the Precision = Recall and Maximum F1 thresholds are shown. *Constr. of Shared Knowledge* construction of shared knowledge, *Neg./Coord.* negotiation/coordination, *Maintain.* maintaining team function

Table 8 Discriminability (Pearson correlations) between human and model scores of the three CPS facets

	Human scores	Precision = Recall	Max F_1
Constr. of shared knowledge and Neg./Coord	− .07	.57	.50
Constr. of shared knowledge and Maintain	− .18	.57	.68
Neg./Coord. and Maintain	.40	.38	.59

Constr. of Shared Knowledge construction of shared knowledge, *Neg.Coord.* negotiation/coordination, *Maintain.* maintaining team function

We found that when comparing the matched precision–recall threshold to the max F_1 threshold, discriminability is either similar or higher. For the matched precision–recall threshold, discriminability between negotiation/coordination and maintaining team function mimics the human codes most similarly. However, discriminability is low between construction of shared knowledge and the other two facets, demonstrating that work remains to be done in increasing discriminability of our models.

7.3 Discussion

We expanded our analysis of the language and task context Random Forest models to better understand validity and the contexts in which they should be used. We first selected a threshold to transform continuous model prediction into binary predictions, which is also required for many real-world applications. We found that selecting a threshold where precision and recall matched is desirable for mimicking the base rate of the facets and thus controlling false-positive rate. This model is useful for situations where proper identification of negative instances is crucial. Conversely, the threshold where F_1 is maximized is most useful for situations where all positive cases are detected (i.e., high recall), even if it means incurring some false alarms. Next, we provided some evidence of the discriminability of our models, but work remains to be done. Specifically, discriminability for negotiation/coordination and maintaining team function mimics that of the human codes. However, correlations between construction of shared knowledge and the other two facets are high. This is most likely due to the model detecting generally positive CPS behaviors, rather than behaviors specific to construction of shared knowledge. Indeed, the behavioral indicators for this facet are related to overall CPS discussion and solution generation (Sect. 4.6).

8 General discussion

Collaborative problem-solving (CPS) is a key twenty-first century skill, crucial for people entering the modern workforce (Graesser et al. 2018; OECD 2016). However, teams often do not perform as well as they theoretically could, a phenomena known as process loss (Kerr and Tindale 2004). Process loss is even more pronounced in computer-mediated interactions where rich social signals available in face-to-face interactions are suppressed or nonexistent (Schulze and Krumm 2017). Our eventual goal is to build real-time systems that mitigate the problems of process loss in computer-mediated interactions by monitoring CPS and intervening appropriately. This requires automatic method to model key CPS facets, which was our goal.

Specifically, we developed fully automated detectors of three key CPS facets: construction of shared knowledge, negotiation/coordination, and maintaining team function. To do this, we leveraged data from teams collaborating in an open-ended virtual environment where they could communicate naturalistically. From this complex social interaction, we extracted a rich multimodal dataset of language, task

context, facial expression, and acoustic–prosodic features, which we used to predict the CPS facets. We compared deep sequential learning models to standard machine learning classifiers and also contrasted unimodal and multimodal models. We analyzed our best-performing model with respect to incorporating it into a real-time system, which is an item for future work.

8.1 Main findings

We had three research questions. Our first question asked: *to what extent can behavioral signals be used to automatically model CPS facets?* We analyzed four behavioral signals to address this question: language, task context, facial expression, and acoustic–prosodic. All of our models performed 8–72% better than a chance baseline (relative improvement), demonstrating that we can indeed automatically model CPS facets from behavioral patterns. Language-based models outperformed nonverbal models, which is unsurprising given that the coding scheme was primarily verbal. That said our best nonverbal model (with extended feature sets) performed 66%, 40%, and 40% better than chance (relative improvement) for the three facets. This is particularly because it suggests that *nonverbal* behavioral signals can be useful for modeling CPS facets based on a *verbal* coding scheme in a *verbally dominated* task.

We further examined the trade-offs of different modeling approaches for our second question: *do deep sequential learning approaches improve prediction accuracy of CPS facets compared to standard machine learning classifiers?* We used long short-term memory (LSTM) neural networks with word embeddings for language-only models, as well as LSTMs and feed-forward neural networks for combined language and nonverbal models. Neither of these deep sequential learning approaches yielded higher accuracies than the standard machine learning classifiers (Random Forest). This finding is particularly important for real-world applications where less time- and resource-intensive models are of value.

We also examined whether *multimodal features improve modeling compared to unimodal feature sets* for our third question. We found that the addition of task context features to language slightly improved classification accuracy for negotiation/coordination and maintaining team function, while accuracy for construction of shared knowledge stayed the same. This suggests our task features do indeed contextualize the language, even if to a small degree and only for some facets. Further, the addition of facial expression and acoustic–prosodic features to language and task context models decreased classification accuracy for all three facets, and are thus not useful when combined with previously predictive features. Finally, when relying on nonverbal features alone, there was also benefit to a multimodal approach with a combination of all nonverbal feature sets outperforming unimodal models.

It is unsurprising that multimodal models yielded slightly better (or at least equivalent) accuracies to unimodal models as different modalities provide different insight into the interaction. For example, language indexes verbal communication, task context features index actions taken by the team, and facial expression and acoustic–prosodic features index nonverbal communication and emotional aspects of the conversation. Improvement over unimodal models was most pronounced

for multimodal nonverbal-only models (task context, facial expression, and acoustic–prosodic), which yielded accuracies an average of 3.2% better than the best-performing unimodal model (acoustic–prosodic). By comparison, the best multimodal language-based model yielded an average boost in accuracy of only 1.3%, when compared to the unimodal language model. Since language is already a powerful modality for our task, adding nonverbal features provided limited insight.

We initially relied on a fairly limited set of features, particularly for facial expression and acoustic–prosodic signals. Accordingly, we expanded our analyses to include a more comprehensive set of features for these two modalities. For nonverbal models (both unimodal and multimodal), the expanded feature set improved prediction accuracy, but still did not outperform the language-only models. Importantly, the unimodal acoustic–prosodic model with expanded features performed equivalently to the multimodal task context, face, and acoustic–prosodic model suggesting that prediction power is primarily in paraverbal signals.

Our best-performing model used a combination of language and task context features, presumably because it combined modalities that measure verbal communication and action in the collaborative environment. Can it be used for real-time interventions, which might require binary predictions? At the threshold that maximizes F_1 score, recall was quite high; however, false-positive rate was inflated and discriminability was low. Thus, this model is most useful in detecting general positive CPS behaviors rather than specific facets. We successfully limited facet over prediction (and false-positive rate) using a threshold where precision and recall were matched. This also improved the discriminability of the model, but it is still somewhat high, an item that needs to be addressed in the future.

8.2 Limitations and future work

Our work has limitations that must be addressed in the future. First, our dataset is relatively small (32 teams) and contains little ethnic, socioeconomic, or age diversity. Thus, this limits claims of generalizability. We are currently working to remedy this limitation by exploring an extended dataset of teams from multiple universities with more ethnic and socioeconomic diversity.

Second, we only model CPS during a single collaborative task. While our method is likely to generalize to other tasks, the specific model might not as the vocabulary used by teams is specific to this task. We are also working to address this limitation by modeling CPS across multiple tasks.

Third, our models were trained in the lab, which limits distractions and produces relatively clean data signals (e.g., background noise is limited). We are currently working to address this concern with artificiality of context by collecting data on teams engaged in remote CPS from their homes using their own equipment.

Fourth, our deep sequential learning approaches did not outperform standard machine learning classifiers. It is likely that gains can be expected when using other state-of-the-art language models, such as Bidirectional Encoder Representations from Transformers (BERT) models (Kenton et al. 2017). We are exploring this approach on our data while also investigating how to improve generalizability,

discriminability, and model accuracy using these state-of-the-art deep sequential learning architectures.

8.3 Conclusion

We developed multimodal, team-generalizable models of three key CPS facets: construction of shared knowledge, negotiation/coordination, and maintaining team function. We modeled these CPS facets in a computer-mediated videoconferencing environment where teams were free to use language, gesture, voice tone, and facial expression to communicate. Thus, we take a critical step forward in automated detection of high-level CPS facets in open-communication collaboration environments. The next step is to deploy these models in real-time systems, for example, by providing teammates formative on CPS facets based on the model assessments.

References

Alterman, R., Harsch, K.: A more reflective form of joint problem solving. Int. J. Comput. Support. Collab. Learn. **12**(1), 9–33 (2017). https://doi.org/10.1007/s11412-017-9250-1

Amon, M.J., Vrzakova, H., D'Mello, S.K.: Beyond dyadic coordination: multimodal behavioral irregularity in triads predicts facets of collaborative problem solving. Cogn. Sci. **43**(10), e12787 (2019). https://doi.org/10.1111/cogs.12787

Andrews-Todd, J., Forsyth, C.M.: Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. Comput. Hum. Behav. (2018). https://doi.org/10.1016/j.chb.2018.10.025

Aran, O., Gatica-Perez, D.: Fusing audio-visual nonverbal cues to detect dominant people in group conversations. In: 2010 20th International Conference on Pattern Recognition, pp. 3687–3690 (2010). https://doi.org/10.1109/ICPR.2010.898

Beyan, C., Capozzi, F., Becchio, C., Murino, V.: Identification of emergent leaders in a meeting scenario using multiple kernel learning. In: Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction, pp. 3–10 (2016a)

Beyan, C., Carissimi, N., Capozzi, F., Vascon, S., Bustreo, M., Pierro, A., Becchio, C., Murino, V.: Detecting emergent leader in a meeting environment using nonverbal visual features only. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 317–324 (2016b). https://doi.org/10.1145/2993148.2993175

Bird, S., Loper, E.: NLTK: The natural language toolkit. In: Proceedings of the Association for Computational Linguistics 2004 on Interactive Poster and Demonstration Sessions, 31-es (2004). https://doi.org/10.3115/1219044.1219075

Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Berlin (2006)

Calacci, D., Lederman, O., Shrier, D., Pentland, A.S.: Breakout: an open measurement and intervention tool for distributed peer learning groups (2016). CoRR, abs/1607.0. http://arxiv.org/abs/1607.01443

Camara, W., O'Connor, R., Mattern, K., Hanson, M.A.: Beyond academics: a holistic framework for enhancing education and workplace success. ACT Research Report Series. 2015(4). ACT, Inc. (2015)

Chopade, P, Edwards, D., Khan, S.M., Andrade, A., Pu, S.: CPSX: using AI-machine learning for mapping human–human interaction and measurement of CPS teamwork skills. In: 2019 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–6 (2019). https://doi.org/10.1109/HST47167.2019.9032906

Chopade, P., Stoeffler, K.M., Khan, S., Rosen, Y., Swartz, S., von Davier, A.: Human-Agent Assessment: Interaction And Sub-Skills Scoring For Collaborative Problem Solving. In: Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B.

(eds.) C Penstein Rosé, pp. 52–57. Artificial Intelligence in Education. Springer International Publishing, Berlin (2018)
Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1), 22–29 (1990)
Code Studio. (n.d.). Retrieved April 1, 2018, from https://studio.code.org/s/mc/stage/1/puzzle/1
Cukurova, M., Luckin, R., Millán, E., Mavrikis, M.: The NISPI framework: analysing collaborative problem-solving from students' physical interactions. Comput. Educ. **116**, 93–109 (2018). https://doi.org/10.1016/j.compedu.2017.08.007
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). https://doi.org/10.1145/3375462.3375484
de Kok, I., Heylen, D.: Multimodal end-of-turn prediction in multi-party meetings. In: Proceedings of the 2009 International Conference on Multimodal Interfaces, pp. 91–98 (2009). https://doi.org/10.1145/1647314.1647332
Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: a survey of evaluation methods across disciplines. IEEE Trans. Affect. Comput. **3**(3), 349–365 (2012). https://doi.org/10.1109/T-AFFC.2012.12
Dielmann, A., Garau, G., Bourlard, H.: Floor holder detection and end of speaker turn prediction in meetings. In: Proceedings of the International Conference on Speech and Language Processing, Interspeech (2010)
Dozat, T.: Incorporating nesterov momentum into adam. In: Proceedings of the International Conference on Learning Representations (2016)
Duran, N.D., Fusaroli, R.: Conversing with a devil's advocate: interpersonal coordination in deception and disagreement. PLoS ONE **12**(6), e0178140 (2017). https://doi.org/10.1371/journal.pone.0178140
Dyke, G., Adamson, D., Howley, I., Penstein Rosé, C.: Towards academically productive talk supported by conversational agents. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) Intelligent Tutoring Systems, pp. 531–540. Springer, Berlin (2012)
Ekman, R.: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press, Oxford (1997)
Evans, A. C., Wobbrock, J. O., Davis, K.: Modeling collaboration patterns on an interactive tabletop in a classroom setting. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 860–871 (2016). https://doi.org/10.1145/2818048.2819972
Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia (MM'13), pp. 835–838 (2013). https://doi.org/10.1145/2502081.2502224
Fan, B., Wang, L., Soong, F.K., Xie, L. Photo-real talking head with deep bidirectional LSTM. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4884–4888 (2015). https://doi.org/10.1109/ICASSP.2015.7178899
Faucett, H.A., Lee, M.L., Carter, S.: I should listen more: real-time sensing and feedback of non-verbal communication in video telehealth. In: Proceedings of the ACM on Human–Computer Interaction 1(CSCW), pp. 44:1–44:19 (2017). https://doi.org/10.1145/3134679
Fiore, S.M., Graesser, A., Greiff, S.: Collaborative problem-solving education for the twenty-first-century workforce. Nat. Hum. Behav. **2**(6), 367–369 (2018). https://doi.org/10.1038/s41562-018-0363-y
Flor, M., Yoon, S.-Y., Hao, J., Liu, L., von Davier, A.: Automated classification of collaborative problem solving interactions in simulated science tasks. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 31–41 (2016)
Fraser, N. (2015). Ten things we've learned from Blockly. In: Proceedings of the 2015 IEEE Blocks and Beyond Workshop, pp. 49–50. https://doi.org/10.1109/BLOCKS.2015.7369000
Fusaroli, R., Rkaczaszek-Leonardi, J., Tylén, K.: Dialog as interpersonal synergy. New Ideas Psychol. **32**, 147–157 (2014)
Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the Big-Five personality domains. J. Res. Pers. **37**(6), 504–528 (2003). https://doi.org/10.1016/S0092-6566(03)00046-1

Graesser, A.C., Fiore, S.M., Greiff, S., Andrews-Todd, J., Foltz, P.W., Hesse, F.W.: Advancing the science of collaborative problem solving. Psychol. Sci. Public Interest **19**(2), 59–92 (2018). https://doi.org/10.1177/1529100618808244

Grafsgaard, J., Duran, N., Randall, A., Tao, C., D'Mello, S.: Generative multimodal models of nonverbal synchrony in close relationships. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 195–202 (2018). https://doi.org/10.1109/FG.2018.00037

Griffin, P., Care, E., McGaw, B.: The changing role of education and schools. In: Griffin, P., McGaw, B., Care, E. (eds.) Assessment and Teaching of 21st Century Skills, pp. 1–15. Springer, Dordrecht (2012). https://doi.org/10.1007/978-94-007-2324-5_1

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 and Knowledge, pp. 516–517 (2016). https://doi.org/10.1145/2883851.2883877

Gutwin, C., Bateman, S., Arora, G., Coveney, A.: Looking away and catching up: dealing with brief attentional disconnection in synchronous groupware. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 2221–2235 (2017). https://doi.org/10.1145/2998181.2998226

Gwet, K.L.: Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Advanced Analytics, LLC, Gaithersburg (2014)

Hall, J.A., Coats, E.J., LeBeau, L.S.: Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. Psychol. Bull. **131**(6), 898–924 (2005). https://doi.org/10.1037/0033-2909.131.6.898

Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1), 29–36 (1982). https://doi.org/10.1148/radiology.143.1.7063747

Hao, J., Chen, L., Flor, M., Liu, L., von Davier, A.A.: CPS-Rater: automated sequential annotation for conversations in collaborative problem-solving activities. ETS Res. Rep. Ser. **2017**(1), 1–9 (2017). https://doi.org/10.1002/ets2.12184

Hesse, F., Care, E., Buder, J., Sassenberg, K., Griffin, P.: A framework for teachable collaborative problem solving skills. In: Griffin, P., Care, E. (eds.) Assessment and Teaching of 21st Century Skills: Methods and Approach, pp. 37–56. Springer, Dordrecht (2015). https://doi.org/10.1007/978-94-017-9395-7_2

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

Hung, H., Gatica-Perez, D.: Estimating cohesion in small groups using audio–visual nonverbal behavior. IEEE Trans. Multimedia **12**(6), 563–575 (2010)

Hunt, M.J.: Figures of merit for assessing connected-word recognisers. Speech Commun. **9**(4), 329–336 (1990). https://doi.org/10.1016/0167-6393(90)90008-W

Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning, pp. 448–456 (2015). PMLR

Ishii, R., Otsuka, K., Kumano, S., Higashinaka, R., Tomita, J.: Analyzing Gaze behavior and dialogue act during turn-taking for estimating empathy skill level. In: Proceedings of the 2018 on International Conference on Multimodal Interaction, pp. 31–39 (2018). https://doi.org/10.1145/3242969.3242978

Jermann, P., Sharma, K.: Gaze as a proxy for cognition and communication. In: 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), pp. 152–154 (2018)

Jokinen, K., Furukawa, H., Nishida, M., Yamamoto, S.: Gaze and turn-taking behavior in casual conversational interactions. ACM Trans. Interact. Intell. Syst. **3**(2), 12:1-12:30 (2013). https://doi.org/10.1145/2499474.2499481

Kelly, J.R., Barsade, S.G.: Mood and emotions in small groups and work teams. Organ. Behav. Hum. Decis. Process. **86**(1), 99–130 (2001)

Kendon, A.: Some functions of gaze-direction in social interaction. Acta Physiol. **26**, 22–63 (1967). https://doi.org/10.1016/0001-6918(67)90005-4

Kenton, M. C., Kristina, L., Devlin, J.: BERT paper (2017). http://arxiv.org/abs/1810.04805 [Cs]

Keras. (n.d.). Retrieved May 2, 2018, from https://github.com/keras-team/keras

Kerr, N.L., Tindale, R.S.: Group performance and decision making. Annu. Rev. Psychol. **55**(1), 623–655 (2004). https://doi.org/10.1146/annurev.psych.55.090902.142009

Krafft, P.M., Baker, C.L., Tenenbaum, J.B., et al.: Modeling human ad hoc coordination. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)

Lai, E., DiCerbo, K., Foltz, P.: Skills for today: what we know about teaching and assessing collaboration. Pearson (2017)

Latif, N., Barbosa, A.V., Vatiokiotis-Bateson, E., Castelhano, M.S., Munhall, K.G.: Movement coordination during conversation. PLoS ONE **9**(8), 1–10 (2014). https://doi.org/10.1371/journal.pone.0105036

Lemaitre, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. **18**(1), 559–563 (2017)

Levitan, R., Gravano, A., Willson, L., Benus, S., Hirschberg, J., Nenkova, A. Acoustic–prosodic entrainment and social behavior. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 11–19 (2012)

Lin, D.: Extracting collocations from text corpora. In: First Workshop on Computational Terminology, pp. 57–63 (1998)

Lipton, Z.C., Kale, D.C., Wetzel, R.: Directly modeling missing data in sequences with RNNs: improved classification of clinical time series. In: Doshi-Velez, F., Fackler, J., Kale, D., Wallace, B., Wiens, J. (eds.) Proceedigns of Machine Learning Research, pp. 253–270. PMLR (2016)

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M.: The computer expression recognition toolbox (CERT). In: 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011, pp. 298–305 (2011). https://doi.org/10.1109/FG.2011.5771414

Lo, R.T.-W., He, B., Ounis, I.: Automatically building a stopword list for an information retrieval system. In: Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR), 5, 17–24 (2005)

Lu, J., Chiu, M.M., Law, N.W.: Collaborative argumentation and justifications: a statistical discourse analysis of online discussions. Comput. Hum. Behav. **27**(2), 946–955 (2011). https://doi.org/10.1016/j.chb.2010.11.021

Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the International Conference on Machine Learning (2013)

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (M-RNN). In: Proceedings of the 2015 International Conference on Learning Representations (2015)

Müller, P., Huang, M.X., Bulling, A.: Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In: 23rd International Conference on Intelligent User Interfaces, pp. 153–164 (2018)

Murray, G., Oertel, C.: Predicting group performance in task-based interaction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 14–20 (2018). https://doi.org/10.1145/3242969.3243027

Nelson, L.M.: Collaborative problem solving. Instr. Des. Theories Models New Paradigm Instr. Theory **2**, 241–267 (1999)

Newell, A., Simon, H.A., et al.: Human Problem Solving, vol. 104, Issue 9. Prentice-Hall, Englewood Cliffs (1972)

Ng, A.Y.: Feature selection, L1 vs. L2 regularization, and rotational invariance. In: Proceedings of the Twenty-First International Conference on Machine Learning, p. 78 (2004). https://doi.org/10.1145/1015330.1015435

Nihei, F., Nakano, Y.I., Hayashi, Y., Hung, H.-H., Okada, S.: Predicting influential statements in group discussions using speech and head motion information. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 136–143 (2014). https://doi.org/10.1145/2663204.2663248

OECD.: PISA 2015 Results (Volume I): excellence and equity in education, PISA, OECD Publishing, Paris. (2016). https://doi.org/10.1787/9789264266490-en

Otsuka, K., Kasuga, K., Köhler, M.: Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 191–199 (2018). https://doi.org/10.1145/3242969.3242973

Oviatt, S., Cohen, A. (2013). Written and multimodal representations as predictors of expertise and problem-solving success in mathematics. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 599–606. https://doi.org/10.1145/2522848.2533793

Palumbo, R.V., Marraccini, M.E., Weyandt, L.L., Wilder-Smith, O., McGee, H.A., Liu, S., Goodwin, M.S.: Interpersonal autonomic physiology: a systematic review of the literature. Personal. Soc. Psychol. Rev. **21**(2), 99–141 (2017). https://doi.org/10.1177/1088868316628405

Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

Pham, H.X., Cheung, S., Pavlovic, V.: Speech-driven 3D facial animation with implicit emotional awareness: a deep learning approach. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops **2017**, 2328–2336 (2017). https://doi.org/10.1109/CVPRW.2017.287

Polyak, S.T., von Davier, A.A., Peterschmidt, K.: Computational psychometrics for the measurement of collaborative problem solving skills. Front. Psychol. **8**, 2029 (2017). https://doi.org/10.3389/fpsyg.2017.02029

Porter, M.F.: Snowball: A Language for Stemming Algorithms (2001). https://api.semanticscholar.org/CorpusID:59634627

Prata, D.N., Baker, R.S.J., Costa, E.B., Rosé, C.P., Cui, Y., De Carvalho, A.M.J.B.: Detecting and understanding the impact of cognitive and interpersonal conflict in computer supported collaborative learning environments. In: International Working Group on Educational Data Mining (2009)

Richardson, D.C., Dale, R., Kirkham, N.Z.: The art of conversation is coordination. Psychol. Sci. **18**(5), 407–413 (2007). https://doi.org/10.1111/j.1467-9280.2007.01914.x

Roschelle, J., Teasley, S.D.: The construction of shared knowledge in collaborative problem solving. In: O'Malley, C. (ed.) Computer Supported Collaborative Learning, pp. 69–97. Springer, Berlin (1995)

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F.: Analyzing collaborative learning processes automatically: exploiting the advances of computational linguistics in computer-supported collaborative learning. Int. J. Comput. Support. Collab. Learn. **3**(3), 237–271 (2008). https://doi.org/10.1007/s11412-007-9034-0

Rosen, Y.: Computer-based assessment of collaborative problem solving: exploring the feasibility of human-to-agent approach. Int. J. Artif. Intell. Educ. **25**(3), 380–406 (2015). https://doi.org/10.1007/s40593-015-0042-3

Salas, E., Sims, D.E., Burke, C.S.: Is there a "Big Five" in teamwork? Small Group Res. **36**(5), 555–599 (2005). https://doi.org/10.1177/1046496405277134

Samrose, S., Zhao, R., White, J., Li, V., Nova, L., Lu, Y., Ali, M.R., Hoque, M.E.: CoCo: collaboration coach for understanding team dynamics during video conferencing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(4), 160:1-160:24 (2018)

Sanchez-Cortes, D., Aran, O., Mast, M.S., Gatica-Perez, D.: Identifying emergent leadership in small groups using nonverbal communicative cues. Int. Conf. Multimodal Interfaces Workshop Mach. Learn. Multimodal Interact. **39**(1–39), 4 (2010). https://doi.org/10.1145/1891903.1891953

Schlösser, C., Harrer, A., Kienle, A.: Supporting dyadic chat communication with eye tracking based reading awareness. In: 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), pp. 149–151 (2018). https://doi.org/10.1109/ICALT.2018.00042

Schulze, J., Krumm, S.: The "virtual team player": a review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. Organ. Psychol. Rev. **7**(1), 66–95 (2017). https://doi.org/10.1177/2041386616675522

Sinha, T., Cassell, J.: We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In: Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And InfLuence, pp. 13–20 (2015)

Spikol, D., Ruffaldi, E., Dabisias, G., Cukurova, M.: Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. J. Comput. Assist. Learn. **34**(4), 366–377 (2018). https://doi.org/10.1111/jcal.12263

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)

Stewart, A.E.B., D'Mello, S.K.: Connecting the dots towards collaborative AIED: linking group makeup to process to learning. In: International Conference on Artificial Intelligence in Education, pp. 545–556 (2018)

Stewart, A.E.B., Keirn, Z.A., D'Mello, S.K.: Multimodal modeling of coordination and coregulation patterns in speech rate during triadic collaborative problem solving. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 21–30 (2018). https://doi.org/10.1145/3242969.3242989

Stewart, A.E.B., Vrzakova, H., Sun, C., Yonehiro, J., Stone, C.A., Duran, N.D., Shute, V., D'Mello, S.K.: I say, you say, we say: using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving. Proc. ACM Hum. Comput. Interact. **3**, 19 (2019). https://doi.org/10.1145/3359296

Stoeffler, K., Rosen, Y., Bolsinova, M., von Davier, A.: Gamified assessment of collaborative skills with chatbots, pp. 343–347 (2018). https://doi.org/10.1007/978-3-319-93846-2_64

Subburaj, S.K., Stewart, A.E.B., Rao, A.R., D'Mello, S.K.: Multimodal, multiparty modeling of collaborative problem solving performance. In: Proceedings of the 2020 Conference on Multimodal Interaction (2020)

Sun, C., Shute, V.J., Stewart, A.E.B., Yonehiro, J., Duran, N., D'Mello, S.K., D'Mello, S., D'Mello, S.K.: Towards a generalized competency model of collaborative problem solving. Comput. Educ. **143**, 103672 (2020). https://doi.org/10.1016/j.compedu.2019.103672

Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010). https://doi.org/10.1177/0261927X09351676

Tegos, S., Demetriadis, S., Karakostas, A.: Promoting academically productive talk with conversational agent interventions in collaborative learning settings. Comput. Educ. **87**, 309–325 (2015). https://doi.org/10.1016/j.compedu.2015.07.014

Tegos, S., Demetriadis, S., Papadopoulos, P.M., Weinberger, A.: Conversational agents for academically productive talk: a comparison of directed and undirected agent interventions. Int. J. Comput. Support. Collab. Learn. **11**(4), 417–440 (2016). https://doi.org/10.1007/s11412-016-9246-2

Vrzakova, H., Amon, M.J., Stewart, A., Duran, N.D., D'Mello, S.K.: Focused or stuck together: multimodal patterns reveal triads' performance in collaborative problem solving. In: Proceedings of the Tenth International Conference on Learning Analytics and Knowledge, pp. 295–304 (2020). https://doi.org/10.1145/3375462.3375467

Webb, M., Gibson, D.: Technology enhanced assessment in complex collaborative settings. Educ. Inf. Technol. **20**(4), 675–695 (2015). https://doi.org/10.1007/s10639-015-9413-5

Weintrop, D.: Minding the gap between blocks-based and text-based programming (Abstract Only). In: Proceedings of the 46th ACM Technical Symposium on Computer Science Education, p. 720 (2015). https://doi.org/10.1145/2676723.2693622

Weintrop, D., Wilensky, U.: Bringing blocks-based programming into high school computer science classrooms. In: Annual Meeting of the American Educational Research Association (AERA). Washington DC, USA (2016)

Westlund, J.K., D'Mello, S.K., Olney, A.M.: Motion Tracker: camera-based monitoring of bodily movements using motion silhouettes. PLoS ONE **10**(6), e0130293 (2015). https://doi.org/10.1371/journal.pone.0130293

Yoo, J., Kim, J.: Can online discussion participation predict group project performance? Investigating the roles of linguistic features and participation patterns. Int. J. Artif. Intell. Educ. **24**(1), 8–32 (2014). https://doi.org/10.1007/s40593-013-0010-8

Angela E. B. Stewart received her Ph.D. in Computer Science from the University of Colorado Boulder in 2020. She is currently a Postdoctoral Fellow in the Human–Computer Interaction Institute at Carnegie Mellon University. Stewart has long been fascinated by how technology and education intersect, to make more equitable, inclusive spaces. Her work investigates how to create educational technologies to support learners from diverse backgrounds, cultures, and life experiences.

Zachary Keirn received his Ph.D. in Electrical Engineering from Colorado State University, 1992. He is currently a consulting engineer with Hitachi Vantara Corporation (47 Lining) working on cloud infrastructure, data mining, and data architecture. He has published numerous papers in bioengineering, magnetic and optical recording. He has recently published work on machine learning and cognitive sciences. Keirn has co-authored over 20 patents in magnetic recording and information storage systems. His research interests include biotechnology, neuroscience, and cognitive science.

Sidney K. D'Mello (Ph.D. in Computer Science) is an Associate Professor in the Institute of Cognitive Science and Department of Computer Science at the University of Colorado Boulder. He is interested in the dynamic interplay between cognition and emotion while individuals and groups engage in complex real-world tasks. He applies insights gleaned from this basic research program to develop intelligent technologies that help people achieve to their fullest potential by coordinating what they think and feel with what they know and do. D'Mello has co-edited seven books and published almost 300 journal papers, book chapters, and conference proceedings. His work has been funded by numerous grants and he currently serves(d) as associate editor for Discourse Processes and PloS ONE. D'Mello is the Principal Investigator for the NSF National Institute for Student-Agent Teaming