User Modeling and User-Adapted Interaction
https://doi.org/10.1007/s11257-023-09387-6

# Improving collaborative problem-solving skills via automated feedback and scaffolding: a quasi-experimental study with CPSCoach 2.0

Sidney K. D'Mello[1] · Nicholas Duran[2] · Amanda Michaels[2] · Angela E. B. Stewart[3]

Received: 10 May 2023 / Accepted in revised form: 24 November 2023

## Abstract

We present CPSCoach 2.0, an automated system that provides feedback, instructional scaffolding, and practice to help individuals improve three collaborative problem-solving (CPS) skills drawn from a theoretical CPS framework: construction of shared knowledge, negotiation/coordination, and maintaining team function. CPSCoach 2.0 was developed and tested in the context of computer-mediated collaboration (video conferencing) with an educational game. It automatically analyzes users' speech during a round of collaborative gameplay to provide personalized feedback and to select a target CPS skill for improvement. After multiple cycles of iterative testing and refinement, we tested CPSCoach 2.0 in a user study where 21 dyads ($n = 42$) completed four rounds of feedback and scaffolding embedded within five rounds of game-play in a single session. Using a quasi-experimental matching procedure, we found that the use of CPSCoach 2.0 was associated with improvement in CPS skill development compared to matched controls. Further, users found the automated feedback to be moderately accurate and had positive perceptions of the system, and these impressions were stronger for those who received higher scores overall. Results demonstrate

✉ Sidney K. D'Mello
sidney.dmello@colorado.edu

Nicholas Duran
nduran4@asu.edu

Amanda Michaels
amanda.michaels@asu.edu

Angela E. B. Stewart
angelas@pitt.edu

[1] University of Colorado Boulder, Boulder, USA

[2] Arizona State University, Glendale, USA

[3] University of Pittsburgh, Pittsburgh, USA

Published online: 14 February 2024

the use of automated feedback and instructional scaffolds to support the development of CPS skills.

**Keywords** Collaborative problem-solving · Automated feedback · Educational technology · User study · Computer-supported collaborative work

## 1 Introduction

What does planning the Artemis 1 mission to the moon, fighting wildfires, scoring the winning goal at the world cup final, and coordinating dinner reservations with fussy eaters have in common? They all involve multiple individuals coming together to identify the steps needed to solve a problem by transforming a given state to a goal state, a process called collaborative problem-solving (CPS) (Fiore et al. 2018). CPS is recognized as a critical twenty-first-century skill in the future workplace and workforce, which is increasingly technologically rich, distributed, and diverse. CPS entails both analytical problem-solving skills and social interpersonal skills. For example, teammates must engage in a number of coordination processes, such as establishing common ground (Graesser et al. 2018), maintaining a joint conception of the problem (Roschelle and Teasley 1995; Graesser et al. 2018), sharing ideas (Graesser et al. 2018), forming a plan (Griffin et al. 2012a; Graesser et al. 2018), negotiating among multiple alternatives (Andrews-Todd and Forsyth 2020), and maintaining a positive team dynamic (Sun et al. 2020).

In theory, teams that exhibit these skills should demonstrate *process gain*—i.e., superior performance that exceeds the joint abilities of the individual team members (Laughlin et al. 1975, 2006; Laughlin and Ellis 1986). The reality, however, is quite different. Decades of research on small group teams have indicated that teams consistently fail to live up to expectations (see reviews (Steiner 1972; Hill 1982; Kerr and Tindale 2004)) and experience *process loss* instead of process gain. Possible causes include coordination losses, such as production blocking during collective ideation (Nijstad et al. 2003), overemphasis on shared vs. individual knowledge (the common-knowledge effect (Gigone and Hastie 1993)), or group-think (Janis 1982) when individual members converge to the dominant view. There are also motivation losses, such as social-loafing (Kerr 1983; Karau and Williams 1993), evaluation apprehension (Diehl and Stroebe 1987), and free-rider effects (Kerr and Bruun 1983).

It is perhaps unsurprising that teams fail to achieve their potential, because they do not have the requisite skills. Recognizing the importance of CPS, in 2015, the Programme for International Student Assessment (PISA) conducted an international CPS assessment among 15-year-old students in 52 countries and regions. They found that only 10% of countries scored at the highest level involving solving complicated problems that require overcoming obstacles and resolving conflict. Even more concerning, less than 30% of students demonstrated success on the lowest complexity problems (OECD 2015a), resulting in a conclusion of a "global deficit in acquiring collaboration competencies" (Fiore et al. 2018b).

In a series of influential reports (Fiore et al. 2018; Graesser et al. 2018), the authors argue that one reason for such deficits is a dearth of training on CPS skills. According to employers, college graduates are ill-equipped for collaborating in the twenty-first-century workplace. As a remedy to this problem, education researchers have called for curricula that focus on developing CPS competencies. This call has remained largely unrealized, as there is little dedicated instruction focused on CPS. Instead, students are expected to master CPS competencies by engaging in small-group collaborations and project-focused teams, where the only guidance and evaluation they receive are on domain knowledge and project outcomes. The assumption that these skills will be learned in some indirect way is faulty. Imagine if math and science were taught this way. To learn these skills, it is necessary to have direct focused instruction and feedback on CPS.

Lastly, whereas much of the classic research on CPS has occurred in face-to-face collaboration (Kerr and Tindale 2004), CPS increasingly occurs in remote, computer-mediated settings, as teams are more distributed (Schulze and Krumm 2017). The COVID-19 pandemic has made the shift to remote work more common, with early reports estimating that approximately 50% of pre-pandemic workers are now doing their jobs from home (Brynjolfsson et al. 2020). CPS is already difficult, and communication barriers inherent to remote work make it even harder (Schulze and Krumm 2017). Social signals are crucial to interpersonal communication, and they are muted or non-existent in virtual interactions (Schulze and Krumm 2017; Alterman and Harsch 2017). Poor audio quality, obstructed visual cues, and lagging signals limit the effectiveness of virtual communication. For example, eye gaze can help regulate turn-taking dynamics (Kendon 1967), yet this signal is suppressed in modern video conferencing interfaces (Vrzakova et al. 2020). Deficiencies in social cues impairs team coordination, communication, cohesion, trust, and even performance (Schulze and Krumm 2017; Virtaneva et al. 2021).

Taken together, there is an ever-increasing prevalence and importance of CPS, yet a conspicuous absence of educational pathways to developing CPS proficiency. This creates an urgent need for scalable solutions for CPS training and mastery especially in remote settings. Accordingly, we propose *CPSCoach 2.0*, an intelligent system that provides feedback and instructional scaffolds to help people practice and improve CPS skills. CPSCoach is a major refinement of an earlier prototype system (Stewart et al. 2023) as detailed in Sect. 1.2. We present the design and implementation of CPSCoach 2.0 followed by a study to investigate whether the use of CPSCoach 2.0 is associated with improvement in corresponding skills and to examine user perceptions of the system.

## 1.1 Background and related work

### 1.1.1 Collaborative problem-solving proficiency

CPS occurs when two or more people coordinate to solve a problem (Roschelle and Teasley 1995, 2015). It involves a myriad of skills (Nelson 1999; Hesse et al. 2015; Cukurova et al. 2018). For example, teammates need to create a collaborative environment by establishing common ground (Roschelle and Teasley 1995, 2015), understanding their teammates' perspectives (Griffin et al. 2012a; Hesse et al. 2015), assigning roles according to team member strengths (Griffin et al. 2012a; Hesse et al. 2015), monitoring for breakdowns in communication (Roschelle and Teasley 1995, 2015), and negotiating for compromise (Griffin et al. 2012a; Hesse et al. 2015). Teams must also engage in effective problem-solving processes—they must analyze the problem and its constraints (Griffin et al. 2012a; Hesse et al. 2015; Webb and Gibson 2015; Graesser et al. 2018), strategize for how to achieve the goal (Griffin et al. 2012a; Hesse et al. 2015), develop and execute solution plans (Griffin et al. 2012a; Hesse et al. 2015), reflect on results (Webb and Gibson 2015; Graesser et al. 2018), and iteratively refine plans (Webb and Gibson 2015; Graesser et al. 2018). Thus, both task- and social-focused norms are important, and people need to master a number of cognitive, social, and metacognitive skills to successfully engage in CPS. Researchers have organized collections of these skills into various frameworks, such as the PISA CPS Framework (OECD 2015b), the ATC21S framework (Friedrich et al. 2015), the CPS Ontology (Andrews-Todd and Forsyth 2020), and the Generalized CPS Framework (Sun et al. 2022). These frameworks provide a theoretical basis for the assessment and development of CPS skills.

### 1.1.2 Assessing and modeling collaborative problem-solving

There is a vast literature on modeling collaborations. One line of work examines low-level behaviors, such as visual focus of attention (Otsuka et al. 2018), turn-taking (de Kok and Heylen 2009; Dielmann et al. 2010; Jokinen et al. 2013), and behavioral coordination and synchrony (Fusaroli et al. 2014; Krafft et al. 2016; Stewart et al. 2018; Eloy et al. 2019; Amon et al. 2019). At a higher level, researchers have also identified important socio-cognitive processes, such as rapport-building (Sinha and Cassell 2015; Müller et al. 2018), speaker dominance and influence (Aran and Gatica-Perez 2010; Nihei et al. 2014), and emergent leadership (Sanchez-Cortes et al. 2010; Mercier et al. 2014; Beyan et al. 2016). Some have examined modeling of CPS outcomes, such as task performance (Murray and Oertel 2018; Chopade et al. 2019; Subburaj et al. 2020) and learning educational content after engaging in CPS (Stewart and D'Mello 2018; Olsen et al. 2020).

Most relevant to our work is research on assessment and computational modeling of CPS skills. One approach is to leverage the above CPS theoretical frameworks to create assessments of CPS. For example, in the large-scale PISA CPS assessment, students interacted with a computer agent by choosing responses from pre-defined options designed to elicit CPS skills based on the PISA CPS framework (OECD 2015b). Although this approach allows for precision of CPS skill assessment at scale, its ecological validity is limited in that communications are constrained to pre-specified responses. Others addressed this limitation by using log data from a CPS virtual environment to infer a mapping between student actions in the environment and levels of CPS proficiency (low and high) (Scoular and Care 2020). They found that the identified behaviors were consistent across students, tasks, and assessments. However, the click stream type of the interaction does not resemble rich, naturalistic social interactions.

As a step toward supporting more naturalistic communications, several researchers have focused on modeling CPS based on chat communications during collaborative tasks (Michael et al. 2016; Stephen et al. 2017; Dowell et al. 2019). For example, researchers have used network analysis techniques to understand how individual contributions are situated in interactive, interdependent, temporal discourse (Swiecki et al. 2020), latent semantic analysis coupled with clustering to understand how emergent roles are related to CPS outcomes and skills (Dowell et al. 2020), and n-grams to predict CPS behaviors from chat communications (Jiangang et al. 2017). However, CPS assessment from open-ended spoken discourse still remains elusive with few exceptions (Stewart et al. 2019; Pugh et al. 2022), ostensibly due to challenges with automatic speech recognition especially in real-world contexts (Pugh et al. 2021; Southwell et al. 2022). Lastly, there have been some efforts toward multimodal modeling of CPS, including the use of video, acoustics, and features of the task (Cukurova et al. 2020; Stewart et al. 2021), but the verbal modality tends to outperform nonverbal modalities in predicting CPS skills given the emphasis of spoken communication in remote CPS (Stewart et al. 2021). Nonverbal signals might be more useful for other outcomes of interest like success at the task or users' subjective perceptions of the interaction (Vrzakova et al. 2019; Shree Krishna et al. 2020).

### 1.1.3 Intelligent systems to support collaboration

Researchers in the fields of computer-supported collaborative work (CSCW) and computer-supported collaborative learning (CSCL) have developed and evaluated several systems that support collaborations. Many are task-specific, in that they augment team processes to more effectively move them toward the end-goal. For example, some interfaces aim to enhance team communication and code sharing in programming tasks (Davor et al. 2005) or support collaborative information seeking by improving group coordination (Hong et al. 2019). Systems have also been built to monitor the unfolding collaboration and automate tasks to increase productivity, such as automated discussion summarization (Gutwin et al. 2017; Tian et al. 2021) and task list generation (McGregor and Tang 2017).

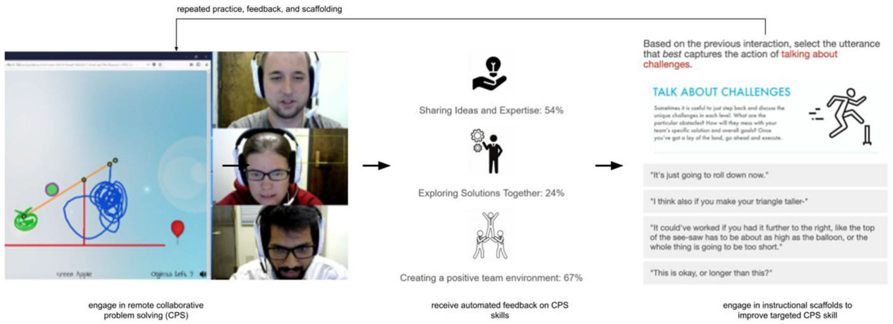
Most closely related to our work are systems that emphasize awareness of individual behaviors in collaborative interactions for reflection (Kori et al. 2014). In this literature, machine-observable behaviors are usually synthesized into meaningful metrics and displayed to the user. For example, collaborator speech has been used to compute metrics of turn-taking behaviors (e.g., verbal participation, interruptions) (Calacci et al. 2016; Faucett et al. 2017), which are then visualized in real time. As another example, visualizations of shared eye gaze among teammates have been used to support shared awareness (Schlösser et al. 2018; Kütt et al. 2020), in turn increasing task performance and decreasing cognitive workload (Kütt et al. 2020). In the context of remote meetings, MeetingCoach automatically analyzes users' audiovisual data, extracts measures such as tone, turn taking, sentiment., and displays these to the user along with suggestions (e.g., "try varying your pitch") via an interactive dashboard (Samrose et al. 2021). A user study indicated that participants reported that MeetingCoach helped improve their awareness of meeting dynamics. Together, these works demonstrate how visualization of behaviors can potentially prompt teammates to engage in more effective collaborative behaviors. However, to our knowledge, researchers have yet to directly address improvement of complex CPS skills.

## 1.2 Novelty, contributions, and research questions

We identified three gaps from our review of the literature. First, researchers have developed theoretical CPS frameworks for the purpose of assessing this critical twenty-first-century skill. However, complementary frameworks and guidance to help *improve* CPS skills are largely missing despite a critical need (Fiore et al. 2018). Second, there is a considerable body of research on modeling CPS and related collaborative behaviors, but these models have yet to be integrated within intervention systems that support the development of CPS skills. Further, computational models of CPS have focused on interactions via chats and interface logs (e.g., (Stephen et al. 2017; Stoeffler et al. 2018)), but with some exceptions (e.g., (Stewart et al. 2019; Pugh et al. 2021, 2022)) have yet to broadly embrace the complexity of open-ended spoken collaborative discourse. Third, whereas there has been considerable work in the CSCW and CSCL communities on improving collaboration outcomes, the focus has been on supporting specific tasks or providing feedback on communicative behaviors (e.g., turn taking, prosody, sentiment). There have yet to be systems that support the development of complex CPS skills.

We addressed these gaps by designing and testing CPSCoach 2.0, an intelligent, personalized system that aims to help users improve their CPS proficiency in remote collaborations via video conferencing. CPSCoach 2.0 has two components: (1) a *feedback* system that uses speech and language processing with deep machine learning to automatically measure high-level CPS skills based during remote CPS and (2) and an *intervention* system that provides instructional scaffolds to help users' improve targeted CPS skills *personalized* to individual users (Fig. 1). The CPS assessments and interventions are grounded in a validated CPS theoretical framework (e.g., (OECD 2015b; Andrews-Todd and Forsyth 2020)) and evidence-based principles from the cognitive science of learning (Bransford et al. 2000; NASEM 2018). Both components are implemented in a collaborative *environment* where people engage in CPS to demonstrate and practice their skills.

The present study makes two contributions. First, we discuss the design and implementation of CPSCoach 2.0, highlighting general design considerations that can be


repeated practice, feedback, and scaffolding
Sharing Ideas and Expertise: 54%
Exploring Solutions Together: 24%
Creating a positive team environment: 67%
TALK ABOUT CHALLENGES
engage in remote collaborative problem solving (CPS)
receive automated feedback on CPS skills
engage in instructional scaffolds to improve targeted CPS skill

Based on the previous interaction, select the utterance that *best* captures the action of talking about challenges.

Sometimes it is useful to just step back and discuss the unique challenges in each level. What are the particular obstacles? How will they mess with your team's specific solution and overall goals? Once you've got a lay of the land, go ahead and execute.

"It's just going to roll down now."

"I think also if you make your triangle taller-"

"It could've worked if you had it further to the right, like the top of the see-saw has to be about as high as the balloon, or the whole thing is going to be too short."

"This is okay, or longer than this?"

**Fig. 1** High-level overview of the CPSCoach 2.0 feedback (middle) & intervention system (right) situated within the collaborative environment (left)

applicable to similar intelligent systems aimed at helping users develop CPS skills. Second, we present a quasi-experimental study that investigates two research questions (RQ1): Are interactions with CPSCoach 2.0 associated with improvements in CPS skills?; and (RQ2): What are users subjective perceptions of CPSCoach 2.0 and what factors predict these perceptions?

CPSCoach 2.0 builds off an earlier prototype system called CPSCoach (Stewart et al. 2023), which also provided automated feedback on CPS skills using the same machine learning models (i.e., item #1 was the same). However, CPSCoach 2.0 includes a new and elaborated intervention system (item #2) than CPSCoach, which has three salient differences. First, CPSCoach 2.0 uses automated feedback to personalize the interventions to individual users by selecting one target CPS skill to focus on, whereas the original interventions targeted all three skills. Second, the CPSCoach 2.0 interventions greatly differ from those used previously in terms of depth, level of interactivity, and length. Specifically, whereas the former interventions were passive (interaction was limited to reading and viewing videos), the CPSCoach 2.0 interventions incorporate constructive and generative activities with feedback and additional opportunities for practice. Third, the CPSCoach 2.0 interventions are dynamic, changing each time, whereas the former interventions were static, resulting in considerable disengagement after the novelty wore off. Further, several insights learned from a previous usability study with the initial system were incorporated into CPSCoach 2.0. Lastly, the previous user study[1] (Stewart et al. 2023) did not investigate whether CPSCoach improved CPS skills (RQ1), which we do here for the first time with CPSCoach 2.0. With the revised intervention design, research addressing RQ2 (user perceptions) is also novel.

## 2 Feedback and intervention design of CPSCoach 2.0

The present CPS feedback and intervention system is intended for any number of contexts; however, to ground its operationalization we instantiate it within a particular collaborative problem-solving environment: an engaging two-dimensional educational game called Physics Playground.

### 2.1 Physics playground: current CPS environment

Physics Playground is designed to introduce basic Newtonian physics concepts (e.g., Newton's laws, energy transfer, and properties of torque) to students (Miguel et al. 2014; Bosch et al. 2015; Shute et al. 2021a, b). The game is composed of multiple game levels organized as "playgrounds" that players can freely navigate. The objective of the game is to draw simple machines (i.e., ramps, levers, pendulums, and springboards) within a level to navigate a green ball to a red balloon while avoiding pre-existing obstacles. Everything in the game obeys the laws of Physics. Players can restart, exit, or change levels at any time, as well as view a tutorial on game mechanics. A team earns

[1] The only publication on CPSCoach includes a short conference paper (Stewart et al. 2023). The present paper reports results on a new study with a substantially revised system.

a gold trophy when they successfully solve a level using fewer objects. A successful solution that uses more objects earns a silver trophy.

The game can be played both individually or collaboratively. In the latter case, one player—designated the Controller—can control interaction within the game. This person's screen is shared with the other players – called Contributors—who discuss and contribute to the solution. Collaborative gameplay can occur face-to-face or via video conferencing. Figure 1 (left) shows a team using a lever (pre-existing in the game) and a weight (drawn by team) to solve a level. This team was virtually collaborating over zoom with video and audio enabled.

## 2.2 Automated feedback generation

The first component of our system automatically assesses CPS skills. We used a validated theoretical framework of CPS to identify focal skills, followed by a deep text classification approach to computationally model these skills.

### 2.2.1 CPS framework

The framework defines three facets (i.e., high-level skills) that comprise CPS: (1) construction of shared knowledge (or shared knowledge construction) occurs when teammates share ideas and expertise with each other to bring about a collective, broad understanding of the problem space; (2) negotiation/coordination is a process of iteratively developing and executing a particular solution, and revising it as necessary; (3) maintaining team function occurs when teammates create a positive group dynamic by eliciting each other's perspectives, providing encouragement, and proactively contributing to the team's success. Each facet is comprised of three positive verbal indicators, which are observable behaviors used to assess the corresponding facet (Table 1). The framework also includes some negative indicators (e.g., making fun of others); however, we only focused on the positive indicators here.

The framework was validated in two studies using two datasets: one of middle school students playing an educational game face-to-face, and another with college students engaging in a visual programming task over a videoconference (Sun et al. 2020, 2022). Together the three facets work in unison to facilitate effective collaboration, and indicators from all three facets have been linked to successful CPS outcomes (Sun et al. 2022; Zhou et al. 2022). Whereas nonverbal indicators do play a role in CPS, the verbal indicators tend to dominate in remote collaborations (Stewart et al. 2021), making the framework very relevant for the present context.

### 2.2.2 Data to train computational models

We used a training dataset where 94 triads collaboratively played Physics Playground for three 15-min blocks in one of two research labs. Teammates were positioned at separate computer workstations, and communicated over videoconference (Zoom). Control of Physics Playground switched each block, so that each person in the triad controlled Physics Playground for one block. We recorded separate audio streams for

**Table 1** CPS framework with facets, positive verbal indicators, and examples from real teams playing Physics Playground

| Facet or skill | Indicator | Human transcribed example |
|---|---|---|
| Construction of Shared Knowledge | Proposing Specific Solutions | *Also draw the two pivots a little higher up on the dolphin* |
| | Talking about the givens and constraints of the task | *So if you want to move yeah so you can just click on the ball, and that will be making the ball move to the right* |
| | Confirming understanding by asking questions/paraphrasing | *So what should we attach it to then?* |
| Negotiation/ Coordination | Providing reasons to support a solution | I guess while you're at it, do that on the other end too, so that it doesn't like roll off. |
| | Responding to others' questions/ideas | *Right. Yeah, right here* |
| | Talking about results | *There's not enough momentum for it* |
| Maintaining Team Function | Asking if others have suggestions | *Like utilizing what is already on the screen, how would we be able to attach something to stop that ball?* |
| | Complimenting or encouraging others | *Great idea* |
| | Giving instructions | *Yeah and then just like do that, and make a little hook at the end* |

each teammate, which were automatically transcribed using the IBM Watson automated transcription service. In total, there were 87,943 utterances across 94 triads.

Three expert humans were trained to code the utterances for the presence of each verbal indicator. Coders watched videos of the collaborations, alongside the automated transcripts and counted the number of times each indicator occurred in an utterance, but because the same indicator rarely (< 1%) occurred twice per utterance, we binarized indicator counts. Coder agreement on the indicators ranged from 0.88 to 1.00 (Gwet's AC1) on 10, 90-s video samples consisting of 406 utterances. After training, videos were randomly assigned for independent coding. We adopted a thin slicing approach (Olsen and Finkelstein 2017) where a random 90 s was coded from the first, second, and third five minutes of a 15-min block (i.e., 30% of the data was coded, or 90 s × 3). In all, 27,019 utterances were coded across all teams and blocks. A total of 22% of the utterances were coded for construction of shared knowledge, 13% for negotiation/coordination, and 9% for maintaining team function.

### 2.2.3 Training and validating computational models

We used the above dataset to develop computational models to predict binary codes for the three CPS facets from the automated transcriptions. We specifically use one modern deep learning architecture called transformers (Vaswani et al. 2017) in a transfer (machine) learning setting where a model trained on one dataset/task is adapted to another (Pan and Yang 2010). This entails two steps: pretraining and fine-tuning. During pretraining, the transformer uses large amounts (i.e., gigabytes) of text to learn the meanings of words from their context via specific tasks. For example, the language modeling task entails predicting the next word from the previous words (Brown et al., 2020), whereas masked language modeling involves predicting a word from both its left and its right sequence of words (Devlin et al. 2018). Training on these tasks enables the model to obtain a representation of the contextual meaning of words, which serves as a starting point for subsequent fine-tuning. Here, the model is augmented to include a task-specific (CPS classification in our case) output layer and then tuned (parameters are updated) on small amounts of task-specific data (the 27,019 annotated utterances).

We used the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al. 2018), which was state of the art when the models were initially developed and validated (i.e., in 2018–2019). We started with the 'bert-base-uncased' pre-trained model weights from huggingface.co and then fine-tuned it on our labeled dataset for four epochs, using hyper-parameters based on recommendations from (Wolf et al. 2019) (e.g., fine-tuning over four epochs, batch size = 32, sequences padded or truncated at 300 words). The BERT model outputs a probability that an utterance is a positive example of each facet.

We used team-level tenfold cross-validation such that all utterances from a team were in the training set, or testing set, but never both; this is critical for generalizability to new teams. The cross-validation process involved training a model on data from 90% of teams, then evaluating its predictive accuracy on a test set (which contains data from the 10% of teams withheld during training). This was repeated ten times, such that every team appeared in the test set exactly once. Finally, test set predictions from each of the ten folds were aggregated to compute accuracy using the receiver operating characteristic curve (AUROC). This metric considers the true-positive/false-positive tradeoff across various classification thresholds, rather than forcing a single binary prediction. We achieved cross-validated AUROC values of 0.88, 0.83, and 0.82 for construction of shared knowledge, negotiation/coordination, and maintaining team function, respectively (chance = 0.5). This metric indicates good classification accuracy and generalizability to new teams in a similar context.

This general modeling approach has been shown to generalize to a different CPS domain (Pugh et al. 2022). It is also moderately robust to speech recognition errors (Southwell et al. 2022), for example achieving AUROCs of 0.80 using noisy speech recognition data compared to 0.91 for human transcripts (Pugh et al. 2021).

### 2.2.4 Model deployment: generating feedback scores in near-real time

Once cross-validated accuracy was established (see above), we retrained (i.e., fine-tuned) a model on all the data (i.e., 27,019 instances preserving the hyper-parameters noted above) and utilized this deployment model for near-real-time assessment in the current study. The deployment model operates at the utterance level (i.e., predicts the probability of three CPS facets for each individual utterance). However, our goal is to provide feedback at the level of a block or round of gameplay comprising multiple utterances. Therefore, we generated an aggregate score for each facet by averaging the utterance-level probabilities in the block/round. To address outliers, we winsorized the top and bottom 1.25% of values (i.e., replacing values outside of these bounds with the closest value within each bound). Because there were differences in scores across the three facets, we norm-referenced the aggregated round-level proportions into percentile scores (i.e., using the cumulative distribution function for the corresponding facet) from the reference (i.e., training) dataset. Further, since there were differences in facet scores between Controllers and Contributors, we computed the percentile score separately for two roles, resulting in six distributions (3 facets × 2 roles). This ensured equivalence of scores relative to differences across facets and roles (e.g., different raw *mean* scores for all three facets and roles would yield the *same* percentile score of 50%).

Participants completed the study from their own personal spaces using their own devices (due to pandemic restrictions Sect. 3), which limited researcher control on software and hardware setup. Therefore, processing of audio to generate feedback scores was semi-automated in that each participant's audio stream was manually recorded using Zoom during the collaborative CPS interactions and submitted to the automated feedback generating pipeline. Because there were latencies in this pipeline, only the first seven minutes of audio from ten-minute blocks of gameplay was recorded and submitted to the processing pipeline. This way, scores were ready for feedback and intervention soon after the block had ended.

## 2.3 Intervention design

### 2.3.1 Design principles

The design of the feedback and intervention systems was guided by a number of principles from the cognitive science of learning (Bransford et al. 2000; NASEM 2018). Briefly, these include:

**Objective feedback.** This principle highlights the importance of providing objective feedback as a necessary condition for learning (Azevedo and Bernard 1995; Bransford et al. 2000). Feedback plays a foundational role in the intervention by providing the starting point to launch subsequent interventions aimed at improving CPS skills.

**Formative feedback**. This principle highlights the virtue of feedback that is formative or aimed towards improvement rather than feedback that is evaluative (Shute 2008). Accordingly, the feedback provided is framed as an opportunity to improve, as opposed to a means to demonstrate competence (Elliot and McGregor 2001). For this reason, feedback is always individualized and provided to users privately rather than collectively.

**Observational learning.** Theories of observational or social learning posit that people learn vicariously through observing examples of behaviors in others (Chi et al. 2001). Accordingly, the intervention provides users with numerous examples of CPS from real teams with prompts to promote reflection, comparison, and assessment.

**ICAP (Interactive-constructive-active–passive) framework**. According to this theory (Chi and Wylie 2014), learning activities that target interactive (e.g., explaining to others) and constructive (e.g., generating a response) processes are more beneficial than active (e.g., taking verbatim notes) and passive (e.g., silently reading) ones. Thus, the intervention is designed to engender constructive and interactive processes via information retrieval, error correction, comparing and contrasting, and generating open-ended responses.

**Instructional scaffolding.** Lastly, the intervention is designed to scaffold students toward mastery in incremental steps (Smagorinsky 2018). This scaffolding is implemented by initially providing direct instruction and examples on individual behaviors that comprise the CPS skills, followed by increasing mastery to consider multiple behaviors and skills, followed by asking users to construct explanations rather than providing corrective feedback and opportunities for revised responses.

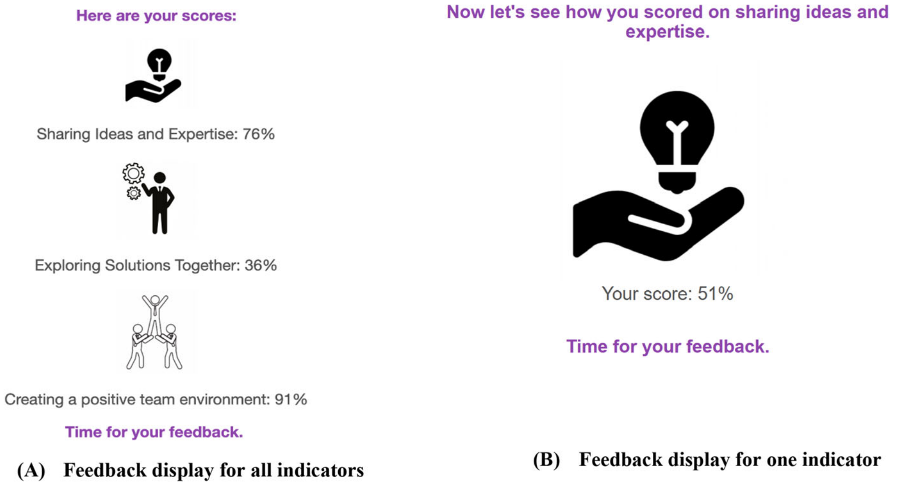
### 2.3.2 Reconfigurable and reusable intervention components

The intervention content was designed as a set of reusable components that could be assembled to create a variety of learning experiences aligned with the above principles. It is intended to be interspersed within subsequent rounds of gameplay where users' reflect on the automated feedback and learn strategies to improve their CPS skills.

**Feedback Display.** This component simply displays users' CPS performance (based on prior interactions) in an easy-to-understand fashion. It displays scores as percentages (see Sect. 2.2.4) along with accompanying media icons for each facet. The three facets were renamed in order to facilitate easier interpretation of the scores as follows: Constructing Shared Knowledge was renamed "Sharing Ideas and Expertise"; Negotiation/Coordination was called "Exploring Solutions Together"; and Maintaining Team Function was renamed "Creating a Positive Team Environment." This component affords display of all three scores (Fig. 2A) or a score for a selected facet (Fig. 2B).

**Facet and Indicator Overview.** The purpose of this component is to introduce participants to the facet selected for feedback (called target facet). Accordingly, it identifies the selected facet (Fig. 3A), provides a short discussion on the importance of the facet for supporting effective CPS (Fig. 3B), followed by descriptions of the three indicators that comprise the facet along with actionable strategies to help users learn how to utilize each indicator (Fig. 3C). Information pertaining to the facet and its indicators was designed to be presented as an initial exposure on a single page (Fig. 3C) or as a reminder with one indicator per page (not shown here) (Fig. 4).

Indicator Identification (Multiple-choice response). This component provides users with an opportunity to identify specific CPS indicators from example CPS sessions. Users are presented with a short video (30 s on average) of a team collaboratively solving a Physics Playground level (see Fig. 1—left), followed by a transcript of the spoken discourse, and a high-level description of the activity (i.e., to identify the



Here are your scores:

Sharing Ideas and Expertise: 76%

Exploring Solutions Together: 36%

Creating a positive team environment: 91%

Time for your feedback.

**(A) Feedback display for all indicators**

Now let's see how you scored on sharing ideas and expertise.

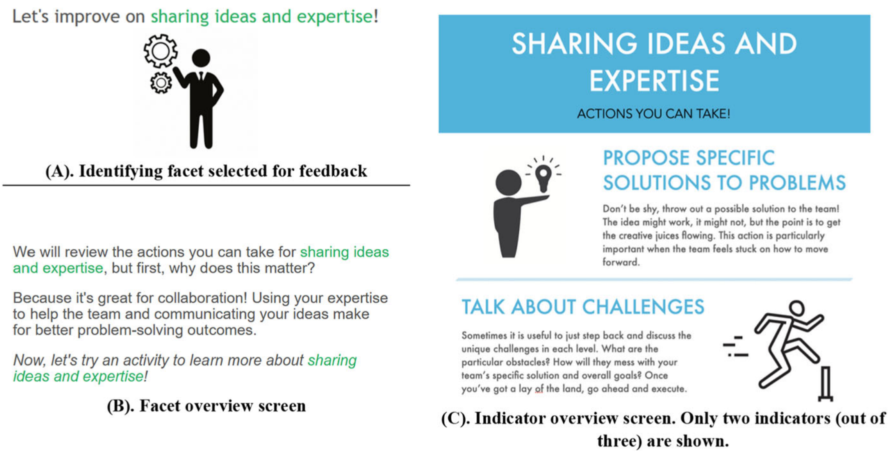
Your score: 51%

Time for your feedback.

**(B) Feedback display for one indicator**

**Fig. 2** Screenshot of the Feedback Display screen

three indicators of the selected facet from the transcript) (Fig. 4A), which consists of selecting an utterance (out of four utterances) from the transcript that best reflects a particular indicator (Fig. 4B). They are given an opportunity to rewatch the video prior to beginning the activity. If their response is correct, they receive positive feedback and an explanation of the correct answer (Fig. 4C), else they receive negative feedback with a hint (Fig. 4D). They are given a second opportunity to respond. If their second response is correct, they are provided the correct feedback screen (Fig. 4C), but if it is incorrect, they are again provided with the incorrect feedback screen, followed by an explanation of the correct response (Fig. 4C). This process repeats for all three indicators.



Let's improve on sharing ideas and expertise!

**(A). Identifying facet selected for feedback**

We will review the actions you can take for sharing ideas and expertise, but first, why does this matter?

Because it's great for collaboration! Using your expertise to help the team and communicating your ideas make for better problem-solving outcomes.

*Now, let's try an activity to learn more about sharing ideas and expertise!*

**(B). Facet overview screen**

SHARING IDEAS AND EXPERTISE

ACTIONS YOU CAN TAKE!

PROPOSE SPECIFIC SOLUTIONS TO PROBLEMS

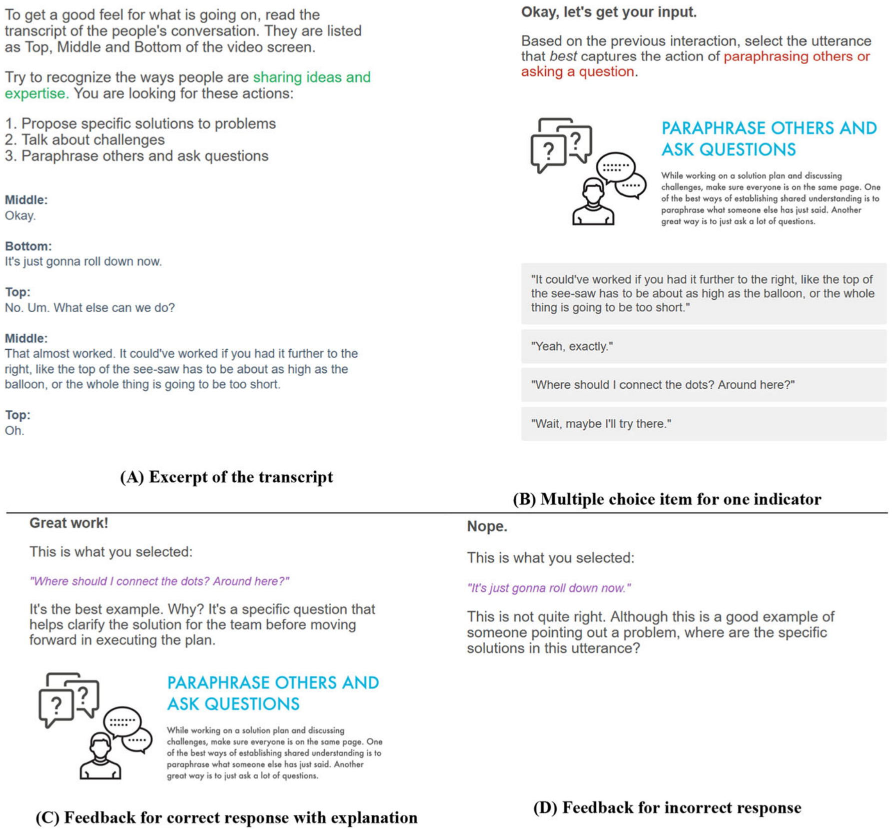
Don't be shy, throw out a possible solution to the team! The idea might work, it might not, but the point is to get the creative juices flowing. This action is particularly important when the team feels stuck on how to move forward.

TALK ABOUT CHALLENGES

Sometimes it is useful to just step back and discuss the unique challenges in each level. What are the particular obstacles? How will they mess with your team's specific solution and overall goals? Once you've got a lay of the land, go ahead and execute.

**(C). Indicator overview screen. Only two indicators (out of three) are shown.**

**Fig. 3** Screenshots of the Facet and Indicator Overview screens



To get a good feel for what is going on, read the transcript of the people's conversation. They are listed as Top, Middle and Bottom of the video screen.

Try to recognize the ways people are sharing ideas and expertise. You are looking for these actions:

1. Propose specific solutions to problems
2. Talk about challenges
3. Paraphrase others and ask questions

**Middle:**
Okay.

**Bottom:**
It's just gonna roll down now.

**Top:**
No. Um. What else can we do?

**Middle:**
That almost worked. It could've worked if you had it further to the right, like the top of the see-saw has to be about as high as the balloon, or the whole thing is going to be too short.

**Top:**
Oh.

**(A) Excerpt of the transcript**

**Okay, let's get your input.**

Based on the previous interaction, select the utterance that *best* captures the action of paraphrasing others or asking a question.

PARAPHRASE OTHERS AND ASK QUESTIONS

While working on a solution plan and discussing challenges, make sure everyone is on the same page. One of the best ways of establishing shared understanding is to paraphrase what someone else has just said. Another great way is to just ask a lot of questions.

"It could've worked if you had it further to the right, like the top of the see-saw has to be about as high as the balloon, or the whole thing is going to be too short."

"Yeah, exactly."

"Where should I connect the dots? Around here?"

"Wait, maybe I'll try there."

**(B) Multiple choice item for one indicator**

**Great work!**

This is what you selected:

*"Where should I connect the dots? Around here?"*

It's the best example. Why? It's a specific question that helps clarify the solution for the team before moving forward in executing the plan.

PARAPHRASE OTHERS AND ASK QUESTIONS

While working on a solution plan and discussing challenges, make sure everyone is on the same page. One of the best ways of establishing shared understanding is to paraphrase what someone else has just said. Another great way is to just ask a lot of questions.

**(C) Feedback for correct response with explanation**

**Nope.**

This is what you selected:

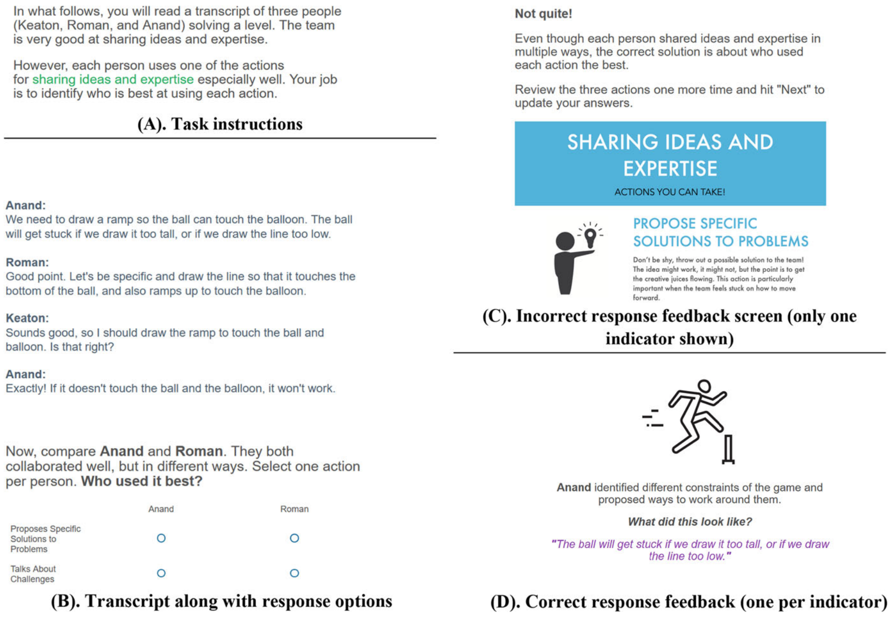
*"It's just gonna roll down now."*

This is not quite right. Although this is a good example of someone pointing out a problem, where are the specific solutions in this utterance?

**(D) Feedback for incorrect response**

**Fig. 4** Screenshots of the Indicator Identification task

**Compare and Contrast.** This activity aims at providing users with an opportunity to compare and contrast effective use of CPS indicators across different hypothetical individuals. Users are informed that they would read a transcript of three people solving a level in Physics Playground who were generally good in demonstrating the selected facet. They are further informed that their task is to identify the person who was particularly adept at demonstrating behaviors (indicators) that map onto the facet (Fig. 5A). Next, they are provided a synthesized transcript clearly demarking the turns associated with each hypothetical individual (e.g., Anand, Keaton, and Roman). Below the transcript, a grid-format matching question (Fig. 5B) asks users to identify which of two hypothetical individuals (e.g., Anand and Roman) more aptly demonstrated each of two indicators of the selected facts; users are not able to select the same individual for both indicators. For incorrect responses, they receive negative feedback, an opportunity to review the facet and indicator overview screen (Fig. 5C), and then correct their response. A second incorrect response yields another feedback screen, but this time they are informed that they will be moved on to another item and that the correct answer will be displayed at the end. The second item asks them to contrast



In what follows, you will read a transcript of three people (Keaton, Roman, and Anand) solving a level. The team is very good at sharing ideas and expertise.

However, each person uses one of the actions for sharing ideas and expertise especially well. Your job is to identify who is best at using each action.

**(A). Task instructions**

**Anand:**
We need to draw a ramp so the ball can touch the balloon. The ball will get stuck if we draw it too tall, or if we draw the line too low.

**Roman:**
Good point. Let's be specific and draw the line so that it touches the bottom of the ball, and also ramps up to touch the balloon.

**Keaton:**
Sounds good, so I should draw the ramp to touch the ball and balloon. Is that right?

**Anand:**
Exactly! If it doesn't touch the ball and the balloon, it won't work.

Now, compare **Anand** and **Roman**. They both collaborated well, but in different ways. Select one action per person. **Who used it best?**

| | Anand | Roman |
|---|---|---|
| Proposes Specific Solutions to Problems | ○ | ○ |
| Talks About Challenges | ○ | ○ |

**(B). Transcript along with response options**

**Not quite!**

Even though each person shared ideas and expertise in multiple ways, the correct solution is about who used each action the best.

Review the three actions one more time and hit "Next" to update your answers.

SHARING IDEAS AND EXPERTISE

ACTIONS YOU CAN TAKE!

PROPOSE SPECIFIC SOLUTIONS TO PROBLEMS

Don't be shy, throw out a possible solution to the team! The idea might work, it might not, but the point is to get the creative juices flowing. This action is particularly important when the team feels stuck on how to move forward.

**(C). Incorrect response feedback screen (only one indicator shown)**

**Anand** identified different constraints of the game and proposed ways to work around them.

***What did this look like?***

*"The ball will get stuck if we draw it too tall, or if we draw the line too low."*

**(D). Correct response feedback (one per indicator)**

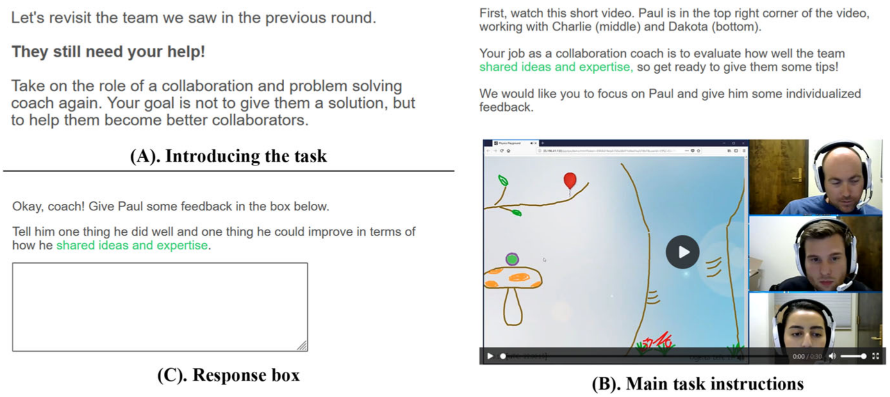
**Fig. 5** Screenshots of the Compare & Contrast task

with another pair (e.g., Anand and Keaton) on two more indicators (one overlapping with the previous item). The feedback, revision, feedback process is repeated for this second item. Lastly, the correct answers are provided by displaying each individual, the indicator for which they performed best, and a corresponding example from the transcript (Fig. 5D).

**Collaboration Coaching.** The purpose of this activity is for users to integrate the knowledge they have gained from the previous informational and scaffolding activities. Users are asked to watch a video clip (40 s on average) of a real CPS team collaborating in a Physics Playground level. They are prompted to adopt the role of a "collaboration coach" (Fig. 6A) who is tasked with providing targeted CPS feedback for a designated individual in the video clip (Fig. 6B). Users can review the video multiple times. The response prompt asks users to provide feedback on one thing they did well and one thing they could improve on. The open-ended feedback is targeted at the facet level (rather than indicator level) with the aim of promoting an integrated response of the three indicators underlying the facet (Fig. 6C). Responses less than 10 characters in length receive an error message and an opportunity to retry. Otherwise, no other feedback is provided.

### 2.3.3 Sequencing of intervention components

We assembled the intervention components to create two levels of scaffolds for each facet with a round of collaborative game-play in-between each scaffold (Fig. 7). Scaffold 1 (Initial exposure) was designed to provide an initial exposure to the facet and



Let's revisit the team we saw in the previous round.

**They still need your help!**

Take on the role of a collaboration and problem solving coach again. Your goal is not to give them a solution, but to help them become better collaborators.

**(A). Introducing the task**

First, watch this short video. Paul is in the top right corner of the video, working with Charlie (middle) and Dakota (bottom).

Your job as a collaboration coach is to evaluate how well the team shared ideas and expertise, so get ready to give them some tips!

We would like you to focus on Paul and give him some individualized feedback.

**(B). Main task instructions**

Okay, coach! Give Paul some feedback in the box below.

Tell him one thing he did well and one thing he could improve in terms of how he shared ideas and expertise.

**(C). Response box**

**Fig. 6** Screenshots of the Collaboration Coaching task

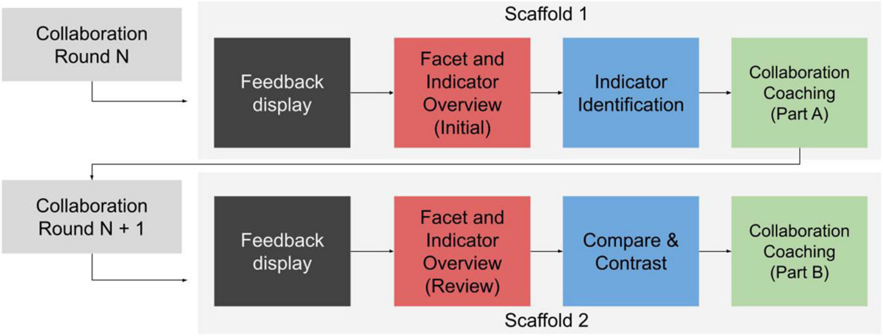

Collaboration Round N
Scaffold 1
Feedback display
Facet and Indicator Overview (Initial)
Indicator Identification
Collaboration Coaching (Part A)
Collaboration Round N + 1
Feedback display
Facet and Indicator Overview (Review)
Compare & Contrast
Collaboration Coaching (Part B)
Scaffold 2

**Fig. 7** Sequence of components for intervention

its behavioral indicators. It consisted of the Facet and Indicator Overview activity to provide an introduction, the Indicator Identification task as an easy step-by-step application, followed by the more challenging Collaboration Coaching task as an initial attempt to produce an integrated and constructive response at the facet level. Scaffold 2 (Complex practice) was designed to provide a second round of instruction and more complex practice after users had an opportunity to apply learnings from Scaffold 1. It included an abbreviated Indicator Identification task as a refresher of the content (not shown here), followed by the Compare and Contrast task which involved multiple indicators, again culminating with the Collaboration Coaching task (using the same video but this time focusing on a different [from Scaffold 1] individual). Both scaffolds were proceeded by versions of the Feedback Display component (see Methods below). At the end of Scaffolds 1 and 2, users are informed to focus on the targeted feedback in subsequent gameplay and as a means to become a better collaborator.

The feedback was administered via Qualtrics surveys (6 total – 3 facets × 2 scaffolds), formatted as a series of guided, step-by-step web pages. There were filler pages between the various tasks and component pages to make the interaction more seamless and user-friendly. Back arrows were available to participants whenever a learning concept was present (text, transcripts, video examples) but were disabled for pages where responses were solicited to avoid repeated attempts.

Overall, the intervention content including instructional materials, assessments, and feedback was designed to be formative (i.e., focusing on improvement) rather than evaluative. Thus, while the complexity of the intervention increased across scaffolds, it was designed such that most users could successfully answer all items and complete each scaffold within a 5- to 10-min interval.

### 2.3.4 Iterative refinement

The intervention was developed over multiple iterations where a version was internally assessed then tested with dyads who were observed and interviewed about their experience. The intervention materials and assessments were then modified based on their feedback. Piloting proceeded over 15 rounds with initial rounds focusing on subcomponents of the intervention and later rounds testing the entire system. Overall, feedback from approximately 20 users was incorporated into the current version of the intervention.

## 3 Method

The study was conducted in Spring 2021 in the midst of the pandemic. As a result, all data collection occurred virtually from participants' homes via video conferencing and the experimenters also participated virtually from their homes. All study procedures were approved by the Institutional Review Board at Anonymous University.

### 3.1 Participants

Participants were 42 individuals (21 dyads) recruited from a large public university in the United States (via flyers, listservs, and online postings). Self-reported demographics were 55% female, 45% male, 0% non-binary, of which 52% were White, 33% Asian, 2% Black, 12% Hispanic, and 0% Native Hawaiian or Pacific Islander. Participants average age was 22.4 years, and 83% self-reported English as their first language.

The criteria for inclusion in the study was that participants must: (1) be an affiliate of the university, (2) be at least 18 years old, (3) be fluent in English (materials and instructions were in English); (4) not have significant and uncorrected vision impairment; (5) have access to a computer with a webcam, microphone, and speakers for video conferencing; and (6) had not previously played Physics Playground or a similar game (e.g., Crayon Physics Deluxe, Magic Pen) for more than an hour. Further, a stable internet connection was recommended but not required.

Participants were compensated with a $40 Amazon electronic gift card for their participation.

## 3.2 Procedure

Participants were assigned to teams of two (i.e., dyads) based on their availability. They were emailed the link to the Zoom video conference 24-h in advance of their study. Teammates were introduced in a virtual room (via Zoom) upon the start of the study. There were two experimenters who also joined via Zoom to facilitate data collection. One experimenter facilitated the back-end tasks pertaining to running the models and configuring the interventions, whereas the other facilitated interactions with the participants. The experimenters communicated via chat and voice/video while interacting with the participants but turned off their camera/microphones while participants were working collaboratively in the main Zoom room. Participants were separated into two breakout rooms to complete various individual tasks (e.g., engagement with the interventions, surveys), and the experimenters did not join these rooms except to deliver instructions and provide help as needed. The study lasted for about 2.5 h (with breaks) and was divided into three main phases, described in turn in the following sections. Appendix A provides additional details on the study procedure.

### 3.2.1 Preliminary surveys and introductory materials

Participants were separated into breakout rooms to complete introductory study materials. These included a short survey, which gathered demographic and background information reported above. We used the Physics Playground problem-solving environment (Sect. 2.1) for the collaborative task. Game levels were organized into a playground, and participants could freely explore and attempt any level within the playground (though they usually explore them in linear fashion). Participants were introduced to the game Physics Playground and completed a few tutorial levels with the game. Participants then rejoined to the main room for the main portion of the study.

### 3.2.2 Collaborative gameplay, intervention, & surveys

Upon joining the main room, participants were verbally provided the following main instructions for the study:

> *"Before we begin with the first round of play, I'm going to give you a quick overview of the study so you know what to expect. [Experimenter shares screen with an overview slide]. You've just completed the introductory survey and game tutorial. Next, we will do five rounds of game play and feedback. During each round of game play, you will collaborate to solve levels in Physics Playground for 10 minutes. Only one of you can control the game at a time, so [Participant A] will control for rounds 1, 2, and 3, then [Participant B] will control for rounds 4 and 5. After each round of game play, we will separate you into breakout rooms, and you will receive feedback on your collaboration.* ***Your goal is to use the feedback you receive to improve your collaboration in the next round****. After 5 rounds of play, we'll have you complete another survey to conclude the study, then we'll let you go"*

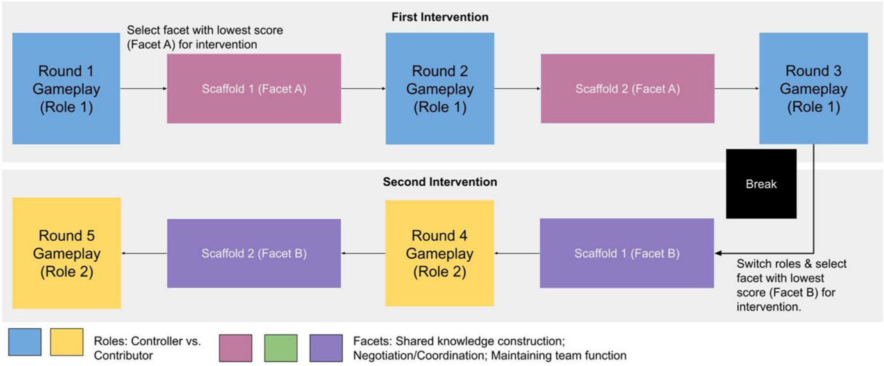

First Intervention
Select facet with lowest score (Facet A) for intervention
Round 1 Gameplay (Role 1)
Scaffold 1 (Facet A)
Round 2 Gameplay (Role 1)
Scaffold 2 (Facet A)
Round 3 Gameplay (Role 1)
Break
Second Intervention
Round 5 Gameplay (Role 2)
Scaffold 2 (Facet B)
Round 4 Gameplay (Role 2)
Scaffold 1 (Facet B)
Switch roles & select facet with lowest score (Facet B) for intervention.
Roles: Controller vs. Contributor
Facets: Shared knowledge construction; Negotiation/Coordination; Maintaining team function

**Fig. 8** Overview of the study procedure

Participants then commenced the first of five 10-min rounds of collaborative gameplay (Fig. 8). Round 1 served as an initial baseline to assess each participant's CPS scores. Participants were first shown a brief four-minute introductory video that explained the three facets and how their scores were generated prior to receiving their initial feedback. The facet with the lowest performing round 1 score (Facet A) was selected for intervention prior to rounds 2 and 3 (e.g., "Scaffold 1 [Facet A]" and "Scaffold 2 [Facet A]" in Fig. 8). Then, after a 5-min break, round 3 served as the second baseline, and interventions to improve the lowest performing facet from round 3 (i.e., Facet B where B $\neq$ A) were provided prior to rounds 4 and 5 (e.g., "Scaffold 2 [Facet B]" and "Scaffold 2 [Facet B] in Fig. 8). Participants also switched roles prior to round 4.

Participants received the following reminder of their goals after engaging with the intervention and before resuming gameplay (i.e., prior to rounds 2–5).

> *"Your goal is to work together as a team to solve as many levels as you can, while using the feedback you received to improve your collaborative problem solving."*

Feedback on all three facets were provided after round 1 so participants could have a sense of their baseline performance. Feedback was only provided on the facets selected for intervention (i.e., Facets A and B) immediately prior to receiving the designated interventions (i.e., before rounds 2 and 3 for Facet A and before rounds 4 and 5 for Facet B). We collected participants' perceptions of the accuracy of the feedback immediately after it was provided using the following single-item measure: "*Your score for < facet name > was < X% >* . *How accurate was this score?* They responded using a five-point unipolar scale: 1 (not at all accurate), 2, 3, 4, and 5 (very accurate).

We also collected participants' perceptions of the intervention using the following six-items from the intrinsic motivation inventory (Deci and Ryan 1982): (1) *I enjoyed doing this activity very much; (2) I am satisfied with my performance at this task; (3) I didn't try very hard to do well on this activity – reverse coded; (4) I did not feel nervous at all while doing this; (5) I believe this activity could be of some value* *to me; and (6) I found my thoughts wandering spontaneously during this activity – reverse coded.* They were instructed to reflect on their "*experience in the previous round, including playing Physics Playground and viewing your scores and feedback*." Participants responded using a five-point Likert scale (strongly disagree, somewhat disagree, neutral, somewhat agree, strongly agree).

#### 3.2.3 Post-intervention surveys & debriefing

After all five rounds of gameplay and intervention, participants individually completed a final survey including five-items modified from the System Usability Scale (Lewis 2018) as a measure of general usability. The specific items were: *(1) I thought the feedback system was easy to use; (2) I imagine that most people would learn to use the feedback system very quickly; (3) I found the feedback system very cumbersome to use; (4) I thought there was too much inconsistency in the feedback system; and (5) I found the feedback system unnecessarily complex*. They were instructed to consider all four rounds of interaction with the feedback system, which was defined as "*The feedback system refers to the Qualtrics surveys you viewed after each round of play, and includes the scores you received, the tips for actions you can take, the video and transcript examples from other teams, and the questions which tested your understanding.*" Participants responded using the same five-point Likert-type scale used for the intrinsic motivation inventory.

They completed a few additional items not relevant to the present study, upon which they were debriefed.

### 3.3 Data treatment

#### 3.3.1 Expert scoring of CPS facets and assessing model accuracy & generalizability

We recruited a trained human coder to manually score the automatically transcribed utterances of the CPS interactions alongside Zoom video recordings to provide the necessary context. This coder was one of the original coders of the reference dataset used to train the computational models. This coder scored the first seven minutes of each round of CPS interactions (i.e., the same data used to provide automated feedback), resulting in 11,608 coded utterances across 105 transcripts (21 teams × 5 rounds per team).

Table 2 contrasts the proportional occurrence of human-coded CPS scores from the training dataset alongside the current data. It also contrasts the mean model predicted scores across data sets before norm-referencing. We note that ground-truth scores for shared knowledge construction were highly similar across datasets, but there was a 27% and 34% reduction in scores for negotiation/coordination and maintaining team function, respectfully. This may be due to the current study which focused on dyads, whereas the training data was on triads. Nevertheless, the model-predicted probabilities were highly consistent across both datasets. Comparisons of model- vs. human- scores for the current data indicated close alignment for shared knowledge

**Table 2** Proportional occurrence of facets across training and current datasets and for ground-truth human coding vs. model predictions

| Facet | Ground truth (human-coded) | | Model predicted | |
|---|---|---|---|---|
| | *Training* | *Current* | *Training* | *Current* |
| Construction of Shared Knowledge | .261 | .254 | .257 | .248 |
| Negotiation/Coordination | .181 | .142 | .163 | .174 |
| Maintaining Team Function | .171 | .102 | .101 | .107 |

construction and maintaining team function, but the model tended to underpredict and overpredict negotiation and coordination by about 23% on the current data.

In terms of AUROCs, model accuracies for the current data were similar to what was achieved on the training set, suggesting evidence for generalizability: shared knowledge construction (AUROCs: 0.85 current vs. 0.88 training); negotiation coordination (0.80 current; 0.83 training); and maintaining team function (0.79 current vs. 0.82 training). Overall, we considered the AUROC scores (0.79 to 0.85) on the current data to be sufficiently accurate for automated feedback, especially since scores were averaged from the utterance level to the round level, thereby increasing reliability due to the principle of aggregation (Li et al. 1996).

To equate for baseline differences across facets and roles (not shown here), we norm-referenced the facet scores for the current data using the same procedure as in Sect. 2.2.4. The model training data were used for norming the automated scores because this reflects the feedback presented to the participants. However, the distributions derived from the current data itself (rather than the training data) were used to norm reference the human-coded CPS scores, which served as our dependent variable for RQ1 (changes in CPS scores).

We also investigated the extent to which the facets selected for intervention, which was based on the automated scores from the trained models, aligned with hypothetical selections based on the ground-truth human-coded scores. Recall that interventions for rounds 2 and 3 were selected based on the facet with the lowest computer-generated round 1 score. This aligned with the lowest- and second-lowest human-coded score 55% and 29% of the time, respectively. Thus, only 17% of the time was the facet with the highest human-coded score selected for intervention. Similarly, interventions for rounds 4 and 5 were selected based on the lowest automated round 3 score with the exception that it could not have been a facet selected for intervention during rounds 2 and 3. Here, 62% and 26% of the facets selected for intervention were also ranked lowest or second-lowest, respectively, based on the human-coded scores. Again, only 12% of the selected facets had the highest human-coded ranking. Thus, the overall conclusion is that facets selected for intervention based on the automatically computed CPS scores were sufficiently, albeit not perfectly, aligned with what would have been selected from the human-coded CPS scores.

**Table 3** Description of main dependent variables analyzed

| Variable | Description |
|---|---|
| RQ1: Ground Truth Score | Human coded norm-referenced (on current data) CPS scores (percentiles); higher scores indicate better performance |
| RQ2: Perceived Accuracy | Participants self-reported perceptions of feedback accuracy |
| RQ2: Intrinsic Motivation | Subjective perceptions of the interaction; higher scores are better |
| Automated Feedback Score | Model-generated norm-referenced (on training data) CPS scores (percentiles); higher scores indicate better performance |
| Automated Feedback Error | Difference between model- and human- non-normed CPS scores; positive error indicates models are providing higher scores compared to humans |
| Intervention Duration | Time (in minutes) spent engaging with the interventions |

### 3.3.2 Other variables

The main variables are listed in Table 3, which were used to answer our main research questions pertaining to improvements in CPS skills (RQ1—ground-truth score) and users subjective perceptions of CPSCoach 2.0 (RQ2—perceived accuracy and intrinsic motivation). The other measures (automated score, feedback error, and intervention duration) were used to further explore the data.

The intervention durations were examined for outliers. We corrected time spent engaging with the interventions by recoding values greater than 10 min (ostensibly due to the experimenter failing to end the intervention after the time limit) to 10 min. For the intrinsic motivation inventory (IMI), we first reverse-scored items 3 and 6 (Sect. 3.2.2). A reliability analysis (Cronbach's alphas) indicated that dropping item 3 increased the reliability from 0.53 to 0.58, thereby bringing it closer to the recommended minimum threshold of 0.6 (Rosenthal and Rosnow 1984). Accordingly, we averaged the remaining five items.

The total dataset was comprised of 630 cases (42 participants × 3 facets × 5 rounds), although the number of cases per variable is lower due to varying measurement schedules. Appendix B provides histograms of the variables analyzed after averaging them to the participant level (i.e., across facets and rounds).

### 3.3.3 Statistical modeling approach

The data included repeated measures (multiple facets and rounds per participant), so we used mixed-effects regression models using the lme4 package in R for model estimation. We used type 3 ANOVAs from the car package to test for significant main effects in the presence of interactions; we reverted to type 2 ANOVAs when the interaction terms were non-significant. We started with maximal random effects structures with random intercepts and slopes by participant and reduced complexity until there were no convergence issues. In cases where the random effect variances were 0, we reverted to basic (i.e., non-multilevel) models. All numeric variables were z-score-standardized so coefficients could be interpreted as standardized effects (i.e., $\beta$ coefficients). Post hoc analyses of significant interactions were conducted using the emmeans and emtrends functions from the emmeans package. We used two-tailed tests with a $p < 0.05$ cutoff for significance and applied a false discovery rate (fdr) adjustment (Benjamini and Hochberg 1995) for multiple comparisons; these are designated as $p_{fdr}$.

# 4 Results

## 4.1 Descriptives and correlations

Table 4 lists descriptives and correlations among the variables analyzed here. Overall, ground-truth human-coded scores were moderately to strongly correlated ($r = 0.64$) with automated feedback scores. Perceptions of feedback accuracy ($M = 3.3$) were above the middle of the 1–5 scale and correlated with both types of feedback scores, more so for the automated scores that were displayed to participants. Importantly, feedback error did not correlate with participants' perceptions of accuracy of the feedback, but did correlate with both ground-truth and automated feedback scores. Participant spent an average of 5 min (out of a maximum of 10-min or about half the available time) engaging with the intervention, which decreased from a mean of 6.5 min after round 1 to 5 min after rounds 2 and 3 and then 4 min after round 4. Time spent engaging with the intervention was related to perceptions of feedback accuracy ($r = 0.25$). Lastly, intrinsic motivation did not correlate with any of the other variables.

**Table 4** Descriptive statistics and correlations among key variables

| Variable | *M (SD)* | Ground truth score | Feedback score | Feedback error | Perceived accuracy | Intervention duration |
|---|---|---|---|---|---|---|
| Ground Truth Score | .49 (.10) | | | | | |
| Automated Feedback Score | .58 (.15) | .64** | | | | |
| Automated Feedback Error | − .03 (.02) | − .44** | .29** | | | |
| Perceived Accuracy | 3.3 (1.1) | .24* | .34** | .12 | | |
| Intervention Duration | 5.0 (1.9) | − .03 | − .01 | .06 | .25* | |
| Intrinsic Motivation | 3.9 (.56) | .19 | .03 | − . 16 | .18 | .18 |

### 4.2 RQ1: changes in CPS scores across rounds

Participants received the intervention for their lowest scoring facet from round 1 prior to rounds 2 and 3 and again on a different facet with lowest round 3 scores prior to rounds 4 and 5 (with the constraint that the same facet was never repeated, if this second facet was the same as the first, the next lowest facet was selected). Thus, round 1 scores serve as a baseline for the subsequent round 2 and 3 interventions, and round 3 scores serve as a baseline for the subsequent round 4 and 5 interventions. The intervention would be successful to the extent that scores for the facet selected for intervention (called the *treated* facet) increased after receiving the intervention. However, any change in scores for the treated facet might not be solely attributed to the intervention itself. Indeed, scores could change if the initial values were outliers and subsequent scores would revert back to the true value (i.e., regression to the mean), they could improve simply due to practice effects, they could decrease due to fatigue effects, or they could remain flat due to ceiling effects. To account for these possibilities, we utilized a matched-control quasi-experimental design (Stuart and Rubin 2008) where intervention cases were paired with *matched* control cases equated on baseline scores.

#### 4.2.1 Creating treatment & matched controls

The matching procedure proceeded as follows. Consider round 1, where each participant has three facet scores, of which the lowest score was selected for intervention. Thus, of the 126 cases (42 participants × 3 facets), a third ($n = 42$) were considered *treated* cases. Of the remaining 84 cases that did not receive an intervention, the matching procedure aims to identify a subset of 42 *matched-control* cases such that the mean difference in their scores *prior* to the intervention (i.e., baseline scores) is statistically equivalent to that of the treated cases while also equating for pertinent covariates such as facet and role. Once matching is done, the key comparison pertains to changes in scores for the treated compared to the matched control cases *after* receiving the intervention. Because baseline scores are equivalent for these two groups and other aspects of the design such as role and facet are balanced, we have more confidence that any observed differences are attributable to the intervention itself rather than incidental factors.

We performed bipartite cardinality matching using the bmatch function from the designmatch package to identify the matched control cases with the constraint that the mean difference in baseline scores for treated vs. matched control cases was less than 0.05 standard deviations and that distributions of roles and facets were close to equivalent (called fine balance). Matching was done twice, first for the first set of interventions (i.e., rounds 1–3 with round 1 as the baseline) and then again for the second set of interventions (i.e., rounds 3–5 with round 3 as the baseline). When matching for the second round of interventions (i.e., rounds 3–5), we removed cases corresponding to facets selected for the first set of interventions (i.e., rounds 1–3). Even though round 3 scores were included twice, the analyzed data pertains to two different facets (one for each set of interventions).

Table 5 provides descriptives on the covariates for the treated (T) and matched-control (C) groups before and after the matching procedure. Prior to matching, the

**Table 5** Covariates before and after matching

| Covariate | Round 1 | | | | Round 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | Non-matched | | Matched | | Non-matched | | Matched | |
| | C | T | C | T | C | T | C | T |
| *n* | 84 | 42 | 42 | 42 | 42 | 42 | 25 | 25 |
| Baseline score*:* M (SD) | **.52 (.30)** | **.33 (.27)** | .35 (.25) | .33 (.27) | **.64 (.24)** | **.35 (.26)** | .50 (.20) | .50 (.23) |
| Observer Role: n (%) | 42 (50) | 21 (50) | 21 (50) | 21 (50) | 21 (50) | 21 (50) | 13 (52) | 13 (52) |
| Shared Knowledge Const: n (%) | 26 (31) | 16 (38) | 16 (38) | 16 (38) | **7 (17)** | **19 (45)** | 7 (28) | 7 (28) |
| Negotiation/Coordination: n (%) | 29 (35) | 13 (31) | 13 (31) | 13 (31) | **20 (48)** | **9 (21)** | 9 (36) | 9 (36) |
| *Maintain.* Team Function: n (%) | 29 (35) | 13 (31) | 13 (31) | 13 (31) | 15 (36) | 14 (33) | *9* (36) | *9* (36) |

T = treated; C = matched-control; Bold = significant difference control group consisted of the baseline scores for the two non-treated facets. By design, it had significantly higher baseline scores than the treated group for both round 1 ($p <$ 0.01) and round 3 ($p < 0.001$). However, the two groups were equivalent on baseline scores (*ps* > 0.73) after matching. Whereas all 42 treatment cases were matched to a control in round 1, only 25 were successfully matched in round 3. Lastly, there were no significant differences across treatment and control groups across roles and facets before matching for round 1. However, for round 3, there was considerable imbalance in facets in the two groups prior to matching; they are perfectly balanced after matching.

### 4.2.2 Changes in scores across rounds for treatment and matched controls

We investigated changes in ground-truth scores via Model 1 below. Here, round was a numeric independent variable, and condition (treatment, matched control [reference group], unmatched) and role (controller [reference group] vs. observer) were moderators. We included facet (shared knowledge construction [reference group], negotiation/coordination, & maintaining team function) and the number of utterances as covariates to account for facet differences and verbosity effects (i.e., higher scores for those who speak more).

$$\begin{aligned} Model1 &: CPSGroundTruthScore \\ &\sim Round \times Condition \times Role + Facet + Verbosity + RandomEffects \end{aligned}$$

This model specification was designed to examine the slope or rate of change in CPS scores across rounds and critically to investigate whether the slopes varied by condition. Even though we were primarily interested in slope differences for the treatment vs. matched control, we included the unmatched cases for unbiased estimation of parameters. The inclusion of role as moderator allowed us to investigate whether changes in slopes by condition (i.e., the round × condition interaction) further varied by role (i.e., separate slopes for each condition for controllers and observers—six total). This model specification also includes all main effects and two- and three-way interactions, though we are mainly interested in effects involving condition. We estimated separate models for the first (i.e., rounds 1–3 with round 1 as the baseline [i.e., pre-intervention score], $n = 378$) and second (i.e., rounds 3–5 with round 3 as the baseline, $n = 201$) set of interventions. Note that assignment of role switched before round 4.

**Changes in scores for rounds 1–3.** The model that converged included random intercepts and slopes for condition (i.e., condition | participant). There was a significant main effect of condition, $X^2(2) = 12$, $p = 0.002$, but this was subsumed by the significant round × condition interaction, $X^2(2) = 15$, $p < 0.001$, suggesting that the rate of change of CPS scores across rounds varied by condition. As shown in Figure 9A, both the treatment and matched controls showed improvements in slopes. However, the slope for the treatment ($\beta = 0.27$ [0.13, 0.42]) was 80% steeper than the matched control ($\beta = 0.15$ [0.01, 0.30]). There were no other significant interactions with condition.

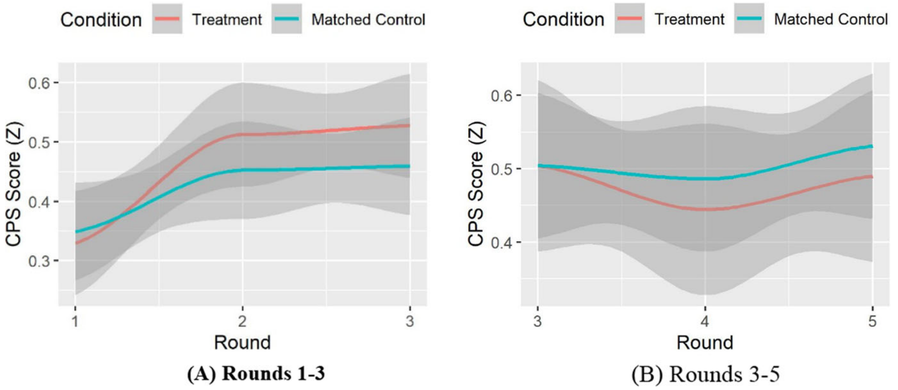


**Fig. 9** CPS scores across rounds **A** 1–3 and **B** 3–5 for treatment and matched control. Note that the values for round 3 are different for **A** and **B** because they pertain to different treatment and matched controls

**Changes in scores for rounds 3–5.** The model that converged only included a random intercept (i.e., 1 | participant). There was a significant condition × round interaction, $X^2(2) = 14$, $p = 0.001$. However, as indicated in Fig. 9B, scores were largely flat for both treatment ($\beta =$ -0.03 [-0.23, 0.17]) and for matched control ($\beta =$ 0.04 [-0.16, 0.24]). There was also a significant condition × role interaction, $X^2(2) =$ 14, $p = 0.001$, but the significant differences only pertained to the unmatched group, which is not of interest here.

## 4.3 RQ2: subjective perceptions of the system

### 4.3.1 Perceptions of feedback accuracy

Participants generally thought the feedback was moderately accurate (mean of 3.3 on a 1–5 scale). We investigated the factors that predicted it starting with Round 1 where participants viewed feedback and self-reported their perceptions of its accuracy for all three facets prior to any intervention ($n = 126$ cases; 42 participants × 1 round × 3 facets). We included role, facet, feedback score, feedback error, and number of utterances as predictors in Model 2. Next, we re-ran the model for rounds 2–4, where participants received feedback on and rated the accuracy of the facet selected for intervention ($n = 126$ cases; 42 participants × 3 rounds × 1 facet). This second model included round as a predictor in addition to the above predictors (Model 3).

$$\begin{aligned} Model2: Perceptionsof Accuracy(Round1) &\sim Role + Facet + FeedbackScore \\ &+ FeedbackError + Verbosity + RandomEffects \end{aligned}$$

$$\begin{aligned} Model3: Perceptionsof Accuracy\,(Rounds2-4) \\ \sim Role + Facet + FeedbackScore + FeedbackError \\ + Verbosity + Round + RandomEffects \end{aligned}$$

Models with intercept only random effects converged (i.e., 1 | participant). In both cases, the only significant effect was for the feedback score, $X^2(1) > 53$, $ps < 0.001$, indicating that participants tended to be more confident in the feedback when they received higher scores ($\beta = 0.55$ [0.40, 0.70] for round 1 and $\beta = 0.59$ [0.45, 0.74] for rounds 2–4). Importantly, their perceptions of feedback remained stable across rounds ($\beta$ = -0.09, $p = 0.11$).

### 4.3.2 Perceptions of the interventions

Overall participants' perceptions of the intervention were positive with a mean of 3.9 ($SD = 0.56$) on the intrinsic motivation inventory (1–5 scale). We investigated factors associated with these scores by regressing them on role, facet, feedback score, feedback error, round, and number of utterances ($n = 163$ due to occasional missing data; Model 3 but with a different dependent variable). Only an intercept-only random effect model converged (i.e., 1 | participant). Results indicated significant main effects for feedback score, $X^2(1) > 6.6$, $p = 0.01$, and number of utterances (verbosity), $X^2(1) > 3.9$, $p < 0.05$. Those who received higher scores had more positive perceptions of the intervention ($\beta = 0.15$ [0.03, 0.26]) as did those who were more verbose ($\beta = 0.15$ [0, 0.30]).

Participants also completed a modified version of the system usability scale as a measure of general usability at the end of the study using a 1–5 scale. Participants were neutral with respect to the ease of use of the system, $M = 2.9$ ($SD = 1.2$), and its complexity, $M = 3.1$ ($SD = 1.1$), but they found it to be somewhat cumbersome, $M = 3.3$ ($SD = 1.0$) and inconsistent, $M = 3.6$ (1.2). They were modestly positive about its learnability, $M = 3.49$ ($SD = 1.1$).

## 5 Discussion

We developed and tested CPSCoach 2.0, an intelligent system that provides automated feedback and personalized instructional scaffolds to improve peoples' collaborative problem-solving (CPS) skills. We integrated our intervention in the context of a virtual game-based collaborative learning environment with naturalistic open-ended interaction. We conducted a pre-post, matched-control quasi-experimental study to examine whether engaging with CPSCoach 2.0 was associated with improvements in targeted CPS skills (RQ1). We also investigated users' subjective perceptions of the system and the factors related to their perceptions (RQ2). We discuss the main findings along with design implications, generalizability, limitations, and future work.

### 5.1 Main findings and implications

Our main finding (RQ1) was that CPSCoach 2.0 had some success in improving CPS skills after the first set of interventions. Specifically, facets that were selected for intervention improved at a higher (80%) rate than matched controls that did not receive the intervention despite having equivalent baseline scores. To our knowledge, this is the first demonstration of improvement of CPS skills via automated feedback and personalized instructional scaffolds. Whereas previous work (Sect. 1.1.3) has targeted improvement on related lower-level behaviors (e.g., turn taking, improving communicative tone) and intermediate-level socio-cognitive constructs (e.g., shared attention, rapport), CPS is more than the sum of these behaviors because it involves both problem-solving and collaboration. For example, equitable verbal participation among teammates is likely related to maintaining a positive team environment as everyone should have the opportunity to voice their ideas. However, this is insufficient as the team could be engaging in hostile conversations, or be off-topic, neither of which might contribute to positive CPS outcomes. Similarly, high verbosity itself might not be indicative of effective CPS if the team does not engage in productive metacognitive reflection when their solutions fail. Accordingly, we moved from the more general collaborative behaviors to higher-level CPS skills and found that principles from the science of learning embedded in intelligent technologies can be leveraged at promoting these skills.

We also had a null finding in that there were no changes in scores for the second set of interventions after roles were reversed and a second facet was selected for feedback. We chose to switch roles in order to give both users an opportunity to control the gameplay. However, it is likely that adjusting to a new role while also focusing on a different facet for feedback resulted in too high of a cognitive load for users, which might explain the lack of any further improvement. Another possibility is that because the intervention materials were role-agnostic (i.e., they were the same for both roles), users had difficulty applying the principles from one role to another. Fatigue effects after three rounds of gameplay and two phases of the first intervention might also have contributed to the lack of effects for the second intervention. These findings suggest that it may be prudent to keep collaboration roles consistent for a given session; role switching might occur for subsequent sessions or when users have demonstrated a sufficient level of mastery of CPS skills in a given role. Second, there may be benefits to customizing the intervention materials for each role, especially given the additional cognitive load on the Controller who has to also manipulate the game environment in addition to problem solving and interacting with the two Collaborators.

Turning to users' perceptions of the feedback and interventions (RQ2), we found that those who received higher scores as feedback, found it to be more accurate and had more positive impressions of the intervention. Interestingly, the model errors did not predict participants' perceptions of the feedback and intervention, suggesting that they might not have been sensitive to these errors. This finding is consistent with prior research on speech-based learning systems, which has found that perfect speech recognition and natural language processing are not necessary for beneficial interactions (D'Mello et al. 2010; Forbes-Riley and Litman 2011). In the present case, we ameliorated some of the errors by using deep learning models that focus on semantics and by increasing reliability by aggregating noisy predictions from multiple utterances across a seven-minute window. Overall, users' perceived CPSCoach 2.0 as being moderately accurate and generally had positive impressions of it.

### 5.2 Scalability and generalizability

Our approach to CPS feedback is highly scalable and has potential to generalize beyond the current context. Regarding scalability, users only need access to a computer with standard webcam, microphone, and Internet connection; no specialized sensors are needed. With respect to generalizability, we expect that this work can be applied to similar contexts because this research was predominantly conducted in-the-wild where participants engaged from their homes. For example, in our user studies we encountered interrupting housemates and distracting notifications from other computer applications. This is part and parcel of remote work in the post-pandemic era (Erik et al. 2020; Breideband et al. 2022), which makes the study highly relevant. One additional way our work supports generalizability is through the use of a CPS framework, which was selected for its ability to generalize across domains (Sun et al. 2020). Although the examples used in the intervention were derived in the context of a specific collaborative task, all other content and activities were domain-independent. We also demonstrate generalizability in our CPS models, which transferred from training to deployment contexts despite key differences. In particular, the models were trained on pre-pandemic data consisting of triads collaborating in a laboratory environment and applied to dyads collaborating from their chosen, remote location post-pandemic with only a slight reduction in accuracy. Taken together, the models and instructional scaffolds can be ported to new domains with some additional data collection for fine-tuning and iterative refinement. Future studies will be needed to examine generalizability to additional contexts.

### 5.3 Limitations and future work

Like all studies, ours has limitations. First, our sample size was small ($n = 42$) and might not have had sufficient power to detect some of the effects. The reliability on the intrinsic motivation inventory was also low, suggesting that additional items or a different instrument should be used for future work. Second, we elected to use a quasi-experimental, pre-post, matched-control design in this initial user study of CPSCoach 2.0. Our goal was to simply demonstrate whether engaging with CPSCoach 2.0 could be associated with improvements in CPS scores and to examine user perceptions of the system. Even though such designs are valid methods for generalized causal inference (Cook et al. 2002) and have been used to evaluate educational applications (Koedinger et al. 2010), the design did not afford assessing improvements in CPS outcomes (i.e., gameplay success) because there was no non-intervention control group. Because findings from this study indicated areas of improvement for CPSCoach 2.0, it is prudent to make these changes before conducting a well-powered randomized controlled trial.

Along those lines, another limitation pertained to some users finding the system to be cumbersome and inconsistent, suggesting some redesign is needed. Additional improvements would include tailoring intervention content to different roles. We also only focused on a single CPS task in one domain, and the collaboration and intervention times were quite short. It is possible that more time and extended periods of targeted feedback are needed to further improve CPS. Relatedly, it is likely that a single session might not provide sufficient practice and opportunities to fully develop CPS skills. Other limitations pertain to the use of college students as study participants and fixed assignment of roles. Generalizability across multiple CPS activities, different populations, and spontaneous emergence of roles should be warranted. We also did not analyze measures of Physics learning due to the relatively short duration of interactions with the game and lack of learning supports; previous studies indicate the need for longer, and multiple gameplay sessions to obtain reliable learning gains (e.g., 4 h over 1.5 weeks (Shute et al. 2013)) and the provision of explicit supports pertaining to the underlying physics concepts (Rahimi et al. 2022).

Taken together, an important item for future work is to extend the intervention into a multi-session use of CPSCoach 2.0, where users receive feedback and scaffolds on all three facets, in multiple roles, different domains and activities, and collaborate with multiple teammates. Such a longitudinal study will also afford collection of additional measures, including learning outcomes.

### 5.4 Concluding remarks

Researchers, practitioners, and policy makers have lamented the dearth of collaborative problem-solving (CPS) skills among people and have advocated for educational interventions that intentionally target the development of these skills (Griffin et al. 2012b; Fiore et al. 2018). Accordingly, we developed CPSCoach 2.0, as a proof-of-concept system for the development of CPS skills where users collaborate to solve complex problems, receive automated feedback on that collaboration, engage in personalized instructional scaffolds, and have opportunities to apply the gained knowledge in subsequent collaboration cycles. A quasi-experimental, pre-post, matched-control, design indicated positive benefits of CPSCoach 2.0, but also areas of improvement to inform the design and testing of the next version of the system. Additionally, the overall approach could be applied to other twenty-first-century skills, where there are also widespread calls for intentional approaches to help people develop these valuable skills (Dede 2010; Griffin et al. 2012b). Thus, the present work demonstrates a proof-of-concept design of the use of automated feedback and instructional scaffolds to support building critical twenty-first-century skills.

**Acknowledgements** This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under DRL 2019805 and by DUE 1745442/1660877. The opinions expressed are those of the authors and do not represent views of the NSF. We also thank Chen Sun, Arjun Rao, Valerie Shute, Sam Pugh, and Quinton Beck-White for their contributions to the research.

**Author contributions** S.D. and A.S wrote the main manuscript text. N.D. and A.G. provided reviews and feedback. A.G. and A.S. conducted the user studies. All authors intellectually contributed to the research.

## Declarations

**Conflict of interest** The authors have no conflicts of interests.

**Ethical approval** All research conducted has been approved by the cognizant IRBs and participants provided informed consent prior to the study.

## Appendix A: Details on Study Procedure

We provide the following additional details to complement the overview provided in Sect. 3.2.

**Round 1–3 gameplay.** A randomly selected participant was chosen as the Controller (i.e., who interacts with the game mechanics) and the other as the Contributor (i.e., who provides suggestions) prior to the start of the first round. After completing 10 min of gameplay (Round 1), participants were separated into breakout rooms again and informed that they would receive feedback on their collaboration.

**Rounds 2 & 3 gameplay & intervention on Facet A.** Upon entering the breakout room, participants were shown a brief four-minute video that explained the three facets of the CPS framework and how their scores were generated. Participants then received feedback on all three facets and self-reported their perceived accuracy of the feedback for each facet. The facet with the lowest Round 1 score (Facet A) was selected for improvement, and this was communicated to the participants along with the Scaffold 1 Intervention. Participants had a maximum of 10 min to engage with the intervention, upon which the experimenter intervened. Participants who completed the intervention before the 10-min interval had elapsed simply informed the experimenter who instructed them to wait for their partner.

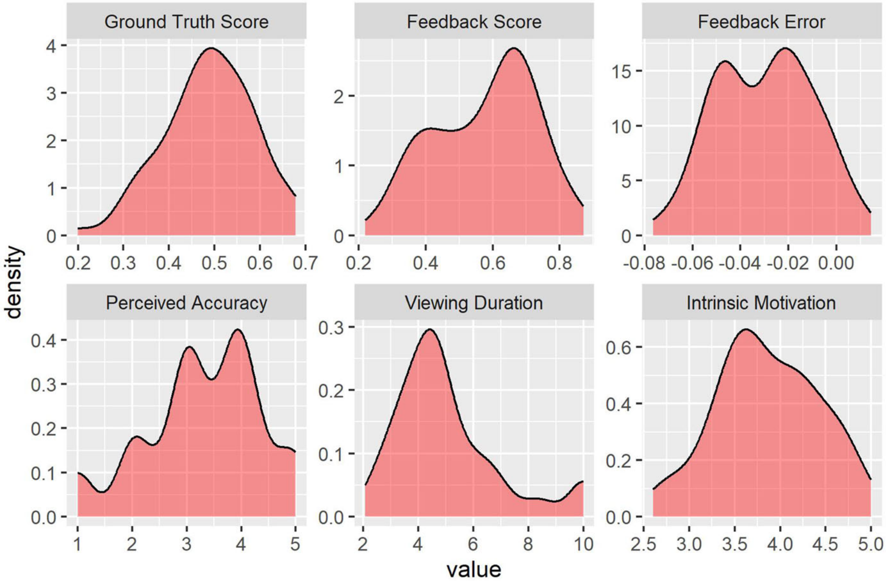
When both partners were back in the main room, a second 10-min round of gameplay (Round 2) commenced with the same participants assigned to the Controller vs. Contributor roles. When 10-min had elapsed, they were sent to separate breakout rooms where they received feedback on the same facet selected for improvement in the previous round (i.e., Facet A). They once again self-reported their perceptions of feedback accuracy for this facet only, upon which they received the Scaffold 2 Intervention to further improve on the same facet (i.e., Facet A). When both participants were done with the intervention or 10 min had elapsed, participants re-entered the main room and completed a third 10-min round of collaborative gameplay (Round 3).

**Rounds 4 & 5 gameplay & intervention on Facet B.** After Round 3, participants once again were moved to separate breakout rooms. They were then given a five-minute break. When they returned, they were informed that they would now focus on improving performance on a different facet (i.e., Facet B – the one with the second lowest score during Round 3) using the following instructions.

> *"In the following round you will change roles with your teammate. If you were controlling the game, you will now be observing your teammate and offering suggestions. If you were previously observing, you will now have control of gameplay. We'd like you to now focus on a different aspect of collaboration: <Facet Name>"*

They received the Scaffold 1 Intervention, but this time for Facet B, and engaged with it as before. Once they were done or time had elapsed, they returned to the main room. They were informed that they would now switch roles with the previous Controller now becoming the Contributor and vice versa. They completed another 10-min of gameplay (Round 4) in their new roles. Next, they again moved to separate breakout rooms where they received feedback for Facet B, self-reported its perceived accuracy, and engaged with the Scaffold 2 Intervention for Facet B. Then, they returned to the main room and completed one last round (Round 5) with their new roles.

## Appendix B: Histogram of Variables Analyzed



## References

Alterman, R., Harsch, K.: A more reflective form of joint problem solving. Int. J. Comput. Support Collab. Learn **12**, 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.. Sci. (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. Human Behav. **104**, 105759 (2020)

Aran, O., Gatica-Perez. D (2010) Fusing Audio-Visual Nonverbal Cues to Detect Dominant People in Group Conversations. In: 2010 20th International Conference on Pattern Recognition. pp 3687–3690

Azevedo, R., Bernard, R.M.: A meta-analysis of the effects of feedback in computer-based instruction. J. Educ. Comput. Res. **13**, 111–127 (1995)

Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc. Series B-Methodol. **57**, 289–300 (1995)

Beyan, C., Carissimi, N., Capozzi, F., et al (2016) Detecting Emergent Leader in a Meeting Environment Using Nonverbal Visual Features Only. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, New York, NY, USA, pp 317–324

Bosch, N., D'Mello, S., Baker, R., et al (2015) Automatic detection of learning-centered affective states in the wild. In: Proceedings of the 20th international conference on intelligent user interfaces. pp 379–388

Bransford, JD., Brown, AL., Cocking RR (2000) How people learn

Breideband, T., Martinez, G., Sukumar, PT., et al (2022) Collaborating from Home during COVID-19: Examining Individual Sleep Patterns and Sleep Alignment

Brynjolfsson, E., Horton, JJ., Ozimek, A., et al (2020) COVID-19 and remote work: An early look at US data

Calacci, D., Lederman, O., Shrier, D., Pentland, A., "Sandy" (2016) Breakout: An Open Measurement and Intervention Tool for Distributed Peer Learning Groups. CoRR abs/1607.0:

Chi, M., Wylie, R.: The ICAP framework: linking cognitive engagement to active learning outcomes. Educ. Psychol. **49**, 219–243 (2014)

Chi, M., Siler, S., Jeong, H., et al.: Learning from human tutoring. Cogn. Sci.. Sci. **25**, 471–533 (2001)

Chopade, P., Edwards, D., Khan, SM., et al (2019) 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

Cook, T.D., Campbell, D.T., Shadish, W.: Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin, Boston, MA (2002)

Cukurova, M., Luckin, R., Millán, E., Mavrikis, M.: The NISPI framework: analysing collaborative problem-solving from students' physical interactions. Comput. Educ.. Educ. **116**, 93–109 (2018). https://doi.org/10.1016/j.compedu.2017.08.007

Cukurova, M., Zhou, Q., Spikol, D., Landolfi, L, (2020) Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough? In: Proceedings of the tenth international conference on learning analytics &amp; knowledge. association for computing machinery, New York, NY, USA, pp 270–275

D'Mello, S.K., King, B., Graesser, A.: Towards spoken human-computer tutorial dialogues. Hum. Comput. Interact.comput. Interact **25**, 289–323 (2010)

Davor, Č., Margaret Anne, DS., Davor, Č., Margaret Anne, DS (2005) Collaboration support for novice team programming. Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work pp 136–139. https://doi.org/10.1145/1099203.1099229

Deci, EL., Ryan, RM (1982) Intrinsic motivation inventory measurement instrument

de Kok, I, Heylen, D. (2009) Multimodal end-of-turn prediction in multi-party meetings. In: Proceedings of the 2009 International Conference on Multimodal Interfaces. ACM, New York, NY, USA, pp 91–98

Dede, C.: Comparing frameworks for 21st century skills. In: Bellanca, J., Brandt, R. (eds.) 21st century skills: Rethinking how students learn, pp. 51–76. Solution Tree Press, Bloomington IN (2010)

Devlin, J., Chang, M-W., Lee, K., Toutanova, K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv arXiv:1810.04805

Diehl, M., Stroebe, W.: Productivity loss in brainstorming groups: toward the solution of a riddle. J. Pers. Soc. Psychol. **53**, 497–509 (1987)

Dielmann, A., Garau, G., Bourlard, H (2010) Floor holder detection and end of speaker turn prediction in meetings. In: Proceedings of the International Conference on Speech and Language Processing, Interspeech. ISCA

Dowell, N.M.M., Nixon, T.M., Graesser, A.C.: Group communication analysis: a computational linguistics approach for detecting sociocognitive roles in multiparty interactions. Behav. Res. Methods. Res. Methods **51**, 1007–1041 (2019)

Dowell, N.M.M., Lin, Y., Godfrey, A., Brooks, C.: Exploring the relationship between emergent sociocognitive roles, collaborative problem-solving skills, and outcomes: a group communication analysis. J. Learn. Anal. **7**, 38–57 (2020)

Elliot, A., McGregor, H.: A 2 x 2 achievement goal framework. J. Pers. Soc. Psychol. **80**, 501–519 (2001). https://doi.org/10.1037//0022-3514.80.3.501

Eloy, L, E.B. Stewart, A., Jean Amon, M., et al (2019) Modeling team-level multimodal dynamics during multiparty collaboration. In: 2019 international conference on multimodal interaction. association for computing machinery, New York, NY, USA, pp 244–258

Erik, B., John, JH., Adam, O., et al (2020) COVID-19 and remote work: An early look at US data

Faucett, H.A., Lee, M.L., Carter, S.: I should listen more: real-time sensing and feedback of non-verbal communication in video telehealth. Proc. ACM Hum-Comput. Interact. **1**(44), 1–44 (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.behav. **2**, 367–369 (2018)

Forbes-Riley, K., Litman, D.J.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Commun.commun. **53**, 1115–1136 (2011). https://doi.org/10.1016/j.specom.2011.02.006

Friedrich, H., Esther, C., Juergen, B., et al.: A Framework for Teachable Collaborative Problem Solving Skills. Springer, Netherlands (2015)

Fusaroli, R., Rkaczaszek-Leonardi, J., Tylén, K.: Dialog as interpersonal synergy. New Ideas Psychol. **32**, 147–157 (2014)

Gigone, D., Hastie, R.: The common knowledge effect: Information sharing and group judgment. J. Pers. Soc. Psychol. **65**, 959–974 (1993)

Graesser, A.C., Fiore, S.M., Greiff, S., et al.: Advancing the science of collaborative problem solving. Psychol. Sci. Public Interest **19**, 59–92 (2018). https://doi.org/10.1177/1529100618808244

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, Netherlands, Dordrecht (2012a)

Griffin, P., McGaw, B., Care, E.: Assessment and teaching of 21st century skills. Springer, New York (2012b)

Gutwin, C., Bateman, S., Arora, G., Coveney, A. (2017) 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. ACM, New York, NY, USA, pp 2221–2235

Hesse, F., Care, E., Buder, J., et al.: 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, Netherlands, Dordrecht (2015)

Hill, GW. (1982) Group Versus Individual Performance : Are TV + 1 Heads Better Than One ? 91:517–539

Hong, S., Suh, M., Kim, T.S., et al.: Design for collaborative information-seeking: understanding user challenges and deploying collaborative dynamic queries. Proc. ACM Hum-Comput. Interact (2019). https://doi.org/10.1145/3359208

Janis, I.L.: Groupthink: Psychological studies of policy decisions and fiascoes. Houghton Mifflin, Boston (1982)

Jiangang, H., Lei, C., Michael, F., et al.: CPS-Rater: automated sequential annotation for conversations in collaborative problem-solving activities. ETS Res. Report Series **2017**, 1–9 (2017). https://doi.org/10.1002/ets2.12184

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

Karau, S.J., Williams, K.D.: Social loafing: a meta-analytic review and theoretical integration. J. Pers. Soc. Psychol. **65**, 681–706 (1993)

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

Kerr, N.L.: Motivation losses in small groups: a social dilemma analysis. J. Pers. Soc. Psychol. **45**, 819–828 (1983)

Kerr, N.L., Bruun, S.E.: Dispensability of member effort and group motivation losses: free-rider effects. J. Pers. Soc. Psychol. **44**, 78–94 (1983)

Kerr, N.L., Tindale, R.S.: Group performance and decision making. Annu. Rev. Psychol.. Rev. Psychol. **55**, 623–655 (2004)

Koedinger, K.R., McLaughlin, E.A., Heffernan, N.T.: A quasi-experimental evaluation of an on-line formative assessment and tutoring system. J. Educ. Comput. Res. **43**, 489–510 (2010)

Kori, K., Pedaste, M., Leijen, Ä., Mäeots, M.: Supporting reflection in technology-enhanced learning. Educ. Res. Rev. **11**, 45–55 (2014)

Krafft, PM., Baker, CL., Tenenbaum, JB. others (2016) Modeling human ad hoc coordination. In: Thirtieth AAAI Conference on Artificial Intelligence

Kütt, G.H., Tanprasert, T., Rodolitz, J., et al.: Effects of shared gaze on audio- versus text-based remote collaborations. Proc. ACM Hum-Comput. Interact (2020). https://doi.org/10.1145/3415207

Laughlin, P.R., Ellis, A.L.: Demonstrability and social combination processes on mathematical intellective tasks. J. Exp. Soc. Psychol. **22**, 177–189 (1986)

Laughlin, P.R., Kerr, N.L., Davis, J.H., et al.: Group size, member ability, and social decision schemes on an intellective task. J. Pers. Soc. Psychol. **31**, 522 (1975)

Laughlin, P.R., Hatch, E.C., Silver, J.S., Boh, L.: Groups perform better than the best individuals on letters-to-numbers problems: effects of group size. J. Pers. Soc. Psychol. **90**, 644 (2006)

Lewis, J.R.: The system usability scale: past, present, and future. Int. J. Hum. Comput. Interact.comput. Interact. **34**, 577–590 (2018)

Li, H., Rosenthal, R., Rubin, D.B.: Reliability of measurement in psychology: from Spearman-Brown to maximal reliability. Psychol. Methods **1**, 98–107 (1996). https://doi.org/10.1037/1082-989X.1.1.98

McGregor, M., Tang, JC. (2017) More to meetings: challenges in using speech-based technology to support meetings. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. ACM, New York, NY, USA, pp 2208–2220

Mercier, E.M., Higgins, S.E., da Costa, L.: Different leaders: emergent organizational and intellectual leadership in children's collaborative learning groups. Int. J. Comput. Support Collab. Learn (2014). https://doi.org/10.1007/s11412-014-9201-z

Michael, F., Su-Youn, Y., Jiangang, H., et al (2016) Automated classification of collaborative problem solving interactions in simulated science tasks. Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications pp 31–41

Miguel, J., Andres, L., Mercedes, M., et al (2014) An exploratory analysis of confusion among students using Newton's Playground. Proceedings of the 22nd International Conference on Computers in Education

Müller, P., Huang, MX., Bulling, A. (2018) Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In: 23rd International Conference on Intelligent User Interfaces. pp 153–164

Murray, G., Oertel, C. (2018) Predicting group performance in task-based interaction. In: Proceedings of the 20th ACM international conference on multimodal interaction. ACM, New York, NY, USA, pp 14–20

NASEM: How people learn II: Learners, contexts, and cultures. National Academies Press, Washington, DC (2018)

Nelson, L.M.: Collaborative problem solving. Inst. Des. Theories and Models **2**, 241–267 (1999)

Nihei, F., Nakano, YI., Hayashi, Y., et al (2014) Predicting influential statements in group discussions using speech and head motion information. In: Proceedings of the 16th international conference on multimodal interaction. ACM, New York, NY, USA, pp 136–143

Nijstad, B.A., Stroebe, W., Lodewijkx, H.F.M.: Production blocking and idea generation: does blocking interfere with cognitive processes? J. Exp. Soc. Psychol. **39**, 531–548 (2003)

OECD (2015a) PISA 2015 Results in Focus. OECD Publishing

OECD (2015b) PISA 2015 Collaborative problem solving framework. organisation for economic co-operation and development (OECD)

Olsen, J.K., Sharma, K., Rummel, N., Aleven, V.: Temporal analysis of multimodal data to predict collaborative learning outcomes. British J. Educ. Technol. **51**, 1527–1547 (2020). https://doi.org/10.1111/bjet.12982

Olsen, JK., Finkelstein, S. (2017) Through the (thin-slice) looking glass: An initial look at rapport and co-construction within peer collaboration. In: Proceedings of the 12th International Conference on Computer Supported Collaborative Work. International Society of the Learning Sciences, Philadelphia, PA, USA

Otsuka, K., Kasuga, K., Köhler, M. (2018) Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In: Proceedings of the 20th ACM international conference on multimodal interaction. ACM, New York, NY, USA, pp 191–199

Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng.knowl. Data Eng. **22**, 1345–1359 (2010)

Pugh, S.L., Rao, A., Stewart, A.E.B., D'Mello, S.K.: Do Speech-Based Collaboration Analytics Generalize Across Task Contexts? In: LAK22: 12th International Learning Analytics and Knowledge Conference, pp. 208–218. ACM, New York (2022)

Pugh, SL., Subburaj, SK., Rao, AR., et al (2021) Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21)

Rahimi, S., Shute, V.J., Fulwider, C., et al.: Timing of learning supports in educational games can impact students' outcomes. Comput. Educ.. Educ. **190**, 104600 (2022)

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 Heidelberg, Berlin, Heidelberg (1995)

Rosenthal, R., Rosnow, R.: Essentials of behavioral research: Methods and data analysis. McGraw-Hill, New York (1984)

Samrose, S., McDuff, D., Sim, R., et al (2021) Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp 1–13

Sanchez-Cortes, D., Aran, O., Mast, MS., Gatica-Perez, D. (2010) Identifying Emergent Leadership in Small Groups Using Nonverbal Communicative Cues. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction. ACM, New York, NY, USA, pp 39:1--39:4

Schlösser, C., Harrer, A., Kienle, A. (2018) Supporting dyadic chat communication with eye tracking based reading awareness. In: 2018 IEEE 18th international conference on advanced learning technologies (ICALT). pp 149–151

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**, 66–95 (2017). https://doi.org/10.1177/2041386616675522

Scoular, C., Care, E.: Monitoring patterns of social and cognitive student behaviors in online collaborative problem solving assessments. Comput. Human Behav. **104**, 105874 (2020). https://doi.org/10.1016/j.chb.2019.01.007

Shree Krishna, S., Angela, EBS., Arjun Ramesh, R., Sidney, KD. (2020) Multimodal, Multiparty Modeling of Collaborative Problem Solving Performance. Proceedings of the 2020 Conference on Multimodal Interaction

Shute, V.: Focus on formative feedback. Rev. Educ. Res. **78**, 153–189 (2008)

Shute, V.J., Ventura, M., Kim, Y.J.: Assessment and learning of qualitative physics in Newton's playground. J. Educ. Res. **106**, 423–430 (2013)

Shute, V., Rahimi, S., Smith, G., et al.: Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. J. Comput. Assist. Learn.comput. Assist. Learn **37**, 127–141 (2021a). https://doi.org/10.1111/jcal.12473

Shute, V.J., Smith, G., Kuba, R., et al.: The design, development, and testing of learning supports for the physics playground game. Int. J. Artif. Intell. Educ.artif. Intell. Educ. **31**, 357–379 (2021b). https://doi.org/10.1007/s40593-020-00196-1

Sinha, T., Cassell. J. (2015) 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

Smagorinsky, P.: Deconflating the ZPD and instructional scaffolding: retranslating and reconceiving the zone of proximal development as the zone of next development. Learn. Cult. Soc. Interact. **16**, 70–75 (2018)

Southwell, R., Pugh, S., Perkoff, M., et al (2022) Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. In: Proceedings of the 15th International Educational Data Mining Conference (EDM 22)

Steiner, I.D.: Group processes and group productivity. Academic, New York (1972)

Stephen, T.P., von Alina, A., D, Kurt P,: Computational psychometrics for the measurement of collaborative problem solving skills. Front. Psychol. **8**, 2029 (2017). https://doi.org/10.3389/fpsyg.2017.02029

Stewart, A., Vrzakova, H., Sun, C., et al.: I say, you say, we say: using spoken language to model sociocognitive processes during computer-supported collaborative problem solving. Proc. ACM Hum. Comput. Interact. **3**(39), 1–19 (2019)

Stewart, A.E.B., Keirn, Z., D'Mello, S.K.: Multimodal modeling of collaborative problem-solving facets in triads. User Model User-Adapt Interact **31**, 713–751 (2021)

Stewart, AEB., D'Mello, SK. (2018) Connecting the dots towards collaborative AIED: linking group makeup to process to learning. In: international conference on artificial intelligence in education. pp 545–556

Stewart, AEB., Keirn, ZA., D'Mello, SK. (2018) 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. ACM, New York, NY, USA, pp 21–30

Stewart, A., Rao, A., Michaels, A., et al (2023) CPSCoach: The design and implementation of intelligent collaborative problem solving feedback. In: Proceedings of the 24th international conference on artificial intelligence in education (AIED 2023), pp 695–700

Stoeffler, K., Rosen, Y., Bolsinova, M., von, DA. (2018) Gamified Assessment of Collaborative Skills with Chatbots. In: Book cover Book cover International Conference on Artificial Intelligence in Education. pp 343–347

Stuart, EA., Rubin, DB. (2008) Best practices in quasi-experimental designs. Best practices in quantitative methods 155–176

Subburaj, SK., Stewart, AEB., Rao, AR., D'Mello, SK. (2020) Multimodal, Multiparty Modeling of Collaborative Problem Solving Performance. In: Proceedings of the 2020 Conference on Multimodal Interaction

Sun, C., Shute, V., Stewart, A., et al.: Toward a generalized competency model of collaborative problem solving. Comput. Educ.. Educ. **143**, 103672 (2020)

Sun, C., Shute, V.J., Stewart, A.E.B., et al.: The relationship between collaborative problem solving processes and objective outcomes in a game-based learning environment. Comput Human Behav **128**, 107120 (2022)

Swiecki, Z., Ruis, A.R., Farrell, C., Shaffer, D.W.: Assessing individual contributions to Collaborative Problem Solving: A network analysis approach. Comput Human Behav **104**, 105876 (2020). https://doi.org/10.1016/j.chb.2019.01.009

Tian, S., Zhang, A.X., Karger, D.: A System for Interleaving Discussion and Summarization in Online Collaboration. Proc. ACM Hum-Comput. Interact. (2021). https://doi.org/10.1145/3432940

Vaswani, A., Shazeer, N., Parmar, N., et al (2017) Attention is all you need. In: Advances in neural information processing systems. pp 5998–6008

Virtaneva, M., Feshchenko, P., Hossain, A., et al (2021) COVID-19 remote work: Body stress, self-efficacy, teamwork, and perceived productivity of knowledge workers. In: Scandinavian Conference on Information Systems. Association for Information Systems

Vrzakova, H., Amon, MJ., Stewart, A., D'Mello, SK. (2019) Dynamics of visual attention in multiparty collaborative problem solving using multidimensional recurrence quantification analysis. In: proceedings of the ACM CHI conference on human factors in computing systems (CHI 2019). ACM, New York

Vrzakova, H., Amon, MJ., Rees, M. et al (2020) Looking for a Deal? Social Visual Attention during Negotiations via Mixed Media Videoconferencing (in press). In: Proceedings of the ACM: Human Computer Interaction, Computer Supported Collaborative Work (CSCW 2020)

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

Wolf, T., Debut, L., Sanh, V., et al (2019) HuggingFace's Transformers: State-of-the-art Natural Language Processing

Zhou, G., Moulder, R., Sun, C., D'Mello, SK. (2022) Investigating temporal dynamics underlying successful collaborative problem solving behaviors with multilevel vector autoregression. In: Proceedings of the 15th international educational data mining conference (EDM 22).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Sidney K. D'Mello** (PhD in Computer Science) is a Professor in the Institute of Cognitive Science and Department of Computer Science at the University of Colorado Boulder. He conducts research on the dynamic interplay between cognition and emotion while individuals and groups engage in complex real-world tasks.

**Nicholas Duran** (PhD in Cognitive Psychology) is an Associate Professor in the School of Social and Behavioral Sciences. His work revolves around behavioral dynamics and cognition as embedded and extended to complex environments.

**Amanda Michaels** (MS in Cognition & Neuroscience) is a learning specialist at The University of Chicago Pritzker School of Medicine.

**Angela E. B. Stewart** (PhD in Computer Science) is an Assistant Professor in the School of Computing and Information at the University of Pittsburgh. Angela partners with learners, teachers, and community organizations to design technology-enabled STEM learning experiences.