Computers in Human Behavior 128 (2022) 107120

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment

Chen Sun [a,*], Valerie J. Shute [b], Angela E.B. Stewart [e], Quinton Beck-White [d], Caroline R. Reinhardt [c], Guojing Zhou [c], Nicholas Duran [d], Sidney K. D'Mello [c]

[a] Johns Hopkins University, Baltimore, MD, 21218, USA
[b] Florida State University, Tallahassee, FL, 32306, USA
[c] University of Colorado Boulder, Boulder, CO, 80309, USA
[d] Arizona State University, Phoenix, AZ, 85069, USA
[e] Carnegie Mellon University, Pittsburgh, PA, 15213, USA

ARTICLE INFO

Keywords:
Collaborative problem solving
Game-based learning
Triads
Human-human interaction

ABSTRACT

Collaborative problem solving (CPS) is an essential skill for the 21st century workforce but remains difficult to assess. Understanding how CPS skills affect CPS performance outcomes can inform CPS training, task design, feedback design, and automated assessment. We investigated CPS behaviors (individually and in co-occurring patterns) in 101 ($N = 303$) remote triads who collaboratively played an educational game called *Physics Playground* for 45-min. Team interactions consisted of open-ended speech occurring over videoconferencing with screen sharing. We coded participant's utterances relative to a CPS framework consisting of three facets (i.e., competencies such as constructing shared knowledge) manifested in 19 specific indicators (e.g., responds to others' questions/ideas). A matching technique was used to isolate the effect of CPS behaviors on CPS outcomes (quality of solution of a game level) controlling for pertinent covariates. Mixed-effects ordinal regression models indicated that proposing solution ideas and discussing results were the major predictors of CPS performance, and that team-member activities surrounding idea generation mattered. These findings highlighted the importance of both individual and collective contributions and social and cognitive skills in successful CPS outcomes.

1. Introduction

From a trio of classmates working on a project, to a crew of firefighters containing a forest fire, to a medical team battling a novel global pandemic, collaborative problem solving (CPS) is part-and-parcel of our everyday experience. CPS generally refers to a situation where two or more people pool their knowledge and skills to solve complex problems without predefined solutions. The set of actions and interactions that occur during the problem-solving effort can be indicative of CPS skills, which include both collaboration and problem-solving skills (see Hesse, Care, Buder, Sassenberg, & Griffin, 2015; OECD, 2017).

CPS skills have increasingly been viewed as essential in many contexts, such as in schools (e.g., OECD, 2017; Scoular & Care, 2020), informal learning settings (e.g., Huang et al., 2018), online learning (e.g., Rosen, Wolf, & Stoeffler, 2020), military settings (e.g., Swiecki, Ruis, Farrell, & Shaffer, 2020), business services (Aarikka-Stenroos & Jaakkola, 2012), and marketing innovations (Heirati & Siahtiri, 2019) to name a few. Indeed, the ever-growing importance of CPS skills in today's interconnected world is acknowledged by multiple frameworks of 21st century skills (e.g., Andrews-Todd & Forsyth, 2020; OECD, 2017). However, a fundamental question needed for understanding, assessing, and training CPS skills remains unanswered: What particular CPS behaviors give rise to successful problem-solving outcomes?

We address this foundational question by identifying: (1) basic and essential *individual* CPS behaviors, and (2) key *interactive patterns* among triads that contribute to successful problem-solving outcomes. Our study makes three novel contributions to the literature. First, we explored collaboration among triads. This goes beyond current research which mainly examines students collaborating in dyads, resulting in precious little knowledge about how larger groups interact with each other and form an effective team. Researchers have argued that dyads have too few degrees of freedom to reflect the complexity of group behaviors (Moreland, 2010; Reiter-Palmon, Sinha, Gevers, Odobez, & Volpe, 2017). For example, triads have seven degrees of freedom (i.e., three

* Corresponding author. McAuley Hall 400, 5801 Smith Avenue, Baltimore, MD, 21209, USA.
E-mail address: csun44@jhu.edu (C. Sun).

https://doi.org/10.1016/j.chb.2021.107120
Received 28 January 2021; Received in revised form 21 November 2021; Accepted 25 November 2021
Available online 27 November 2021

individuals, three possible dyads, and one triad) compared to three degrees of freedom with dyads. Second, we tackled the complexity of natural spoken communication among the triads instead of asking them to interact with a pre-programmed computer agent (e.g., Stoeffler, Rosen, Bolsinova, & von Davier, 2020) or type in a chat box (e.g., Scoular & Care, 2020). The chat box is convenient for documenting all dialogue in log files, but the interruption of communication required by typing potentially hinders the understanding of team dynamics in real-life scenarios. Conversely, natural discourse allows for a wide range of opportunities for communication patterns to occur. Thus, the current literature does not provide us with a deep understanding of natural communication relative to CPS outcomes. Third, we examined CPS while people collaborated while playing an engaging digital game, which mirrors a real-life CPS scenario (i.e., friends hanging out and playing games together). Currently, many existing CPS tasks are situated in simulation-based learning tasks (e.g., Andrews-Todd & Forsyth, 2020; Hao, Liu, Kyllonen, Flor, & von Davier, 2019), which may be less engaging than digital games.

1.1. Status quo of collaborative problem solving assessments

In line with its expanding importance as an essential skill for people to master in the 21st century, researchers have been developing various theoretical models and frameworks specifying CPS skills, associated behaviors, and assessments (e.g., Andrews-Todd & Forsyth, 2020; Hesse et al., 2015; OECD, 2017; Sun et al., 2020; von Davier, Hao, Liu, & Kyllonen, 2017). For instance, in 2015, the Programme for International Student Assessment (PISA) conducted an international assessment of CPS among 15-year-old students in over 70 countries and regions. In this large-scale CPS assessment, students were tasked with interacting with a computer agent (not another human) by choosing appropriate responses from pre-defined response options. Students' CPS skills were then evaluated based on the quality of their submitted choices. Critics of the PISA assessment have focused on the constrained nature of human-agent interactions, which does not represent naturalistic communication among humans as the responses from the agent are pre-programmed and the human responses are limited to the provided choices (Graesser, Greiff, Stadler, & Shubeck, 2020; Herborn, Stadler, Mustafić, & Greiff, 2020). However, for a large-scale standardized assessment of CPS, interacting with computer agents is a feasible, consistent, and pragmatic way to measure students' CPS skills (Hao et al., 2019).

Beyond standardized assessments, in real-life scenarios, CPS requires complex open-ended human-human interactions where people draw on each other's knowledge, skills, and other abilities to solve problems (Care et al., 2017). Consequently, there has recently been an increase in research on how to accurately assess CPS skills in authentic human-human interactive environments. For example, recent assessments have attempted to capture naturalistic dyadic interactions within various CPS tasks (e.g., Ostrander et al., 2020; Scoular & Care, 2020). However, much less research has explored how to accurately assess naturalistic human-human CPS interactions in triads (e.g., Andrews-Todd & Forsyth, 2020; Sun et al., 2020) as we do in the present study. In addition, researchers have argued that dyads and triads represent qualitatively different types of collaboration. For example, triads can generate more complex group phenomena than dyads, such as formation of a coalition (Moreland, 2010). Also, communication and negotiation are more complicated in triads than with dyads (Reiter-Palmon et al., 2017) because of the added degrees of freedom where the referent of a communication needs to be made more explicit. Because many CPS real-world tasks entail groups larger than dyads, investigating triads is itself a worthy endeavor.

Beyond group size, the communicative medium is also an important factor to consider. Current CPS research mainly takes two approaches to analyzing naturalistic human-human CPS interactions. One approach used in computer-supported CPS environments is to ask people to type in their comments/responses into a chat box, which are automatically logged by the computer (e.g., Andrews-Todd & Forsyth, 2020; Scoular & Care, 2020). The other approach is to audio/video record the CPS interaction which might unfold either face-to-face or remotely (e.g., Sun et al., 2020; Swiecki et al., 2020). Compared with chat box conversations, the recordings produce much richer and more abundant streams of naturalistic communication data. Typing also tends to be more deliberate than spontaneous speech (D'Mello, Dowell, & Graesser, 2011). In our study, we adopted the latter approach where we video recorded participants' open-ended spoken interactions while they engaged in CPS tasks. We also opted for remote collaborations, which have increased tremendously in education and in the workplace (Dowell, Lin, Godfrey, & Brooks, 2020; Schulze & Krumm, 2017). The demand for remote collaborations is especially prominent, given current events, as with the COVID-19 pandemic. However, remote collaborations have additional challenges, including technological limitations (e.g., low bandwidth), loss of some non-verbal information, cultural differences, and even time differences (Schulze and Krumm, 2017; Vrzakova, Amon, Rees, Faber, & D'Mello, 2020). Thus, investigating remote human-human collaboration is an additional contribution to the CPS literature.

1.2. CPS analysis frameworks

Several frameworks reported in the literature have been used to analyze CPS behaviors in learning contexts. Table 1 summarizes the categorization of CPS skills by three main frameworks. All three frameworks consistently view CPS as two separate skill sets: cognitive and social. The three frameworks also show commonalities in essential CPS behaviors, such as establishing shared understanding, negotiation, and carrying out solution plans.

One established CPS framework is the Assessment and Teaching of 21st century Skills (ATC21S), by Hesse et al. (2015), which explicates the individual (rather than collective) contributions of the social and cognitive skills to the problem-solving space (Scoular, Care, & Hesse, 2017). Similarly, the well-known PISA framework of CPS (OECD, 2017) dissects the CPS construct into the interplay of collaboration (social) and problem-solving skills (cognitive). The three collaborative skills and the four problem-solving skills comprise a matrix of 12 CPS skills (see OECD, 2017 for details). As with the other two frameworks, the CPS ontology (Andrews-Todd & Forsyth, 2020) also categorizes the CPS construct into social and cognitive dimensions, with the social dimension focusing on collaboration and teamwork, and the cognitive on problem-solving processes. This ontology included relevant data that can be collected from chat messages and logged events (e.g., students modifying parameter input) during computer-supported CPS tasks.

Table 1
Summary of three main CPS frameworks.

	PISA CPS Framework (OECD, 2017)	ATC21S CPS Framework (Hesse et al., 2015)	CPS Ontology (Andrews-Todd & Forsyth, 2020)
Social Skills	Establishing and maintaining shared understanding	Participation	Maintaining communication
	Taking action to solve problems	Perspective taking	Sharing information
	Establishing and maintaining team organization	Social regulation	Establishing shared understanding
			Negotiating
Cognitive Skills	Exploring and understanding	Task regulation	Exploring and understanding
	Representing and formulating	Knowledge building	Representing and formulating
	Planning and executing		Planning
	Monitoring and reflecting		Executing
			Monitoring

2. Processes and outcomes in collaborative problem solving

Researchers have been assessing people's CPS skills from two angles: CPS processes and subsequent outcomes. When people are engaged in CPS tasks, they generate CPS process data, which refers to sequential events (e.g., actions and utterances) related to dynamic interactions (von Davier et al., 2017). CPS outcomes are generally measured by the correctness of responses to a question or the overall success of problem solving (Hao et al., 2019; von Davier et al., 2017). CPS outcomes can also be subjective, for example, peoples' perceptions of the collaboration process and their teammates (e.g., Meier, Spada, & Rummel, 2007).

The paths connecting CPS processes to problem-solving outcomes are many. A team can demonstrate appropriate CPS behaviors but still have an unsuccessful outcome, especially if the problem is too difficult, the team lacks sufficient knowledge to solve the problem, or pursues an incorrect strategy (e.g., Hmelo, Nagarajan, & Day, 2000; Hmelo-Silver, 2003). Alternatively, a team may be successful at solving the problem, yet demonstrate poor CPS behaviors, like when a dominant member of the team solves the problem without input from others, or there is considerable conflict within the team (Rosen et al., 2020). In short, because CPS processes involve both cognitive and social behaviors, strengths in one but not the other can lead to varying outcomes depending on how the outcome is defined.

As there are multiple ways that CPS processes can contribute to successful outcomes, much depends on the ways that CPS behaviors and outcomes are operationalized. For instance, Hao, Liu, von Davier, Kyllonen, and Kitchen (2016) operationalized CPS into four general skills: sharing ideas, negotiation, regulating problem solving, and maintaining communication. They analyzed dyadic typed dialogues by categorizing each turn of the conversation into one of the four skills, and found that successful teams demonstrated significantly greater negotiation skills compared with unsuccessful teams. The other three skills did not predict the CPS outcome.

Some researchers have attempted to use linguistic features to interpret CPS communication. For example, Reilly and Schneider (2019) used linguistic features (e.g., length of sentences and part of speech) to predict collaboration and learning when dyads interacted face-to-face. The length of utterances positively correlated with collaboration quality. Using domain-specific words and clear references in communication correlated with learning. One drawback of the study relates to the accurate interpretation of the content of communications. Recently, taking semantic meanings into account, Dowell et al. (2020) applied a computational linguistic analysis to analyze dyadic text-based communication in a CPS simulation task on volcano eruption. They analyzed the text in terms of participation, social impact, overall responsivity, newness, internal cohesion, and communication density. They identified five emergent roles adopted by participants during collaboration (i.e., influential actors, drivers, followers, lurkers, and socially detached learners). They found that socially active roles (i.e., influential actors and drivers) helped the team to obtain better outcomes than those with socially disengaged roles (e.g., socially detached learners).

Forsyth, Andrews-Todd, and Steinberg (2020) examined fine-grained CPS behaviors consisting of 23 CPS subskills. They used cluster analyses on those coded CPS behaviors to identify four types of collaborators (i.e., active collaborators, super socials, low collaborators, and social loafers) based on interactions in a computer-mediated CPS environment with typed chat among triads. They found significant correlations between collaborator type with various measures (e.g., number of levels attempted and self-reported CPS skills). In the analysis, the researchers did not address confounding variables such as verbosity, so it is unclear if the clusters contributed additional information beyond the volume of content. Similarly, Chang et al. (2017) analyzed CPS patterns, based on the PISA framework, in a simulation task with typed chat. With a small sample size of 10 triads, they identified four groups that successfully solved simulation tasks, and six groups that did not solve the simulation tasks. A lag sequential analysis indicated that the unsuccessful groups repeatedly and unsystematically tested different values in the simulation task and failed to come up with executable solutions from their discussions. Meanwhile, successful teams demonstrated analytical and reasoning strategies where they iteratively shared understanding, executed possible solutions, and monitored results.

In conclusion, current studies have mainly focus on typed chat and dyadic CPS, which motivated us to investigate natural dialogues generated by triadic interactions. Although these studies have provided some initial insights, they have not provided a clear picture and consistent findings regarding the link between CPS processes and task performance. Further research is needed to understand the complexity and dynamics of triadic CPS interaction with respect to problem solving performance, especially when triads are communicating freely and verbally in open-ended problem-solving environments. Therefore, our study contributes to the understanding of triadic CPS behaviors in relation to outcomes by analyzing CPS communications at a fine-grained size.

2.1. Current study

The goal of our study was to shed light on the relationship between CPS behaviors and problem-solving success as triads interacted remotely (i.e., via videoconferencing) to collaboratively solve problems within a computer-based educational game called Physics Playground (Shute, Almond, & Rahimi, 2019). In our CPS task, all members of a triad interacted with each other using spoken language rather than using a chat box to communicate. Such natural triadic interactions posed some challenges regarding understanding and interpreting the inter-connected actions. Accordingly, we adopted and updated a generalized CPS model developed to analyze peoples' behaviors in CPS tasks (Sun et al., 2020) and trained human raters to code a subset of utterances generated by the triads as they engaged in CPS tasks. And despite being quite labor intensive, human ratings can help to ensure the accuracy of classifying utterances into relevant CPS behaviors (indicators in our case), which is important before proceeding to automated modeling of triadic interactions.

Our analyses examined specific CPS indicators that were related to team outcomes (coin earned for solving a game level in our case) after accounting for baseline communication context (e.g., length of time, verbosity). Further, inspired by research investigating communicative patterns, such as epistemic network analysis (Csanadi, Eagan, Kollar, Shaffer, & Fischer, 2018) and group communication analysis (Dowell, Nixon, & Graesser, 2019), we conducted a preliminary analysis to identify frequently occurring discourse patterns (i.e., co-occurring indicators) that predicted CPS outcomes beyond the individual indicators. In sum, our work contributes to understanding how CPS behaviors – at both the indicator and pattern level – predict successful CPS outcomes in an open-ended, spoken, triadic, computer-mediated, remote CPS environment.

3. Method

The data were collected as part of a larger study on collaborative problem solving (Eloy et al., 2019), but only the details pertinent to the present study are reported here. The primary data source reported here consisted of CPS codes of verbal utterances; these data have not been previously analyzed or published elsewhere.

3.1. Participants

Participants were 303 undergraduates (56% female, average age = 22 years) from two large public universities (39% from University 1). Participants self-reported the following race/ethnicities: 47% Caucasian, 28% Hispanic/Latino, 18% Asian, 2% Black or African American, 1% American Indian or Alaska Native, and 4% "other." Participants were assigned to 101 triads based on scheduling constraints. Thirty

participants from 18 teams (26%) indicated they knew at least one person from their team prior to participation. Participants were compensated with a $50 Amazon gift card (96%) or course credit (4%) at the end of the study.

3.2. CPS task

We used a digital game, *Physics Playground* (Shute et al., 2019) intended to help young adults learn Newtonian physics (e.g., Newton's laws of force and motion). The overarching goal of this game is to direct a green ball to hit a red balloon. To solve game levels, participants need to draw appropriate simple machines (e.g., lever and springboard) on the screen using a mouse (Fig. 1). The simple machines come "alive" on the screen after being drawn as they obey the laws of physics (as does everything in the game). Players receive a gold coin when they solve the level efficiently (i.e., with minimal objects), and they receive a silver coin for a less efficient solution using more objects. No coin is rewarded for unsolved levels.

We selected 17 game levels covering two physics concepts: nine levels related to "energy can transfer" (EcT), and eight levels related to "properties of torque" (PoT). Example subconcepts include kinetic energy, gravitational potential energy, angular acceleration, and angular momentum. The 17 levels varied in terms of difficulty (as rated by two physics experts). The levels were organized within three "playgrounds" (detailed below). Players could freely navigate through levels in their current playground and pick which levels to solve. They could also restart a particular level within their playground as many times as they liked as well as quit a level at will. They could also revisit tutorials illustrating the game mechanics at any time, but no additional hints were provided.

3.3. Procedure

There was an at-home and in-lab portion to the experiment. Participants were emailed a Qualtrics survey with an embedded short tutorial on how to use *Physics Playground* at least 24 h prior to participating in the lab session. Participants needed to complete a pretest on their knowledge of the targeted physics concepts—energy can transfer and properties of torque. The pretest had two parallel forms (form A and form B) of ten items created by physics experts. In the tutorial, participants were instructed on the object of the game, as well as how to draw simple machines. After completing the tutorial, participants were given 15 min to complete five easy levels to familiarize themselves with the game. Participants also completed a battery of individual difference measures, not analyzed here.

Participants were scheduled in groups of three based on availability. Upon arrival in the lab, participants were individually assigned to one of three computer-enabled workstations equipped with a webcam and headset microphone, either partitioned in different corners of the same room or located in different rooms, depending on the University where the data were collected. All collaborations occurred via Zoom video-conferencing software irrespective of the layout since the goal was to study remote collaborations. There were additional sensors at each workstation which are not germane to our current analyses. Zoom recordings of all collaborations were retained for analysis.

Teams (consisting of three participants) collaboratively solved levels in *Physics Playground* across three 15-min blocks, totaling 45-min of collaborative gameplay. During each block, one teammate was randomly assigned the role of controller and the other two were contributors. The controller was in charge of all mouse interactions with *Physics Playground* (Fig. 2). The controller's screen was shared via Zoom, such that the contributors could view gameplay and contribute to the solution of the level. The role of controller randomly rotated so each teammate served as controller for one block. There was a fourth block of a separate collaborative task not analyzed here.

The first block served as a warmup. Participants were instructed verbally and with on-screen instructions to use the time to familiarize themselves with their teammates and play a few levels together. During this time, they were given five easy-to-medium levels in a playground involving a mix of EcT and PoT concepts. Teams then completed two 15-min experimental blocks, where each block was assigned a different CPS goal (this was an experimental manipulation for another purpose). In one goal manipulation, teams were instructed to "solve as many levels as possible." The purpose of this was to prioritize the quantity of levels solved. In the other manipulation, teams were instructed to "get as many gold coins as possible." In that case, the purpose was to focus teams on solution quality. Teams were reminded that gold coins are earned by using fewer objects in their solutions indicating more elegant solutions. Instructions for the experimental block were provided verbally and on screen.

We also manipulated the physics concept. Teams were either presented with seven EcT levels or six PoT levels in separate playgrounds (all were rated as medium to hard difficulty) within the two experimental blocks. The particular goal and physics concept were counterbalanced across teams in a 2 × 2 (goal × concept) within-subject design.

Across all three blocks, teams received on-screen warnings when they had ten and 5 min left in the block. They were also reminded of their assigned goal (levels or gold coins) along with the warning.

3.4. Measures

In-game performance. There were three possible outcomes for each level attempt in the game: (1) the team did *not* solve the level within the allotted time or they quit the level, and thus no coin was awarded; (2) they solved the level using a limited number of objects (i.e., efficient/creative solution) and received a gold coin; or (3) they solved the level

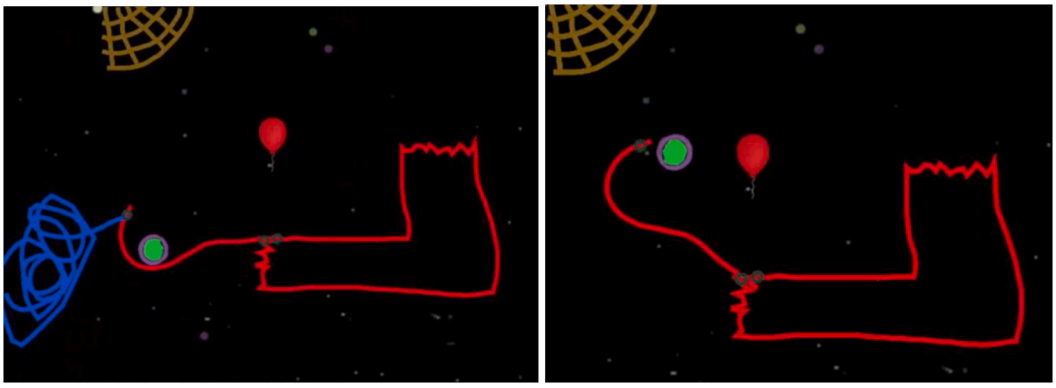

Fig. 1. Spider web: A level involving the physics concept of energy can transfer a springboard drawn with a weight at the end (Left); The ball shooting for the balloon when the weight was released (Right).

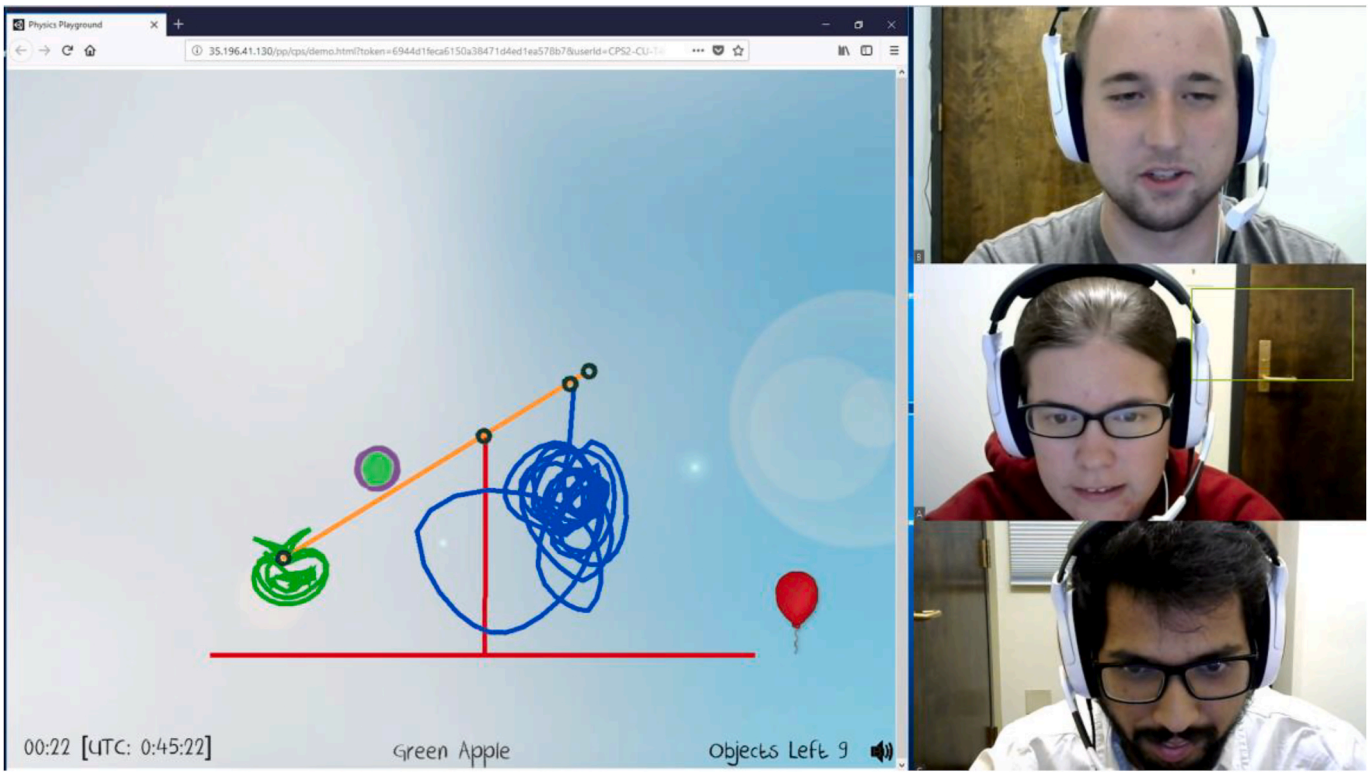

Fig. 2. A triad working collaboratively to solve a level using a lever and weights.

with a less efficient solution and received a silver coin. Earning either a silver or gold coin was considered a successful level attempt.

3.5. Coding CPS behaviors

Our study focused on coding CPS behaviors from free verbal communication while collaboratively playing the game. We used a validated CPS framework and a level matching scheme to code a subset of the in-game data as elaborated below.

Coding Framework. To measure participants' CPS behaviors, we adapted and revised a validated CPS framework suitable for coding open-ended trialogues as in the present study (Sun et al., 2020). The model was derived from existing CPS frameworks (see the CPS frameworks section above) and consists of three main CPS facets: constructing shared knowledge, negotiation/coordination, and maintaining team function. Constructing shared knowledge refers to (a) disseminating knowledge, ideas, and resources among team members, and (b) establishing common ground for understanding the task and solutions, both of which have been emphasized in the literature (Andrews-Todd & Forsyth, 2020; OECD, 2017; Roschelle & Teasley, 1995). Negotiation and coordination relate to the processes involved with reaching a consensus on a solution plan to be carried out. This includes the dividing of labor, resolving conflicts, integrating different perspectives, and monitoring execution (Andrews-Todd & Forsyth, 2020; Hesse et al., 2015; Rummel & Spada, 2005). The third facet emphasizes efforts to maintain a functional team via assuming individual responsibilities, taking initiative, and co-regulation (Care, Scoular, & Griffin, 2016; Hesse et al., 2015; Rosen, 2017).

In the current paper, we slightly refined the earlier model regarding particular indicators that are associated with each facet (see Table 2, for details). The model, then and now, viewed the social and cognitive aspects as closely intertwined, and aimed to analyze CPS skills of individuals playing digital games together. There are 19 indicators aligned to the various CPS facets (Table 2). Most (68%) of the indicators are positive, suggesting desirable behaviors (and higher CPS skills), whereas others are negative (denoted by "R" implying they are reverse coded). Based on previous data using this model and task (Sun et al., 2020), we focused solely on *verbal* indicators because nonverbal indicators (e.g., visibly not focused on the task) rarely occurred.

We refined certain indicators to reflect the quality of problem solving. Specifically, we divided the previous indicator (*suggests potential ideas*) into (a) *suggests appropriate ideas,* and (b) *suggests inappropriate* ideas. Appropriate ideas refer to proposed suggestions that are relevant for the given the circumstances in a level although they are not guaranteed to succeed. For example, a student may suggest drawing a heavier weight if the current weight in Fig. 1 did not launch the ball close enough to the balloon. Alternatively, a student may suggest using a lever to launch the ball, which is also appropriate to solve the level. Thus, appropriate ideas reflect students' current understanding of the problem and their ability to help others build knowledge, whereas inappropriate ideas reflect lack of knowledge in tackling the current situation which can potentially mislead team members into pursuing a futile path. Importantly, the indicators now capture the *appropriateness* of potential ideas – not implemented attempts, but again, do not guarantee a successful solution. As such, they do not encode problem solving success. We include comparison models to investigate whether the mere quantity (not differentiating among appropriate/inappropriate ideas) is sufficient for predicting CPS outcomes.

In addition to the two new indicators described above (i.e., suggests appropriate vs. inappropriate ideas), we added four more indicators to the new CPS framework: *questions/corrects others' mistakes*, *strategizes to achieve task goals*, *tries to quickly save almost successful attempts*, and *apologizes for one's mistakes*, since we observed their occurrence during preliminary analyses of the data. For detailed descriptions and example utterances per indicator, see Table 2.

Coding procedure. To code the utterances generated during gameplay, we used IBM Watson—speech recognition software—to segment each participants' audio stream into individual utterances. We then merged utterances spoken by the same speaker within 2 s to address segmentation errors, and identified duplicates in the transcript arising

Table 2
Coding scheme-descriptions and example utterances for each indicator.

Indicators/Facets	Description and Coding Notes
Constructing shared knowledge	
1. Talks about the challenge situation	● Talks about the challenge/game environment (e.g., *"What does that do?"; "Where is the start?"*) ● Talks about the challenge/game mechanics (e.g., *"How do I delete this?"; "How do I restart the level?"*) ● Talks about the challenge criteria (e.g., *"We need to get gold coins"*; *"Use as few objects as possible"*) ● Talks about something that's already on the screen when the player enters a level (e.g., *"What's that?", "Is that a spider?", "Can we delete that?"*) ● Talks about time (e.g., *"10 min left"*) ● Talks about computer error, program glitches (e.g., *"it's lagging", "it's not letting me draw XX."*)
2. Suggests appropriate ideas	● Proposes appropriate ideas to solve the level (e.g., *"Try to make a weight attached"*) ● Proposes appropriate ways to fix a failed solution (e.g., *"Make it shorter", "This didn't work because …"*) ● Appropriate means that the idea is consistent with the underlying Physics in game context.
3. Suggests inappropriate ideas	● Proposes inappropriate ideas to solve the level (e.g., ideas do not make physics sense) ● Proposes inappropriate ways to fix a failed solution (e.g., suggests lowering pendulum arm when height should be increased).
4. Confirms understanding	● Asks questions for clarification (e.g., *"What?", "Is this what you are asking?"," What's next?"*) ● Reiterates or paraphrases another person's idea (e.g., *"Do you want me to …", "Ok, make it heavier"*) ● Should occur after the proposal of a solution.
5. Interrupts others (R)	● Anytime a person is in the middle of speaking and another person jumps in. ● Does not count if two people start talking at the same time.
Negotiation & Coordination	
6. Provides reasons to support a solution	● Reasons should be substantial and offer clear logic (e.g., "*Hopefully, it will spring upwards and hit the balloon*".) ● Pay attention to signaling words *"because", "and then it will …" Do not code: "Cuz, ya know."; "Because, yeah"*
7. Questions/Corrects others' mistakes	● Tries to point out and/or correct the mistakes in others' ideas/solutions (e.g., "*I think this would get stuck on the green line*", "*Wouldn't it hit the wall*?"). ● Code if the player draws incorrectly, someone proposes a solution to correct their mistake.
8. Responds to other's questions/ideas	● Responds to another's ideas/questions (e.g., *"That's what I was thinking", "No I don't agree"*) ● Responds to a Yes/No question (e.g., *"yes", "no", "not sure", "I don't know"*) ● Responds to "*what/which do you think*?" or similar questions ● Code if the answer is simply, "*up to you", "I don't mind", etc.* If the answer is an elaboration, code it with respect to other appropriate indicators. ● If the answer is *"you can try it", "go ahead and try it"*, code it as "compliments or encourages others".
9. Criticizes, makes fun of, or being rude to other (R)	● Makes disparaging or rude remarks about other player's or their ideas.
10. Discusses the results	● Provides substantial and specific comments about the results (e.g., *"The ball fell off the screen", "It stuck on the pink thing"*). ● Do not code general comments: *"What happened", "oh no", "that worked", "we are close"*.
11. Brings up giving up the challenge (R)	● Identifies the cause of a failed solution and reflect on what has been done. (e.g., *"The line was too short", "Not heavy enough"*). ● Talks about quitting or moving to a different/easier level (e.g., *"Can we try another level?"*) ● Do not count if a person simply talks about the level being difficult or hard. They must bring up quitting.
12. Strategizes to accomplish task goals	● Explicitly states that choosing a different/easier level to achieve task goals (golds or levels) (e.g., *"We only solved one level. Should we move to an easier level?", "Want to go to that level to get a gold?"*) ● Suggests redoing a level to achieve a gold (e.g., *"But we only got silver. How can we get gold?"*) ● Suggests using fewer objects (e.g., *"Restart. You have drawn so many things"*)
13. Tries to quickly save an almost successful attempt	● When the ball almost touches the balloon, uses quick remedy solutions (e.g., *"Click the ball, click the ball!"*)
Maintaining team function	
14. Asks others for suggestions	● Asks others for possible ideas to facilitate collaboration. (Those are general questions for others to state their ideas/solutions, usually at the beginning of a level or when they are stuck) (e.g., *"What do you think?", "How do we do that?", "I don't know what to do"*.) ● Asks the group to choose between two previously discussed ideas.
15. Compliments or encourages others	● Shows support for one another's ideas/solutions (e.g., "Let's try it and see", "That's a good idea", "perfect!", "Yay, great job!", "Yay we did/made it", "we are almost there") ● Encourages others after a solution is implemented (e.g., "Aww, we are so close!" "Ah, almost (there)!") ● Empathizes with others (e.g., "yeah, it's hard to draw")
16. Initiates off-topic conversation (R)	● Talks about anything unrelated to the task at hand or the challenge environment (e.g., "Is it cold in here?", "I'm so tired", "have you ever played XX game?")
17. Joins in off-topic conversation (R)	● Engages with another person's off-topic conversation. ● Simply acknowledging the other person ("Yeah", "uh-huh") doesn't count.
18. Provides instructional support	● Provides instructions to the controlling player on how to implement a solution (e.g., "Start drawing here", "You'll make a hook shape") ● Code for each individual instructional step. ● Double code if a player provides instructions and proposes a new solution at the same time. ● Do not code utterances like "just make a hook".
19. Apologizes for one's mistakes	● Apologizes after a suggested solution failed (e.g., "My bad. It didn't work.", "Oops, I missed it") ● Apologizes after accidently interrupting others (e.g., "Sorry, go ahead", "sorry for interrupting. What were you saying?") ● Apologizes for bad drawing (e.g., "Sorry, my drawing is bad", "sorry, I'm too clumsy")

from audio interference. The utterances from the three speakers were then merged into a cohesive transcript based on timestamps. Although we used speech recognition software to generate the transcripts as it is less resource intensive, the coders had access to the full audio and screen-capture videos to verify transcribed utterances throughout the process.

The coders rated each of the utterances in terms of whether the specific utterance contained evidence of any of the indicators and coded the frequency of occurrence for each indicator. The coders viewed the video recordings (with included audio as in Fig. 2) of gameplay while

coding to understand the context of the utterance as well as the group dynamics. Coders did not know the level attempt results—i.e., solved or unsolved—until they viewed the end of the video.

Table 3 below provides a sample exchange among participants along with codes (assigned indicators), and Fig. 3 shows a screenshot of the game level that generated the sample exchange. In this situation, the three students just started the level and were discussing how to solve it. As shown in Table 3, one student (PA) pointed out that timing was a critical element to solving the level, so that was coded as suggesting an appropriate idea although much more had to occur to yield a solution. Another student (PC) added onto PA's idea and emphasized that they needed to drop a weight to hit the lever at the right time. All the players started to engage in a conversation to come up with a solution plan. Then they drew objects on the screen to execute their solution plan, as shown in the screenshot to the right in Fig. 3. After about 7 min, they could not solve this level so they quit the level and moved on to another one.

Training and interrater reliability. Three human coders received two rounds of training from the first author. In each round of training, the three coders coded three different level attempts randomly selected from different teams. We adopted two indices of interrater reliability: Gwet's AC1[1] and percentage agreement among the three coders. After the second round of training, Gwet's AC1 values across indicators ranged from 0.91 to 1.00 and the percentage agreement was high (0.89–1.00) for all indicators.

Next, we randomly selected nine level attempts (i.e., three level attempts for each round) that all three coders individually coded. The purpose was to ensure the quality of coding. We assessed interrater reliability after each round and the coders discussed the indicators with low (<0.90) reliability indices prior to the next round. Gwet's AC1 values for the three rounds across indicators were 0.91–1.00, 0.84–1.00, and 0.93–1.00 with corresponding percentage agreements of 0.89–1.00, 0.80–1.00, and 0.90–1.00. Because the coders maintained good interrater reliability (0.85–1) in the quality check, they proceeded with individual coding a total of 209 randomly-selected level attempts (see below). The coders followed the same coding schedule (e.g., code X number of level attempts within Y number of days).

Level matching. Because the nature of the collaborative interaction (e.g., what is spoken, game-play dynamics) is largely determined by unique game levels, we sought to compare CPS behaviors across teams who achieved different in-game performance outcomes (i.e., gold, silver, or no coin) within the same level. We used a quasi-experimental design procedure—matching—to isolate the effect of CPS behaviors on CPS outcomes after accounting for pertinent covariates detailed below. To do so, we generated matched sets of level attempts, such that each level attempt in the set had a different in-game outcome (i.e., gold, silver, or no coin).

Level attempts were segmented from the *Physics Playground* logs, which recorded when a team entered a level, earned a coin, exited without earning a coin, or reentered a level. An attempt began when the team entered a level and ended when they solved the level, began a different level, or time ran out in the block. Note that a team could enter the same level multiple times in one block. However, if another level was visited in between these visits, they were considered separate attempts. In total, we segmented 1164 level attempts (27% gold, 29% silver, and 44% no coin) from the data.

Prior to matching, we checked all the level attempt durations in seconds. Then we removed level attempts ($n = 356$, 31% of total) that were less than 60 s since these largely reflected cases where teams were investigating a level ostensibly to decide if they wanted to attempt it, resulting in 808 level attempts for matching.

Solution rates significantly differed for the two blocks with exclusive Energy can Transfer (EcT) (18% successful levels; 7% with gold trophies out of all the attempts) and Properties of Torque (PoT) (63% successful with 40% gold out of all the attempts) levels, two-tailed paired-samples, $t(88) = 12.95$, $p < .001$ (this analysis only included teams with complete log data). Because of the lower success rate for EcT, we coded its CPS performance as a binary paired outcome (i.e., coin [gold or silver] or no coin) for matching to ensure a sufficient number of matches. Success was coded as a triplet (silver, gold, no coin) for PoT and for the warmup levels (average completion rates of 34%; 10% gold).

Next, we used the *bmatch* function in the *designmatch* (Zubizarreta, Kilcioglu, & Vielma, 2018) R package to form matched triplets (i.e., gold, silver, or no coin) or pairs (i.e., coin or no coin) of level attempts. Matching was done separately for each of the three blocks and was based on the following five covariates. The four categorical covariates were school, level identifier, manipulation (i.e., gold coins vs. solve many levels), and block number (first or second) for the experimental blocks. One continuous covariate was the duration of the level attempt, and we constrained the level attempt duration (in seconds) to be at most 0.25 standard deviations of the mean duration of all the level attempts.

In total, we formed 324 level attempt matches, including 33 Warmup and 29 PoT triplets (gold, silver, no coin for both), and 69 EcT pairs (coin or no coin) from our candidate set of 808 level attempts. Human coding all the level attempts would be very time-consuming and labor-intensive. Based on available coding resources, we randomly sampled a subset of 209 matches for analysis. These included 22 Warmup triplets (i.e., 66 level attempts), 34 EcT pairs (i.e., 68 level attempts), and 25 PoT triplets (i.e., 75 level attempts). A total of 47, 49, and 54 unique teams were included in the matched pairs/triplets for warmup, EcT, and PoT, respectively (corresponding to 141, 147, and 162 participants). One EcT pair was not coded because the video recording was missing.

For the subset of 209 level attempts, we assessed the success of our matches across the pertinent covariates (Appendix A). Matches were considered successful if covariates were similar across our outcome groups (coins in this case), which was indeed the case for school, manipulation, and block, which were the same within a matched pair or triplet. The differences in duration (up to 77 s – Appendix A) were most pronounced for the warmup levels where level attempts resulting in silver coins required the least amount of time and unsuccessful attempts taking the most time. We included duration as a covariate in the subsequent models to control for these differences.

We also compared average team physics pretest scores (range from 0 to 10), which served as a proxy for prior knowledge (Appendix A). Average team pretest scores were similar across coin type with a maximum difference of less than 1 point. We did not include pretest scores as a covariate in the matching process because it is a person-level variable (matching was done at the team level) and prior work found no

Table 3
A sample exchange among participants with associated indicators.

Participant	Transcripts	Associated indicators
PA	We gotta time something.	*Suggests appropriate ideas*
PC	Yeah you gonna have to drop something at the right time. It's like …	*Confirms understanding*
PB	Yeah the exact same time.	*Confirms understanding; interrupts others*
PA	Oh, yeah, right, I guess …	*Responds to others' questions/ideas*
PB	Or you could do something like the last time. Make the left one, uh, put like a weight on it so it evens it out.	*Suggests appropriate ideas* *Provides reasons to support a solution*
PA	So like it evens it out.	*Confirms understanding*

[1] Gwet's AC1 provides consistent estimates of interrater reliability regardless of sample sizes and does not assume independence between raters. It is particularly useful in cases where agreement is high, which is a known problem for other metrics like Cohen's kappa (Gwet, 2008).

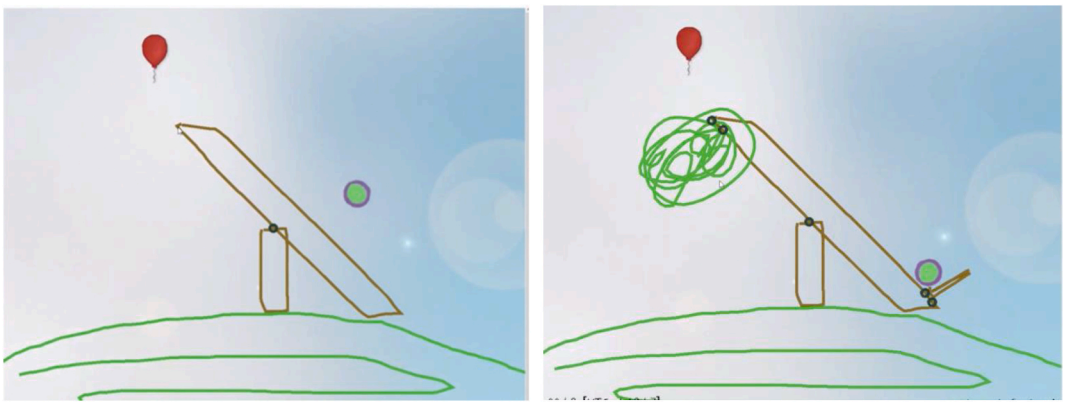

Fig. 3. A screenshot of the game level the team was trying to solve in Table 3 exchange.

relationship between pretest scores and task performance (Stewart, Amon, Duran, & D'Mello, 2020).

3.6. Analysis

Our goal was to investigate how CPS behaviors predicted in-game performance within the matched levels. In total, the three coders coded 16,446 utterances from the 209 levels, after excluding duplicated utterances (as rated by human coders). We removed the last 10% of utterances from each level attempt out of a concern that the language/indicators might have focused on the success or failure of the outcome rather than the problem-solving process. This removal resulted in 14,689 utterances available for our analyses. These data were analyzed at the indicator- and pattern-level as noted below. See Appendix B for a sample of the dataset.

Indicator-level analysis. The first analysis was an utterance-level, multilevel ordinal regression with problem solving outcome (no coin [0] vs. silver coin [1] vs. gold coin [2]) as the dependent variable and level attempt match identifier (match ID) as the grouping factor (and random intercept). Note that although EcT levels were coded as a binary outcome (coin or no-coin) for level matching (see above), we reverted to their original trichotomous codes for the models. We simultaneously included the utterance-level counts of all indicators as our predictor variables, thereby addressing the influence of each indicator relative to the others.

We also controlled for the following six variables (covariates). The first covariate was the *relative utterance index* (i.e., the relative position of the utterance within the level). This was determined by dividing the index of the current utterance by the total number of utterances within the level attempt. The position of an utterance matters. That is, utterances at the beginning of a level attempt tend to relate to figuring out the problem situation and brainstorming solutions, whereas those in the middle of problem solving tend to relate to solution implementation and refinement. The second covariate was *concept,* with two levels for EcT (set as the reference group) and PoT. We added this covariate because success rates differed between the two concepts (i.e., 19% for EcT vs. 63% for PoT). For the third covariate, *level duration*, we computed the time spent per level attempt by subtracting the start time from the end time and z-scoring it across level attempts. This was included because longer levels were generally associated with unsuccessful attempts and perfect matching could not be achieved (Appendix A). The fourth covariate, *verbosity*, was computed by first obtaining the length (i.e., the total number of words spoken) of each utterance, and then computing z-scores across all utterances. We added this covariate to ascertain the incremental predictive validity of the CPS codes over the simple volume of language production. The fifth covariate, *relative start time*, was recorded when participants engaged with the level within the 15-min block. This covariate was also z-scored across all utterances. This was included because it encodes expertise effects, fatigue effects, and time pressure. The last covariate was *manipulation*, with three levels (none, solve as many levels, and maximize gold coins [reference group]).

Pattern Analysis. For this analysis, we identified frequently occurring clusters of indicators (called patterns) and used these as predictors. Specifically, we extracted patterns from the indicator sequences using a sliding window of five utterances (i.e., utterances 1–5 formed the first window and utterances 2–6 formed the second one). On average, the five utterances corresponded to 15.1 s (SD = 8.61) of discourse, which we deemed sufficient for our purposes. For each five-utterance window, we extracted the set of indicators that occurred therein and designated it as a pattern, ignoring duplicated indicators and occurrence order. For example, the five utterances coded with the indicators—[1] *suggests appropriate ideas*, [2] *responds to others' questions/ideas*, [3] *confirms understanding*, [4] *responds to others' questions/ideas*, and [5] *responds to others' questions/ideas* would form the following pattern: *suggests appropriate ideas* + *confirms understanding* + *responds to others' questions/ideas*. Whereas this method ignores temporal ordering of utterances, we chose this approach because we were interested in indicator co-occurrences and not specific sequences. The approach also yields more general patterns and reduces the number of candidate patterns.

Applying the sliding window to the 14,689 utterances across 209 levels resulted in 13,853 five-utterances windows (14,689–4*209 to account for the ending boundaries). We identified 1361 distinct patterns with a range of 0–7 indicators per pattern ($M = 3$; SD = 1). We focused on the three-indicator patterns (i.e., the mean) as they were sufficiently frequent and provided an opportunity to capture utterances from the three participants who were engaged in the trialogue. We included a subset of these patterns (details below) as predictors in the mixed-effects ordinal regression model along with the six covariates mentioned above.

4. Results

Table 4 shows proportional occurrence of indicators, computed by summing the counts for each indicator across all utterances and dividing by the total number of utterances (i.e., 14,689). The table is sorted in decreasing order of frequency. The most frequently (10.7%) occurring indicator was *responds to other's ideas/questions*, which includes short responses like "yes/no," "I agree," and "that makes sense." The second (10%) most frequent indicator, *confirms understanding*, showed that team members checked their understanding by asking questions or paraphrasing. Another frequent indicator (8.5%) was *provides instructional* support, which occurred when the player who controlled the mouse was unsure of what to draw on the screen, and the other players provided step-by-step instructions. Interestingly, the occurrence of *suggests appropriate ideas* (6.3%) was only slightly higher than *suggests inappropriate ideas* (5.2%); collectively suggesting ideas was frequent (10.7%). Team members also tended to compliment or encourage each other (6%)

Table 4
Descriptive statistics of utterances (n = 14,689) analyzed in the mixed-effect models.

Indicators	Mean	SD	Frequency Range
[NEGO] Responds to others' ideas/questions	0.107	0.309	0–1
[CONST] Confirms understanding	0.100	0.304	0–4
[MAINTAIN] Provides instructional support	0.085	0.297	0–3
[CONST] Suggests appropriate ideas	0.063	0.261	0–3
[MAINTAIN] Compliments or encourages others	0.060	0.238	0–1
[CONST] Talks about challenge situation	0.057	0.232	0–2
[CONST] Suggests inappropriate ideas (R)	0.052	0.235	0–3
[NEGO] Provide reasons to support a solution	0.036	0.190	0–3
[CONST] Interrupts others (R)	0.029	0.168	0–1
[NEGO] Discusses the results	0.027	0.161	0–2
[NEGO] Questions/Corrects others' mistakes	0.022	0.146	0–2
[MAINTAIN] Asks for suggestions	0.011	0.238	0–1
[MAINTAIN] Apologizes for one's mistakes	0.007	0.083	0–1
[MAINTAIN] Initiative off-topic conversation (R)	0.004	0.059	0–1
[MAINTAIN] Joins off-topic conversation (R)	0.004	0.059	0–1
[NEGO] Strategizes to achieve task goals	0.003	0.058	0–1
[NEGO] Brings up giving up the challenge (R)	0.003	0.056	0–1
[NEGO] Tries to quickly save almost successful attempts	0.001	0.036	0–1
[NEGO] Criticizes, makes fun of others (R)	0.000	0.014	0–1
Facets			
Constructing shared knowledge (CONST)	0.291	0.507	0–4
Negotiation/coordination (NEGO)	0.195	0.401	0–3
Maintaining team function (MAINTAIN)	0.170	0.394	0–3

and were generally polite (the indicator, *criticizes or makes fun of others*, occurred only three times).

When mapped to the facet-level (See Table 2), the majority of utterances involved shared knowledge construction (29.1%), followed by negotiation/coordination (19.5%), and maintaining team function (17%), suggesting that the team was quite focused on problem solving.

We also computed the proportional occurrence of all three-indicator patterns. We used binary labels to indicate the pattern that occurred within each window (i.e., if a pattern occurred within a window, it was labeled as 1; otherwise, 0). The frequency was calculated by dividing the sum of occurrence of each pattern by the total number of windows (i.e., 13,853).

We further analyzed the top three (out of 347) frequently-occurring three-indicator patterns as these occurred in approximately 1% of the windows: [P1] *confirms understanding* + *responds to others' questions/ideas* + *provides instructional* support (2.82%); [P2] *suggests appropriate ideas* + *confirms understanding* + *responds to others' questions/ideas* (0.95%); and [P3] *suggests inappropriate ideas* + *confirms understanding* + *responds to others' questions/ideas* (0.86%). [P1] suggests the team members were mutually checking each other's understanding, responding to others' statements and/or questions, and offering help when needed. [P2] and [P3] shows that when someone proposes an idea (either appropriate or inappropriate), others listen attentively and respond, such as acknowledging, asking questions, and paraphrasing.

4.1. CPS indicators to predict in-game performance

We removed the "Criticizes, makes fun of others" indicator from model since it rarely occurred ($M = 0.0002$). All variables in our model had VIFs lower than 2, so we concluded that there were no multicollinearity issues.

Table 5 shows the model results, 95% confidence intervals, and p values, computed based on the z-distribution. Significant odds ratios (ORs) greater than 1 indicate a positive predictor; whereas significant ORs less than 1 indicate a negative indicator. The OR itself is an effect size metric. For example, the OR for "suggests appropriate ideas" was 1.19, indicating that a one unit increase in this indicator makes it 1.19 times more likely to result in a positive outcome. The table also includes the following random effects: within match ID variance (σ^2), between match ID variance (τ_{00}), and the intra-class correlation coefficient (ICC, measuring the proportion of variance in the outcome explained by the nesting factor match ID). Finally, the marginal R^2 value indicates the proportion of variance explained by the fixed effects in the statistical model, while the conditional R^2 shows the proportion of variance explained by the fixed and mixed effects.

Table 5
Ordinal Mixed Effects Model: Using Specific Indicators to Predict No Coin vs. Silver vs. Gold [Level Attempts].

	Model 1		Model 2	
	Odds Ratio	p	Odds Ratio	p
Predictors				
Talks about challenge situation	0.97	0.526	0.98	0.696
Suggests appropriate ideas	1.19	**0.001**	–	–
Suggests inappropriate ideas	0.86	**0.002**	–	–
Suggests potential ideas	–	–	1.05	0.148
Confirms understanding	1.07	0.059	1.08	**0.031**
Interrupts others	1.08	0.192	1.08	0.200
Provides reasons to support a solution	1.03	0.562	1.05	0.367
Questions/corrects others' mistakes	1.06	0.396	1.08	0.279
Responds to others' ideas/questions	1.08	**0.033**	1.08	**0.022**
Discusses results	1.15	**0.025**	1.16	**0.018**
Strategizes to achieve task goals	0.94	0.713	0.95	0.783
Brings up giving up the challenge	0.43	**0.001**	–	–
Tries to quickly save almost successful attempts	1.18	0.523	–	–
Asks for suggestions	0.82	0.056	0.83	0.071
Compliments or encourages others	1.13	**0.004**	1.14	**0.002**
Initiates off-topic conversation	0.97	0.880	0.98	0.923
Joins off-topic conversation	1.07	0.699	1.08	0.664
Provides instructional support	1.01	0.860	1.02	0.680
Apologizes for one's mistakes	0.92	0.517	0.93	0.563
Covariates				
Relative Utterance Index	1.00	0.926	0.99	0.887
Concept [PoT]	2.93	**0.001**	2.91	**0.001**
Duration [Z Score]	0.48	**0.001**	0.48	**0.001**
Verbosity [Z Score]	0.99	0.665	0.99	0.345
Relative Start Time [Z Score]	0.46	**0.001**	0.46	**0.001**
Manipulation [Warmup]	1.87	**0.003**	1.87	**0.003**
Manipulation [Levels]	0.86	0.454	0.86	0.460
Random effects				
σ^2	1.00		1.00	
$\tau_{00\ matchID}$	0.57		0.57	
ICC	0.36		0.36	
Marginal R^2/Conditional R^2	0.36/0.59		0.36/0.59	

We found that 6 of the 18 indicators significantly predicted the CPS outcome. As expected, *suggests appropriate ideas*, *compliments or encourages others*, *responds to others' ideas/questions,* and *discusses the results* positively predicted CPS outcomes. Additionally, *suggests inappropriate ideas* and *brings up giving up the challenge* negatively predicted the outcome. Two indicators, *confirms understanding* and *asks for suggestions* were marginally significant predictors.

Four of our indicators provide some indication of the team's progress in arriving at a solution (i.e., *suggests appropriate/inappropriate ideas*, *brings up giving up the challenge, tries to quickly save almost successful attempts)*. And even though these indicators do not directly code the CPS outcome and coders were blind to the outcome (until coding for a level was complete), we ran an additional model to address possible confounding effects that the ratings were biased by these indicators. Specifically, we removed the *brings up giving up the challenge* and *tries to quickly save almost successful attempts* indicator. We also combined *suggests appropriate ideas* and *suggests inappropriate ideas* into a new *suggests potential ideas* indicator to test whether the quality of ideas matters or if quantity is sufficient. The results are shown in Model 2 (Table 5).

Indeed, we found that the combined indicator *suggests potential ideas* did not predict the outcome. This shows that the effects of suggesting appropriate and inappropriate ideas canceled out when they were

combined, as *suggests appropriate ideas* positively influenced the outcome whereas *suggests inappropriate ideas* negatively affected the outcome. However, suggesting any kind of ideas still mattered to some degree, as indicated by the odds ratio larger than 1. Four indicators (bold font in the table) were significant predictors in both models. They were: *responds to others ideas/questions, compliments or encourages others* and *discusses results*. *Confirms understanding* was marginally significant in Model 1 and significant in Model 2. These might be essential behaviors for successful CPS outcomes.

With respect to the covariates, as expected, PoT levels were much easier to solve than EcT levels, and the more time spent on a level, the less likely the team was at successfully solving the level. Relative start time was also inversely related to the outcome, suggesting that participants tended to solve levels at the earlier stages of gameplay within the block, perhaps when they were more refreshed. Verbosity and relative utterance index did not significantly predict the CPS outcome indicating that the content of the communication mattered more than the length of utterances. In terms of manipulation, the reference group was the sessions with the goal of getting as many gold coins as possible (i.e., high-quality solutions), which did not differ from the other experimental condition (earn as many silver coins by solving as many levels as possible), presumably because this variable was used as a covariate in the matching. Unsurprisingly, both yielded lower outcomes than the easier warmup levels with no manipulations.

The ICC value indicated that 40% of the variance in the in-game performance was explained by the matched sets of level attempts. Further, the fixed effects (i.e., the predictors and covariates) explained 36% of the variance in the in-game performance (marginal R^2), suggesting that there are other variables that influence the in-game performance in addition to the variables included in our model. The random and the fixed effects collectively explained about 59% of the variance in participants' success in gameplay (conditional R^2).

4.2. Pattern analysis: CPS indicators to predict binary in-game performance

In the pattern analysis, we built an ordinal regression model using the three frequently occurring three-indicator patterns to predict level success. In addition, we included six significant individual indicators from Model 1 to examine whether the patterns provide additional information and one additional indicator (*provides instructional* support) that was part of the patterns itself. Additionally, we included the same five covariates as in the indicator-level model (calculated using the five-utterance window).

Model 3 (Table 6) shows that among the three most frequent patterns, one significantly predicted the outcome (none vs. silver vs. gold) [P2] (i.e., *suggests appropriate ideas + confirms understanding + responds to others' questions/ideas)*: predicted the outcome whereas three individual indicators were not significant. The discourse pattern [P3] (i.e., *suggests inappropriate ideas + confirms understanding + responds to others' questions/ideas*) was not a significant predictor with respect to the type of coin earned, although *suggests inappropriate ideas* alone had predictive power. It appears that the negative impact of suggesting inappropriate ideas could be mitigated when other team members attempted to paraphrase or ask clarification questions and formed a conversation cycle to check and establish mutual understanding. Interestingly, non-significant individual indicators (*suggests appropriate ideas, confirms understanding, and responds to others' ideas/questions*) formed a constructive communication pattern [P2] that significantly predicted the CPS outcome. Interestingly, the most frequent pattern ([P1], *confirms understanding + responds to others' ideas/questions + provides instructional* support) was not a significant predictor in this model.

Similar to the indicator-level analysis, we re-ran the pattern analysis by removing *brings up giving up the challenge* and combining the two indicators – *suggests appropriate ideas* and *suggests inappropriate ideas* into a new pattern *suggests potential ideas*. These adjustments increased the

Table 6
Ordinal Mixed Effects Model: Using Patterns to Predict No Coin vs. Silver vs. Gold [Level Attempts].

	Model 3			Model 4	
	Odds Ratio	*p*		Odds Ratio	*p*
Predictors			**Predictors**		
[P1] confirms understanding + responds to others' ideas/questions + provides instructional support	0.92	0.207	[P1] confirms understanding + responds to others' ideas/questions + provides instructional support	0.92	0.201
[P2] suggests appropriate ideas + confirms understanding + responds to others' questions/ideas	1.62	**0.001**	[P2] suggests potential ideas + confirms understanding + responds to others' ideas/questions	1.19	**0.024**
[P3] suggests inappropriate ideas + confirms understanding + responds to others' questions/ideas	0.82	0.096	[P3] suggests potential ideas + confirms understanding + provides instructional support	0.73	**0.004**
			[P4] suggests potential ideas + provides reasons to support an idea + responds to others' questions/ideas	0.89	0.283
			[P5] suggests potential ideas + responds to others' questions/ideas + compliments or encourages others	1.19	0.125
Confirms understanding	0.88	0.255			
Responds to others' ideas/questions	1.16	0.145	Confirms understanding	0.89	0.266
Provides instructional support	1.07	0.456	Responds to others' ideas/questions	1.16	0.142
Suggests appropriate ideas	1.16	0.107	Provides reasons to support a solution	0.79	0.496
Suggests inappropriate ideas	0.54	**0.001**	Provides instructional support	1.06	0.469
Compliments or encourages others	1.10	0.162	Compliments or encourages others	1.10	0.153
Discusses results	1.25	**0.020**	Suggests potential ideas	0.81	**0.001**
Brings giving up the challenge	0.38	**0.016**	Discusses results	1.25	**0.021**
Covariates			**Covariates**		
Relative Utterance Index	1.00	0.903	Relative Utterance Index	0.99	0.875
Concept [PoT]	3.00	**0.001**	Concept [PoT]	2.99	**0.001**
Duration [Z Score]	0.48	**0.001**	Duration [Z Score]	0.48	**0.001**
Verbosity [Z Score]	1.00	0.893	Verbosity [Z Score]	1.00	0.686
Relative Start Time [Z Score]	0.47	**0.001**	Relative Start Time [Z Score]	0.47	**0.001**
Manipulation [Warmup]	1.88	**0.004**	Manipulation [Warmup]	1.88	**0.004**
Manipulation [Levels]	0.86	0.470	Manipulation [Levels]	0.86	0.465
Random effects			**Random effects**		
σ^2	1.00		σ^2	1.00	
$\tau_{00\ matchID}$	0.60		$\tau_{00\ matchID}$	0.60	
ICC	0.37		ICC	0.37	
Marginal R^2/ Conditional R^2	0.35/0.60		Marginal R^2/ Conditional R^2	0.35/0.60	

Note: P is short for "Pattern", so P1 means Pattern 1.

number of frequent patterns (roughly 1% or greater occurrence) to five: [P1; 2.82%] *confirms understanding + responds to others' ideas/questions + provides instructional* support, [P2; 1.88%] *suggests potential ideas + confirms understanding + responds to others' ideas/questions*, [P3; 1.08%] *suggests potential ideas + confirms understanding + provides instructional* support, [P4; 0.99%] *suggests potential ideas + provides reasons to* support *an idea + responds to others' questions/ideas*, and [P5; 0.84%] *suggests potential ideas + responds to others' questions/ideas + compliments or encourages others*.

These five patterns were included in the new model along with seven relevant individual indicators (Model 4, Table 6). As before, *discusses results* positively predicted the outcome as it suggests metacognitive reflection. We were initially puzzled to find that suggesting potential ideas was negatively related to CPS outcomes. However, it might be the case that potential inappropriate ideas had more of an effect than appropriate ideas (see Model 3). The results were more illuminating when potential ideas were examined within the context of the two significant patterns [P2] and [P3], which were significant positive and negative predictors, respectively. Both patterns showed the importance of building on a potential idea by confirming understanding, but responding when clarifications were needed [P2] was productive whereas simply providing instructional support [P3] was not. Further, the other patterns accompanying this indicator [P4] and [P5] were not significant predictors of CPS outcomes. In summary, these findings suggest that team member interactions add another layer contributing to outcome quality. That is, when ideas are suggested, team members should build on them with constructive and responsive communications to achieve high quality outcomes.

5. Discussion

We investigated how CPS skills influence objective CPS outcomes in a game-based collaborative learning environment. Our main findings along with directions for future work are summarized below.

5.1. Main findings

We identified the relationships between CPS measures (at the specific indicator level and pattern level) and in-game performance when triads engaged in CPS using a physics game. The indicator-level model revealed that conversations that involved talking about appropriate ideas contributed to desirable outcomes whereas discussing inappropriate ideas tended to divert the team to a nonproductive direction. Simply suggesting ideas, however, was not a significant predictor, which is unsurprising since successful CPS entails both collaboration and problem-solving skills. The problem-solving part of CPS requires team members to have basic background knowledge so that appropriate solution plans can be devised. It is possible that a team persists on applying an inappropriate idea which can lead to unsuccessful results, particularly when no one has sufficient knowledge to rectify the situation. But knowledge is itself insufficient in that a knowledgeable team member may not be able to apply their knowledge if they are shy or if the other members are too dominant. Indeed, CPS entails dynamic and constructive interactions among team members in addition to individual contributions.

In addition to suggesting ideas, other indicators were identified as essential for successful CPS task performance. Complimenting and encouraging fellow team members helps to create a positive collaborative environment, which in turn stimulates good quality collaboration. Additionally, discussing results from implemented solutions benefits performance. That is, monitoring and reflecting on the results from a recent solution attempt might encourage team members to refine their solutions (if warranted) (Andrews-Todd & Forsyth, 2020; Care et al., 2016). Additionally, confirming understanding and responding to others' ideas/questions also predicted high-quality level solutions. This suggests that checking in with team members is crucial to establishing common ground and generating executable solutions. Simply acknowledging others' ideas and being responsive could facilitate negotiation processes that are needed for a successful outcome. This is consistent with the literature on the importance of reciprocal exchanges of communication in collaboration contexts (e.g., Barron, 2000; Hesse et al., 2015).

The pattern analysis revealed additional interactive patterns that contributed to team outcomes. First, team members should frequently check each other's understanding and respond to others' statements or questions, to ensure mutual understanding and establish common ground. From a conversation perspective, the pattern analysis indicates multiple rounds of turn taking is needed to establish a shared understanding. If someone suggests an idea, then others should follow up and discuss the feasibility of the idea, instead of simply doing what was instructed. Thus, multiple conversational turn taking helps develop meaning among team members. It is also a sign of active participation, not passive acceptance of a suggested idea.

Mapping our indicator-level results to the CPS facets (Table 7) indicates that aspects of all three CPS facets were predictive of the outcome. CPS requires each individual to share knowledge, skills, and resources, monitor team progress, and maintain a functional team environment (Andrews-Todd & Forsyth, 2020; OECD, 2017). This is done by sharing one's expertise, coordinating with others, and keeping the team spirit positive (via complements). It further reinforces that cognitive and social skills are interconnected in CPS as demonstrated by the fact that each identified patterns involve the combination of social (e.g., responding to others and asking clarification questions) and cognitive (e.g., contributing ideas and talking about tasks) elements. In short, our findings demonstrate that CPS is a socio-cognitive construct (Dowell et al., 2020), and separating the problem solving and collaboration aspects may not be desirable for CPS assessment.

5.2. Limitations & future work

Our results should be interpreted in light of some limitations. Our study was conducted in a laboratory setting which may not mimic the real-world CPS activity environment. The CPS task used in our study related to learning physics through gameplay. As a result, some aspects of the coding scheme may not fully capture CPS behaviors that may occur in ill-structured CPS problems. Related to that, different features of different games may influence how people behave during CPS tasks. Thus, analyzing and comparing different game structures in CPS environments can be informative for future studies. Another limitation relates to the simplicity of our pattern analysis method in this initial investigation. We applied five-utterance sliding windows in our analysis to see how co-occurring of indicators would influence CPS performance outcomes. Even though the coders were not informed of the performance outcome in advance, some level of knowledge was needed for judging whether an idea was appropriate vs. inappropriate to solve a particular game level. Lastly, the present study was exploratory in terms of understanding the relationship between CPS behaviors and

Table 7
Mapping of CPS indicators/patterns to facets for main models (Models 1 and 3).

Facet	Indicator/Pattern
Constructing shared knowledge	Suggests appropriate ideas (positive) Suggests inappropriate ideas (negative) Confirms understanding (marginally positive)
Negotiation/coordination	Responds to others' ideas/questions (positive) Discusses results (positive) Brings up giving up the challenge (negative)
Maintaining team function	Compliments or encourages others (positive) Asks for suggestions (marginally negative)
Constructing shared knowledge + Negotiation/coordination	[P2] suggests appropriate ideas + confirms understanding + responds to others' questions/ideas

performance outcomes in triadic, spoken, and game-based learning environments. Although the CPS framework we used has been validated across multiple tasks (Sun et al., 2020) and we used an objective measure of performance, it is important to replicate our findings across multiple contexts. Whereas we do not expect that the same set of indicators or patterns will be predictive of task performance in all contexts, as similar studies emerge, the field should be in a position to identify a set of generalizable CPS behaviors that underly performance outcomes.

Our results also indicate several potential areas of future work. We suggest investigating the non-significant indicators in the models we tested. For example, *provides instructional* support frequently occurred during students' communications, but it did not directly relate to the outcome quality. Too much instructional support from other team members may lead to passive participation of the person controlling the interface, or to unsolved levels if the instructional support involves in inappropriate idea. Another indicator—*interrupts others*—might be seen as a double-edge sword. That is, someone could interrupt due to being aggressive and thus impede the CPS processes (Chiu, 2008). But someone could also interrupt to seek clarification, rectify a misunderstanding, or share their excitement (Roschelle & Teasley, 1995). It is also possible that CPS indicators differentially contribute to other CPS outcomes (e.g., subjective perceptions of the interaction), which was not examined in the current study.

Another possible reason that not all indicators related to our performance outcome is that our experimental design did not provide sufficient time for some of the indicator effects to unfold. The literature does not provide empirical evidence regarding how long it takes to form an effective and efficient team in CPS environments. Our design of three 15-min blocks may have been inadequate for triads to fully demonstrate their CPS skills. To this point, we found that teams were less likely to succeed in levels as they were approaching the end of a block. Moreover, the triads in our design switched roles in each block (i.e., controller vs. contributors), which could affect team dynamics. Future studies could also investigate the longitudinal development of CPS skills among team members across days, weeks, or even months. Furthermore, due to the complexity of the CPS construct, skill development might not be linear, so future studies could focus on the dynamics of CPS skills development.

In general, additional research is needed to get a comprehensive understanding of the relationship between specific indicators and subsequent CPS outcomes (Hao et al., 2019). Such an understanding can benefit tailored CPS training (Andrews-Todd & Forsyth, 2020). For instance, based on the results from the current study with the goal to improve CPS task performance, training that emphasizes the significant indicators and patterns listed in Tables 5 and 6 would be impactful. If the goal was to enhance content knowledge or subjective perceptions, then other indicators may be focal.

One other issue to consider is the time-consuming nature of human coding of CPS behaviors, which motivated us to examine a subset of the video recordings. With advances in artificial intelligence, specifically automatic speech recognition and natural language processing (Blanchard et al., 2015; Devlin, Chang, Lee, & Toutanova, 2019; Le Cun, Bengio, & Hinton, 2015), the data generated from human coding could be further utilized towards automated coding of CPS indicators. This would permit testing of the prediction accuracy of various AI techniques using human coding data (e.g., Stewart et al., 2020; Hao et al., 2019; von Davier et al., 2017). In the same vein, automated assessment of CPS in human-human interactions can enable timely feedback. For example, when real-time assessment detects that group members are ignoring each other, then an appropriate intervention/message could be deployed to facilitate communication among team members. Future research should investigate the effectiveness of timely feedback on participants' CPS skill development. In addition, researchers could consider the best way to report CPS skills to stakeholders (e.g., teachers, employers, and team members). Simply presenting current CPS facet scores is likely inadequate. Instead of showing scores, perhaps a progress bar could be displayed, along with descriptions related to particular strengths and weaknesses – as well as ways to improve certain CPS skills. Moreover, researchers should examine whether to provide feedback to the team or to individual team members.

6. Conclusions

Our study examined how CPS behaviors and interactions affected performance while triads engaged in CPS tasks in a game-based learning environment. We found associations among fine-grained indicators as well as patterns of co-occurring indicators and CPS success. The findings emphasized that CPS requires individual contributions along with constructive interactions. Also, the cognitive and social aspects are integral to CPS. Existing CPS models (e.g., PISA and ATC21S) clearly distinguish the two aspects, which tend to deemphasize the interconnectedness between social and cognitive skills. The findings can inform intervention designs to improve students' CPS skills in future research.

Author contribution

Chen Sun: Conceptualization, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. Valerie Shute: Conceptualization, Software, Resources, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. Angela E. B. Stewart: Software, Investigation, Data curation, Review & Editing, Project administration. Quinton Beck-White: Data curation. Caroline R. Reinhardt: Data curation, Investigation. Guojing Zhou: Formal analysis, Visualization, Review & Editing. Nicholas Duran: Review & Editing, Data curation, Supervision, Project administration. Sidney K. D'Mello: Conceptualization, Methodology, Data curation, Resources, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this manuscript.

Acknowledgements

This work was supported by the US National Science Foundation [award number 1660859].

Appendix A

Tables for Level-matching Statistics.

Level-matching: Distribution of the 209 level attempts across schools, experimental blocks, and goal manipulation.

	Warmup			PoT			EcT	
	Gold	Silver	No Coin	Gold	Silver	No Coin	Coin	No Coin
School1	11	11	11	9	9	9	20	20
School2	11	11	11	16	16	16	15	15
Block1	–	–	–	8	8	8	13	13
Block2	–	–	–	17	17	17	22	22
Levels	–	–	–	11	11	11	13	13
Golds	–	–	–	14	14	14	22	22

Means (and Standard Deviations) of pretest score and duration of the 209 level attempts across coin types.

		Gold	Silver	Coin	No coin
Average Team Pretest Score	Warmup	6.2 (1.2)	6.3 (1.3)	–	7.0 (1.2)
	PoT	6.5 (1.3)	6.9 (1.4)	–	6.5 (1.4)
	EcT	–	–	6.7 (1.2)	6.9 (1.2)
Level Duration	Warmup	235.6 (152.4)	186.8 (119.0)	–	263.4 (189.7)
	PoT	218.1 (129.2)	270.2 (210.8)	–	215.7 (172.3)
	EcT	–	–	292.4 (175.8)	357.0 (257.8)

Appendix B

A Simplified Excerpt of the Dataset.

This table shows a simplified excerpt of the major content included in the dataset, specifically, transcript information and associated indicators. For each transcribed utterance, the expert coders would label which indicator(s) occurred in the utterance. Each utterance can be coded to multiple indicators and an indicator can occur multiple times for one utterance.

Partici-pant	Start Time	End time	Transcript	I1	I2	I3	…	I19	duplicate
PA	70.12	71.03	Utterance 1	0	2	1	…	0	0
PC	71.33	73.4	Utterance 2	1	0	1	…	0	0
PB	80.23	83.57	Utterance 3	0	1	0	…	1	0
PC	82.58	83.33	Utterance 4	0	0	3	…	0	0
PB	84.60	90.12	Utterance 5	2	1	0		0	0

Appendix C

All Measures in the Main Study.

Home Measures	Demographics	Gender, age, major, GPA, etc
	Big-five personality (Brief)	Gosling, Rentfrow, and Swann (2003)
	Leadership self-efficacy	Hoyt and Blascovich (2010)
	Individual satisfaction with teamwork	De la Torre-Ruiz, Ferron-Vilchez, and Ortiz-de-Mandojana (2014)
	Physics self-efficacy	Lindstrøm and Sharma (2011)
	Physics pretest test (form X/Y)	Developed by physics experts
	Intrinsic motivation inventory (IMI) for Physics Playground; IMI for Minecraft	Deci and Ryan (1982)
	Minecraft tutorial check	Researcher-developed items
Lab Measures	Valence/Arousal (for each 15-min block)	Researcher-developed items
	Team collaborative problem solving quality (for each 15-min block)	Researcher-developed items
	Inclusiveness and Team Norms (for each 15-min session)	Gardner and Pierce (2016); Whitton and Fletcher (2014)
	Physics posttest (form X/Y)	Developed by physics experts

References

Aarikka-Stenroos, L., & Jaakkola, E. (2012). Value co-creation in knowledge intensive business services: A dyadic perspective on the joint problem solving process. *Industrial marketing management, 41*(1), 15–26.

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

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *The Journal of the Learning Sciences, 9*, 403–436.

Blanchard, N., Brady, M., Olney, A. M., Glaus, M., Sun, X., Nystrand, M., et al. (2015). A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial intelligence in education* (pp. 23–33). Berlin Heidelberg: Springer-Verlag.

Care, E., Scoular, C., & Griffin, P. (2016). Assessment of collaborative problem solving in education environments. *Applied Measurement in Education, 29*(4), 250–264. https://doi.org/10.1080/08957347.2016.1209204

Chang, C. J., Chang, M. H., Chiu, B. C., Liu, C. C., Chiang, S. H. F., Wen, C. T., et al. (2017). An analysis of student collaborative problem solving activities mediated by collaborative simulations. *Computers & Education, 114*, 222–235. https://doi.org/10.1016/j.compedu.2017.07.008

Chiu, M. M. (2008). Effects of argumentation on group micro-creativity: Statistical discourse analyses of algebra students' collaborative problem solving. *Contempo Educational Psychology, 33*, 382–402.

Csanadi, A., Eagan, B., Kollar, I., Shaffer, D. W., & Fischer, F. (2018). When coding-and-counting is not enough: Using epistemic network analysis (ENA) to analyze verbal data in CSCL research. *International Journal of Computer-Supported Collaborative Learning, 13*(4), 419–438.

D'Mello, S., Dowell, N., & Graesser, A. (2011). Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology: Applied, 17*(1), 1–17.

von Davier, A. A., Hao, J., Liu, L., & Kyllonen, P. (2017). Interdisciplinary research agenda in support of assessment of collaborative problem solving: Lessons learned from developing a collaborative science assessment prototype. *Computers in Human Behavior, 76*, 631–640. https://doi.org/10.1016/j.chb.2017.04.059

Deci, E. L., & Ryan, R. M. (1982). *Intrinsic motivation inventory [measurement instrument]*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*.

Dowell, N. M., Lin, Y., Godfrey, A., & Brooks, C. (2020). Exploring the relationship between emergent sociocognitive roles, collaborative problem-solving skills, and outcomes: A group communication analysis. *Journal of Learning Analytics, 7*(1), 38–57.

Dowell, N. M., Nixon, T. M., & Graesser, A. C. (2019). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior Research Methods, 51*(3), 1007–1041.

Eloy, L., EB Stewart, A., Jean Amon, M., Reinhardt, C., Michaels, A., Sun, C., D'Mello, S., et al. (2019, October). Modeling team-level multimodal dynamics during multiparty collaboration. In *2019 International Conference on Multimodal Interaction* (pp. 244–258).

Forsyth, C., Andrews-Todd, J., & Steinberg, J. (2020). Are you really a team player?: Profiles of collaborative problem solvers in an online environment. In A. N. Rafferty, J. Whitehill, V. Cavalli-Sforza, & C. Romero (Eds.), *Proceedings of the 13th international Conference on educational data mining (EDM 2020)* (pp. 403–408).

Gardner, D. G., & Pierce, J. L. (2016). Organization-based self-esteem in work teams. *Group Processes & Intergroup Relations, 19*(3), 394–408.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504–528.

Graesser, A. C., Greiff, S., Stadler, M., & Shubeck, K. T. (2020). Collaboration in the 21st century: The theory, assessment, and teaching of collaborative problem solving. *Computers in Human Behavior, 104*. https://doi.org/10.1016/j.chb.2019.09.010

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*(1), 29–48.

Hao, J., Liu, L., Kyllonen, P., Flor, M., & von Davier, A. A. (2019). Psychometric considerations and a general scoring strategy for assessments of collaborative problem solving. *ETS Research Report Series, 2019*(1), 1–17. https://doi.org/10.1002/ets2.12276

Hao, J., Liu, L., von Davier, A., Kyllonen, P. C., & Kitchen, C. (2016). Collaborative problem solving skills versus collaboration outcomes: Findings from statistical analysis and data mining. In *Proceedings of the 9th international conference of educational data mining (EDM)* (pp. 382–387).

Heirati, N., & Siahtiri, V. (2019). Driving service innovativeness via collaboration with customers and suppliers: Evidence from business-to-business services. *Industrial Marketing Management, 78*, 6–16.

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans. *Computers in Human Behavior, 104*. https://doi.org/10.1016/j.chb.2018.07.035

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

Hmelo-Silver, C. E. (2003). Analyzing collaborative knowledge construction: Multiple methods for integrated understanding. *Computers & Education, 41*(4), 397–420.

Hmelo, C. E., Nagarajan, A., & Day, R. S. (2000). Effects of high and low prior knowledge on construction of a joint problem space. *The Journal of Experimental Education, 69*(1), 36–56.

Hoyt, C. L., & Blascovich, J. (2010). The role of leadership self-efficacy and stereotype activation on cardiovascular, behavioral and self-report responses in the leadership domain. *The Leadership Quarterly, 21*(1), 89–103.

Huang, J., Hmelo-Silver, C. E., Jordan, R., Gray, S., Frensley, T., Newman, G., et al. (2018). Scientific discourse of citizen scientists: Models as a boundary object for collaborative problem solving. *Computers in Human Behavior, 87*, 480–492. https://doi.org/10.1016/j.chb.2018.04.004

Le Cun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature, 521*, 436–444.

Lindstrøm, C., & Sharma, M. D. (2011). Self-efficacy of first year university physics students: Do gender and prior formal instruction in physics matter? *International Journal of Innovative Science and Modern Engineering, 19*(2), 1–19.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning, 2*(1), 63–86. https://doi.org/10.1007/s11412-006-9005-x

Moreland, R. L. (2010). Are dyads really groups? *Small Group Research, 41*(2), 251–267. https://doi.org/10.1177/1046496409358618

OECD. (2017). *PISA 2015 collaborative problem-solving framework*. Retrieved from https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf.

Ostrander, A., Bonner, D., Walton, J., Slavina, A., Ouverson, K., Kohl, A., et al. (2020). Evaluation of an intelligent team tutoring system for a collaborative two-person problem: Surveillance. *Computers in Human Behavior, 104*. https://doi.org/10.1016/j.chb.2019.01.006

Reilly, J. M., & Schneider, B. (2019). Predicting the quality of collaborative problem solving through linguistic analysis of discourse. In *Proceedings of the 12th international conference on educational data mining (EDM)* (pp. 149–157).

Reiter-Palmon, R., Sinha, T., Gevers, J., Odobez, J. M., & Volpe, G. (2017). Theories and models of teams and groups. *Small Group Research, 48*(5), 544–567.

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. E. O'Malley (Ed.), *Computer supported collaborative learning* (pp. 69–97). Heidelberg: Springer.

Rosen, Y. (2017). Assessing students in human-to-agent settings to inform collaborative problem-solving learning. *Journal of Educational Measurement, 54*, 36–53.

Rosen, Y., Wolf, I., & Stoeffler, K. (2020). Fostering collaborative problem solving skills in science: The Animalia project. *Computers in Human Behavior, 104*. https://doi.org/10.1016/j.chb.2019.02.018

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The Journal of the Learning Sciences, 14*(2), 201–241.

Schulze, J., & Krumm, S. (2017). The "virtual team player" A review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. *Organizational Psychology Review, 7*(1), 66–95.

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

Stewart, A. E., Amon, M. J., Duran, N. D., & D'Mello, S. K. (2020, April). Beyond team makeup: Diversity in teams predicts valued outcomes in computer-mediated collaborations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). https://doi.org/10.1145/3313831.3376279.

Stoeffler, K., Rosen, Y., Bolsinova, M., & von Davier, A. A. (2020). Gamified performance assessment of collaborative problem solving skills. *Computers in Human Behavior, 104*. https://doi.org/10.1016/j.chb.2019.05.033

Shute, V., Almond, R., & Rahimi, S. (2019). *Physics Playground (1.3) [Computer software]*. https://pluto.coe.fsu.edu/ppteam/pp-links/.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education, 143*, 1–17. https://doi.org/10.1016/j.compedu.2019.103672.

Swiecki, Z., Ruis, A. R., Farrell, C., & Shaffer, D. W. (2020). Assessing individual contributions to collaborative problem solving: A network analysis approach. *Computers in Human Behavior, 104*. https://doi.org/10.1016/j.chb.2019.01.009

de la Torre-Ruiz, J. M., Ferrón-Vílchez, V., & Ortiz-de-Mandojana, N. (2014). Team decision making and individual satisfaction with the team. *Small Group Research, 45*(2), 198–216.

Vrzakova, H., Amon, M. J., Rees, M., Faber, M., & D'Mello, S. K. (2020). Looking for a deal? Social visual attention during negotiations via mixed media videoconferencing. *Proceedings of the ACM: Human Computer Interaction, Computer Supported Collaborative Work, 4*(CSCW3), 1–35.

Whitton, S. M., & Fletcher, R. B. (2014). The group environment questionnaire: A multilevel confirmatory factor analysis. *Small Group Research, 45*(1), 68–88.

Zubizarreta, J. R., Kilcioglu, C., & Vielma, J. P. (2018). *Matched samples that are balanced and representative by design*. R package version 0.3.1.