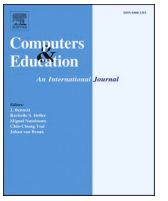
Computers & Education 207 (2023) 104928

Contents lists available at ScienceDirect

Computers & Education

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



Investigating collaborative problem solving skills and outcomes across computer-based tasks☆

Jessica Andrews-Todd [a,*], Yang Jiang [a], Jonathan Steinberg [a], Samuel L. Pugh [b], Sidney K. D'Mello [b]

[a] Educational Testing Service, Princeton, NJ, USA
[b] University of Colorado Boulder, Boulder, CO, USA

ARTICLE INFO

Keywords:
Cooperative/collaborative learning
Secondary education
21st century abilities
Games

ABSTRACT

Collaborative problem solving (CPS) is a critical competency for the modern workforce, as many of todays' problems require groups to come together to find innovative solutions to complex problems. This has motivated increased interest in work dedicated to assessing and developing CPS skills. However, there has been limited attention in prior CPS assessment research on potential differences in how CPS behaviors are exhibited across task contexts. In the current study, we investigated associations among middle- and high-school students' displayed CPS skills across two online (i.e., via videoconferencing) tasks (Physics Playground and the T-Shirt Math Task) and the extent to which different skills were related to CPS outcomes across those tasks. Results showed variation in associations of CPS skills across the tasks, contributing further evidence to our understanding of how different CPS task designs can give students the opportunity to demonstrate different CPS skills. Our findings highlight the potential of incorporating multiple tasks during CPS assessments and can inform future research on CPS task design and computer-based CPS assessment.

1. Introduction

Recent economic, health, technological, and social changes and challenges (e.g., climate change, energy sustainability, the COVID-19 pandemic) have substantively changed the way we live, work, and learn. Such changes have particularly created new demand for competencies like collaborative problem solving (CPS), as solutions to many of these complex problems require groups of individuals with different perspectives and expertise to work together (Graesser et al., 2018; Griffin & Care, 2014; Rosen et al., 2020; Stadler et al., 2020). Correspondingly, many organizations, government agencies, and employers (e.g., the Organization for Economic Cooperation and Development (OECD), Partnership for 21st Century Learning, National Center for Education Statistics, National Research Council) have named CPS and related competencies (e.g., teamwork) as important and necessary for workplace and career success in the 21st century (Casner-Lotto & Barrington, 2006; Fiore et al., 2017; Griffin et al., 2012; McGunagle & Zizka, 2020; National Research Council, 2008; OECD, 2013b; Partnership of 21st Century Learning, 2016; Whorton et al., 2017), more so with the rise of powerful generative AI that risks taking over aspects of individualized cognitive work (Trivedi et al., 2023).

☆ This work was completed when Jonathan Steinberg was with Educational Testing Service. He is now at EurekaFacts.

* Corresponding author. Educational Testing Service, 660 Rosedale Road, MS 16-R, Princeton, NJ, 08541, USA.
E-mail address: jandrewstodd@ets.org (J. Andrews-Todd).

https://doi.org/10.1016/j.compedu.2023.104928
Received 30 June 2022; Received in revised form 11 September 2023; Accepted 14 September 2023
Available online 16 September 2023

Furthermore, decades of research have pointed to the benefits of engaging in collaborative activities in many contexts, including school, the workplace, and the military. These benefits include positive impacts on learning, performance, productivity, efficiency, engagement, the quality of solutions to problems, and social, emotional, and psychological well-being (Andrews & Rapp, 2015; Barron, 2000; Gillies, 2004; Graesser et al., 2018; Jeong et al., 2019; LePine et al., 2008; Lou et al., 2001). As such, in K-12 education, higher education, and workforce contexts, there is increased interest in CPS, including how to assess individuals' skills and develop such skills (Oliveri et al., 2017; Rojas et al., 2021; Tang et al., 2021). However, in much of the work on CPS assessment little attention has been given to how individuals' skills might be exhibited differently across task contexts, which may vary across numerous dimensions including prior knowledge, interest, self-efficacy, etc. This begs the question - are CPS skills task-specific or more task-general?

In the current study, we investigate middle and high school students' CPS skills across two online task contexts. Specifically, student dyads collaborated remotely (i.e., via videoconferencing) on two different computer-based CPS tasks in the Physics and Mathematics domains. CPS skills were scored by trained raters using a fine-grained (i.e., conversational turn-level) ontology-based framework that lays out concepts associated with the CPS competency, their relationships, and indicators or behaviors demonstrating evidence of skills. The ontology includes nine high-level CPS skills and those skills are further grouped into three facets. Our research questions pertain to the degree of associations between students' display of CPS skills across the tasks that differ in various design characteristics, and the extent to which different skills are related to performance outcomes (e.g., task performance) across the tasks.

2. Background

2.1. Collaborative problem solving assessment

CPS competency involves individuals working together by sharing information and pooling knowledge and effort to reach a solution to a problem (OECD, 2013b). As such, CPS is complex and includes individuals' social or collaborative skills (e.g., sharing information) as well as cognitive or problem solving skills (e.g., planning a solution) (Hesse et al., 2015). This complexity has made the measurement of CPS skills a challenging undertaking (Gao et al., 2022; von Davier et al., 2017).

In many instances, methods such as questionnaires, observations, checklists/rubrics, interviews, peer-rated scales, or think-aloud protocols have been utilized to assess skills related to CPS (Aguado et al., 2014; Kyllonen et al., 2017; Oliveri et al., 2017). However, these assessment methods are sometimes not widely applicable to different domains (e.g., some are created for use in specialized medical settings; Britton et al., 2017), have ill-defined items that may complicate interpretation (Hastie et al., 2014), or can introduce self-report biases (socially desirable responding; Paulhus, 1991). Furthermore, traditional assessment orientations like multiple-choice questions are not well suited for a process-oriented construct like CPS (Davey et al., 2015). This has led to an increased interest in using computer-based environments to support the assessment of complex constructs like CPS. In particular, these kinds of environments (e.g., games, simulations, scenario-based tasks) allow individuals to engage in CPS in interactive situations which can resemble real-world activities (Herborn et al., 2020; Shute & Becker, 2010). Importantly, these environments also allow for capturing process data, including actions (e.g., via log files) and discourse (e.g., transcripts of communication) among participating individuals, which are critical sources of evidence regarding individuals' skills (Honey & Hilton, 2011; Quellmalz & Pellegrino, 2009). The process data are essential to providing information beyond just the final answer or solution, which is not always a good indicator of CPS competence (Nouri et al., 2017), such as when a single individual reaches the solution without consulting their teammates.

However, utilizing computer environments for assessment purposes presents many challenges that do not exist for traditional assessments (Andrews-Todd & Forsyth, 2022). For example, the targeted construct needs to be operationalized at the level of granularity of the data captured from the computer environment, and the targeted skills need to be identified in the large streams of individuals' action and discourse data (Andrews-Todd & Forsyth, 2020; Gobert et al., 2012; Romero et al., 2009). Fortunately, there have been recent methodological advances that can help address such challenges, making computer environments a viable option for CPS assessment. For example, a number of frameworks have been proposed for identifying targeted skills and behaviors from open-ended team communications during CPS (Kerr et al., 2016, Andrews-Todd & Kerr, 2019, Andrews-Todd & Forsyth, 2020; Sun et al., 2020, 2022; Hesse et al., 2015; Liu et al., 2015). While such approaches typically depend on human raters to view and annotate the interactions, additional research has shown that the annotation process can be automated using natural language processing (NLP) approaches (Stewart et al., 2019; Pugh et al., 2021, 2022, Flor & Andrews-Todd, 2022; Flor et al., 2016; Hao et al., 2017). Such advances can support the use of open digital environments for CPS assessments that neither constrain communication among teammates nor the problem space, thus affording detailed measurement of individuals' actions and communication. In the current study, we investigate CPS assessment within computer-enabled learning environments across two task contexts.

2.2. Collaborative problem solving across task contexts

CPS can be considered a domain-independent competence (Graesser et al., 2017; Greiff, 2012); however, in practice certain CPS skills may be exhibited differently in different task contexts (Hao et al., 2017) – i.e., the set of variables which defines a task such as the subject domain, goal, format, medium, instructions, etc. One reason is that a given task may present differing complexity or challenges for some individuals based on their prior knowledge (Crippen & Antonenko, 2018). For example, a student with low prior knowledge in a mathematics topic area may exhibit different patterns of CPS skills on a collaborative mathematics task relative to another student with high math prior knowledge. It may be that the low knowledge student does not have sufficient understanding to contribute to the team (e.g., sharing ideas, proposing solutions) compared to others with higher knowledge. This is not to say that each student does not have the targeted CPS skill; however, it may be that the task context does not provide sufficient opportunity for each student to display

the CPS skills of interest.

Previous work has shown some evidence for potential differences in CPS as a function of prior knowledge when examining prior knowledge differences in students characterized according to different CPS skill profiles in an electronics CPS task (Forsyth et al., 2020). Specifically, students characterized as Active Collaborators and Super Socials had higher electronics content knowledge than those characterized as Social Loafers and Low Collaborators. The authors speculated that the higher prior knowledge of Active Collaborators and Super Socials relative to the other groups could have afforded these students the ability to engage in more communication behaviors that contributed to solving the problem. In contrast, the authors also speculated that the Social Loafers and Low Collaborators, with their relatively lower prior knowledge, may not have wanted to collaborate with others or preferred working alone due to embarrassment or discomfort with their level of electronics knowledge (Forsyth et al., 2020).

Additional studies have also found a relationship between prior knowledge or expertise and collaborative behaviors and outcomes (Gijlers & De Jong, 2005; Nokes-Malach et al., 2012; Uz-Bilgin et al., 2020; Zambrano et al., 2019). Such studies have suggested that sufficient prior knowledge can afford detection and correction of errors (Laughlin et al., 2003; Schriver et al., 2008), interpretation, recognition, and evaluation of viable problem solutions (Gu et al., 2015; Nokes-Malach et al., 2012), quality contributions (Resta & Laferrière, 2007), and more substantive discussion of content (Chung et al., 1999). Furthermore, recent research has shown students in higher grades (i.e., more years of education) display better CPS skills, ostensibly because they have better mastery of different disciplines (Ahonen & Harding, 2018; Tang et al., 2021).

Other characteristics of a task or problem space can also potentially relate to how individuals exhibit CPS skills, as certain skills may be exhibited in accordance with different situational needs (Hesse et al., 2015). Specifically, task complexity, structure/item type, or problem representation can vary in CPS tasks even if the subject domain is kept consistent. For instance, a task can be structured such that the problem is represented as a series of text-based multiple-choice questions, a simulation of a real-world activity, or an open-ended game-based environment, and these design variations can potentially evoke different CPS skills due to one or more factors. Beyond task structure, one factor could correspond to differing interests, manifested by different motivations to engage in CPS (Järvenoja et al., 2020). In particular, social loafing, a known problem contributing to motivational loss in small group work (Karau & Williams, 1993), might vary by task context. For example, students identified as Social Loafers in the electronics task in (Forsyth et al., 2020) might be more like the Active Collaborators if, for example, they had a particular interest in the domain (e.g., a student who is curious about biology but uninterested in electronics) or affinity for the type of task (e.g., interested in game-like environments, but uninterested in tasks represented as traditional multiple-choice items). Another factor could be the level of interdependence in the task design (Swiecki, 2021). For instance, a given task could have a high degree of task interdependence (i.e., the completion of some subcomponent of the task depends on the prior completion of another subcomponent), which may elicit different CPS skills than a task without this type of interdependence. Collaboration differences have also been noted in prior work with tasks that differ based on complexity and difficulty. Specifically, previous research has shown individuals engage in more collaborative behaviors while engaging in complex and difficult problems relative to simple and easy problems (Andrews-Todd & Toscano, 2020; Fernández et al., 2001; Gilabert et al., 2009). Other factors that might influence CPS skills and outcomes include the composition of the group, the size of the groups, whether the collaboration is face-to-face or computer-mediated, and so on (see reviews by Graesser, Fiore, et al., 2018; N. L. Kerr & Tindale, 2004).

In prior CPS assessment research, little attention has been given to exploring potential differences in how CPS skills are exhibited across task contexts, as most studies tend to explore CPS in the context of one task or academic domain. For example, prior work has concerned assessments that can assess broad, domain-general skills (Griffin et al., 2012; Hesse et al., 2015; OECD, 2013b; Stoeffler et al., 2020) while others have focused on a single domain-specific assessment (e.g., in the domain of science, reading, or mathematics; Kuo et al., 2020; Liu et al., 2015), with most focusing on one or multiple tasks without variation by task characteristics. Furthermore, little attention is given to such explorations with teams utilizing open, free-flowing dialogue, a common feature in everyday collaborative contexts, but see (Andrews-Todd & Forsyth, 2022; Sun et al., 2022) for exceptions. For example, the PISA 2015 CPS assessment included six units with multiple content-free CPS tasks where a student communicated with computer-simulated agents using predefined chat messages to solve problems (e.g., engaging in a contest to answer questions about a fictional country) (OECD, 2017). In the reporting of results from the PISA assessment, there was no focus on how CPS skills may have varied across the tasks when students responded to computer agents, though it was noted how the difficulty of items within a task may have required students to engage in different CPS skills (OECD, 2017). Other work has explored whether speech-based computational models of CPS skills can generalize across contexts (e.g., Pugh et al., 2022), though the focus is not typically on variation in CPS skills across the contexts, but rather on variations in language across contexts.

2.3. The current study

In the current study, we aim to contribute to the accumulating evidence concerning how CPS skills might vary by task context, particularly for computer-based tasks which are becoming commonplace for CPS assessments. Additionally, we aim to understand differences (between task contexts) in the relationship between these CPS skills and task performance. Exploration of individuals' CPS skills across task contexts can have important implications for CPS assessment. Though CPS in general might be independent of domain knowledge and other individual difference measures (e.g., intelligence, personality; Sun et al., 2020), certain situations may require different kinds of CPS skills to be displayed. For example, in some instances, some degree of prior domain knowledge might be needed to effectively instantiate CPS skills for a given domain. As such, certain task contexts may differentially evoke certain CPS skills. Likewise, certain CPS skills may be more strongly associated with success in one task context than another. These potential differences in displays of CPS skills across task contexts are important considerations that may influence design decisions when seeking to assess

CPS skills in online contexts.

Our primary contributions in this study are two-fold; (1) we explore associations between students' displays of CPS skills during computer-based online tasks that differ according to various task characteristics (RQ1), and (2) we investigate the extent to which different skills are related to performance outcomes across these online tasks (RQ2). We utilized two different online tasks, which differed across multiple dimensions: A more traditional worksheet-like mathematics problem solving task (the T-Shirt Math Task) and an open-ended game-based Physics simulation environment (Physics Playground). Based on results from prior studies (e.g., Forsyth et al., 2020; Sun et al., 2020), we expected there would be differences in the demonstration of CPS skills across tasks, but, given that we intentionally selected tasks that differed across numerous factors (format, subject domain, familiarity, engagement), the goal was not to compare CPS skills across tasks, but rather to examine associations across tasks. Thus, our primary research question was: What CPS skills do students engage in while completing two different online collaborative tasks, and how are the displays of the CPS skills associated across tasks? (RQ1).

Second, given our hypothesis that CPS skills may be differentially evoked across online tasks, we further expected there would be differences in which CPS skills were related to performance outcomes across the tasks. We made careful consideration in our task selection that the T-Shirt Math Task was explicitly selected to closely mirror the kinds of collaborative tasks used in secondary classrooms (e.g., content-based worksheets of problems) and Physics Playground was selected because it was a novel game-based environment supporting creative exploration. We hypothesized that the T-Shirt Math Task may rely more on prior content knowledge than the Physics Playground, so our expectation was that CPS skills related to sharing prior knowledge might be more predictive of performance on the mathematics task. Furthermore, we expected that the Physics Playground may rely more on procedural activities than the T-Shirt Math Task, so our expectation was that CPS skills related to problem solving processes might be better predictors of performance on the physics task. Thus, our second research question was: what CPS skills predict student CPS performance and how do they differ across the online tasks? (RQ2).

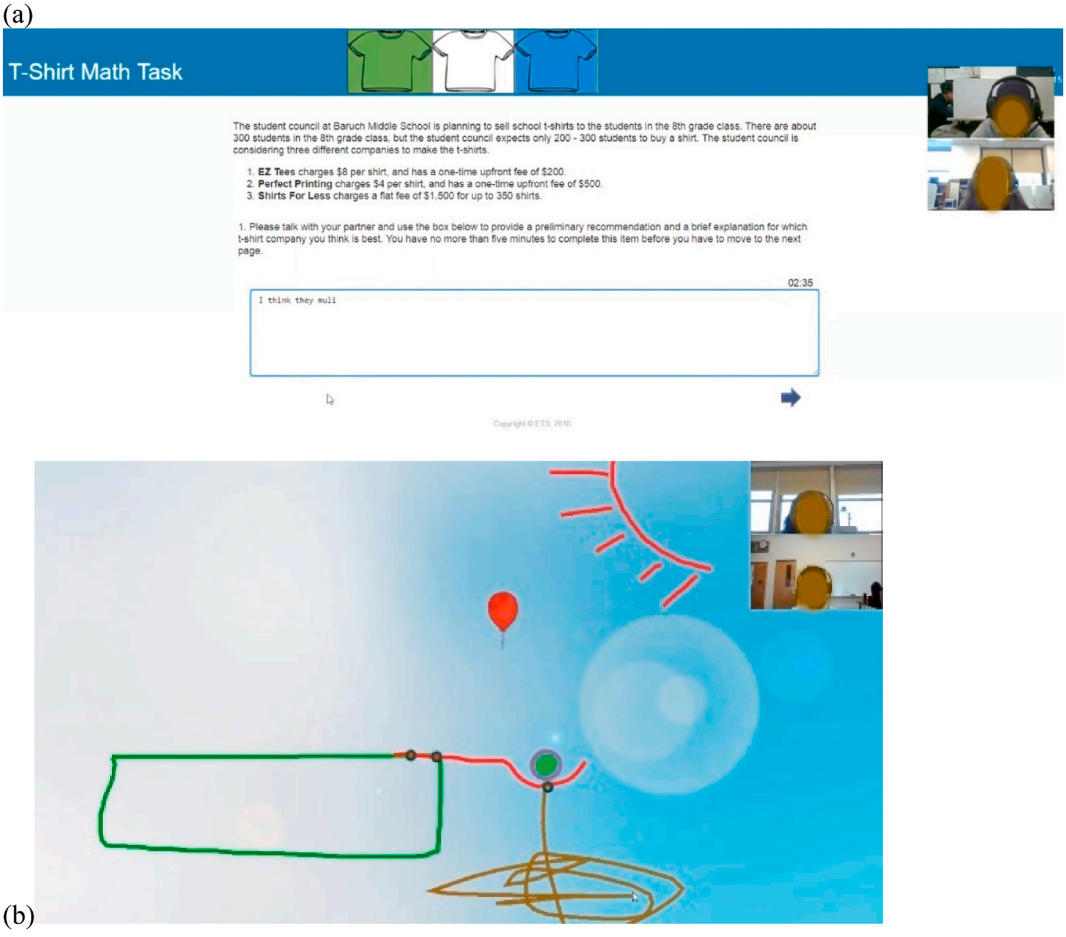


Fig. 1. Screenshot of a constructed response item from the T-Shirt Math Task (a) and the Sunny Day game level from Physics Playground (b).

3. Method

3.1. Participants

Data were obtained from 100 Northeastern U.S. students (82 from an urban school district, 18 through open recruitment outside of school) randomly assigned into teams of two. The majority of the participants were female (60%) and were very diverse by race/ethnicity (3% American Indian/Alaskan Native; 12% Asian; 24% Black; 23% Hispanic/Latino; 12% White; 15% Multi-Racial; 11% Other/Missing/Prefer not to Answer). There was a mix of participants by grade level: 7th grade (11%); 8th grade (30%); 9th grade (57%). All participants were ages 12–15: 12 (12%), 13 (30%), 14 (44%), 15 (13%). For analyses, due to attrition or insufficient quality of the resulting data collected (detailed below), some teams were removed for one or more tasks. The resulting analytical sample sizes were 82 students (41 teams) for the mathematics task, 78 students (39 teams) for the physics task, and 70 students (35 teams) who completed both tasks.

We used self-reported gender information to determine whether a team was comprised of students of the same gender (1) or mixed gender (0) as a measure of gender composition, as group gender composition has been found to influence collaborative behaviors (Andrews et al., 2017; Barrett & Lally, 1999; Fenwick & Neal, 2001; Prinsen et al., 2007). Based on available data, the relative proportion of same-gender teams were 58% (23 of 40 teams) for the T-Shirt Math Task, 61% (23 of 38 teams) for Physics Playground, and 56% (19 of 34 teams) across both tasks. Appendix A contains a complete summary of participant demographics.

3.2. Tasks

3.2.1. T-Shirt Math Task

The study consisted of two online tasks. The T-Shirt Math Task (Andrews-Todd et al., 2019) is a mathematics task concerning linear functions and argumentation. Students worked together on questions to determine which of three companies was the best choice to purchase t-shirts for classmates. Students compared the companies which have different variable and fixed costs to determine which to choose based on a particular number of t-shirts needed. The task comprised 10 questions that included constructed response items (see Fig. 1a), dropdown items, and multiple-choice items. Specifically, three multiple-choice items asked students to select the graph line associated with each company's cost equation. One dropdown item asked students to select problem variables corresponding to the $y = mx + b$ cost equation (e.g., y is the total number of t-shirts). Three additional dropdown items asked students to select the values corresponding to the equation for the costs associated with each company. Two constructed response items asked students to provide a recommendation for which company provided the best deal, one item at the beginning of the task and one item at the end of the task after students had the opportunity to complete the items previously listed. One final constructed response item asked students to provide a recommendation for the best t-shirt company given a new set of conditions (i.e., making a recommendation on the best company to choose when taking into account the percentage of students who ordered t-shirts in the previous year). This kind of task was selected to mimic the kinds of mathematics activities often implemented in classrooms for our targeted grade levels. We wanted to be able to compare such an activity to a more dynamic game-based activity like the one described next.

3.2.2. Physics Playground

Physics Playground (Shute et al., 2013) is an educational game supporting learning of Newtonian physics. Students were tasked with drawing objects (e.g., ramp, springboard, pendulum) to make a ball hit a balloon target (see Fig. 1b in which students are drawing a weight attached to a springboard to launch the ball towards the balloon). The pre-existing objects and the objects drawn by students in the game obeyed the laws of physics. Students started by completing a tutorial demonstrating how the game mechanics worked. Subsequently, students completed up to six game levels covering concepts associated with Newton's 1st Law, energy transfer, properties of torque, and conservation of linear momentum. The six game levels were split into two blocks. In the first block, the three levels, which increased in difficulty, included the Downhill level which required use of a ramp to solve and corresponded to concepts of Newton's 1st Law and energy transfer, the Yippie level which required a springboard to solve and corresponded to concepts of energy transfer and properties of torque, and the Scale level which required use of a lever to solve and corresponded to the concept of properties of torque. In the second block, the three levels, which again increased in difficulty, included the Through the Cracks level which could be solved using a ramp or lever and corresponded to concepts of Newton's 1st Law, energy transfer, and properties of torque, the Sunny Day level which could be solved using a lever, pendulum, or springboard and corresponded to concepts of energy transfer, properties of torque, and conservation of linear momentum, and the Little Mermaid level which required use of a springboard to solve and corresponded to the concept of energy transfer. Appendix C contains screenshots depicting each Physics Playground level.

3.2.3. Task characteristics

Both tasks were administered for remote collaboration via Zoom with screen sharing; however, the tasks differed according to how students were able to control the cursor. For the T-Shirt Math task, only one student could control the cursor at a time, but students could alternate control at will. For the Physics Playground task, one student was randomly chosen to control the cursor first, and after completing three levels (or after half the allotted time), control was switched for the remaining levels.

Both tasks meet many of the conditions that Szewkis et al. (2011) identified as necessary for a successful collaborative activity. Specifically, both tasks involved a common goal that team members needed to achieve (to answer items correctly or to complete game levels) through coordination and communication. Through screen sharing, actions performed by students with control were observed by their partners and thus they were accountable for these actions (i.e., individual accountability), and there was awareness among

members about the current state of their partners and the team. In both tasks, team members received joint outcomes or rewards when completing the tasks, though the rewards in Physics Playground were more salient (e.g., trophies awarded, levels completed – see 3.3.1) compared to the T-Shirt Math Task, where the outcomes were correctness on items and not directly shown to the members for all items. Partners submitting a joint answer and receiving joint outcomes or rewards (as opposed to submitting separate answers or receiving individual outcomes or rewards) indicated that positive interdependence exists in both tasks, while Physics Playground may have a higher level of positive interdependence than the T-Shirt Math task given its more salient rewards.

In all, the tasks differed across four main characteristics: academic domain (mathematics vs. physics), structure (traditional assessment items vs. game-based environment), problem representation (text vs. visual), and method of control (alternating at will vs. alternating at the halfway point). Specifically, the T-Shirt Math task was a mathematics task that was more like traditional assessments with respect to item types available (i.e., multiple-choice, constructed response). The problems were represented through text as shown in Fig. 1a and partners could control the cursor at will. On the other hand, Physics Playground was a physics task that featured an open game-based environment that supported creative exploration. Further, the problem was represented visually as shown in Fig. 1b and only one partner could control the cursor per level. Whereas these major differences afford an investigation into associations of CPS skills across tasks and whether the skills differentially predict CPS outcomes (our main research questions), they preclude direct comparisons of skills across tasks (which is an item for future work).

3.3. Measures

3.3.1. Task performance measures

The T-Shirt Math Task consisted of 10 questions (multiple-choice and open-ended responses), some having multiple parts. The 13 multiple choice questions were scored as incorrect (0) or correct (1), with item correctness as an item-level performance measure. The three open-ended questions were human scored on a scale of 0–3 points according to a pre-defined rubric based on whether the correct company was named and the strength of the explanation (none, weak, strong). The human scoring with three raters produced exact agreement from 72% to 94% across items and Fleiss kappa values from 0.75 to 0.95, all considered acceptable (Landis & Koch, 1977). Scores were summed across items and item parts to obtain a task-level measure, ranging from 0 to 22, which was rescaled to 0–1 for analyses.

For Physics Playground, performance was operationalized according to whether a gold or silver trophy was awarded on a game level. Students were awarded a gold trophy if they solved a level using a more optimal solution (drawing few objects to complete the level). Otherwise, students were awarded a silver trophy for completing a level. No trophy was awarded if students quit a level without solving it or if time ran out. At the task level, we assigned 2 points for a gold trophy, 1 point for a silver trophy, and 0 points for no trophy. These values were summed across the six game levels (range = 0–12) and then rescaled to 0–1 for analyses. As noted in Table 1, students were moderately successful at both tasks, more so for the T-Shirt Math Task.

3.3.2. Prior content knowledge measures

The mathematics pre-test consisted of two multiple-choice items and five true/false items on linear functions (see Fig. 2 for an example item). For physics, the pre-test consisted of four multiple-choice questions depicting scenarios in Physics Playground for students to solve (see Fig. 3 for an example item). For both pre-tests, each item was scored as correct (1) or incorrect (0). Total sum scores on the content assessments were converted to percent correct values for analyses. As shown in Table 1, prior knowledge was moderate (between 50% and 60%), but there was no difference in relative success across content pre-tests ($p = .40$).

3.3.3. Individual teamwork scale

An individual teamwork questionnaire (de la Torre-Ruiz et al., 2014) with seven items on a 7-point Likert response scale ranging from “Strongly Disagree” to “Strongly Agree” was collected as part of the pre-task surveys as an individual difference measure meant to serve as a covariate in relating observed CPS skills to task performance outcomes. Example items included, “I can work effectively in a group setting” and “I am able to resolve conflicts between individuals effectively.” A single overall score was computed from the average of a subset of five of the seven items. The others (“Given a choice, I would rather do a job where I can work alone than do a job where I have to work with others in a group” and “I like it when my work group does things on their own rather than working with others all the time”) did not effectively contribute to the overall score due to poor discrimination and were therefore removed. The Cronbach’s alpha reliability for the five retained items was .86. As noted in Table 1, students self-reported high teamwork values (i.e., scores >5).

Table 1
Descriptive statistics for external measures.

External Measures	T-Shirt Math Task (n = 82)		Physics Playground (n = 78)	
	Mean (SD)	Observed Range	Mean (SD)	Observed Range
Prior Knowledge (% Correct)	59.06 (18.34)	14.29–100.00	56.09 (25.22)	0.00–100.00
Individual Teamwork Scale	5.64 (0.82)	2.40–7.00	5.54 (1.01)	1.00–7.00
Team Gender (proportion same gender)	0.57 (0.50)	0–1	0.61 (0.49)	0–1
Task Performance (proportion maximum points)	0.64 (0.20)	0.32–1.00	0.46 (0.20)	0.08–1.00

Mitchell walks home from school at a steady rate. His sister Karyn rides her bike home from school at a steady rate. They both leave school at the same time, but Karyn is faster than Mitchell. Which graph best represents this information?

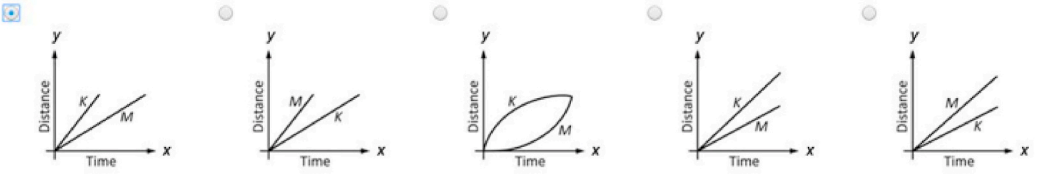
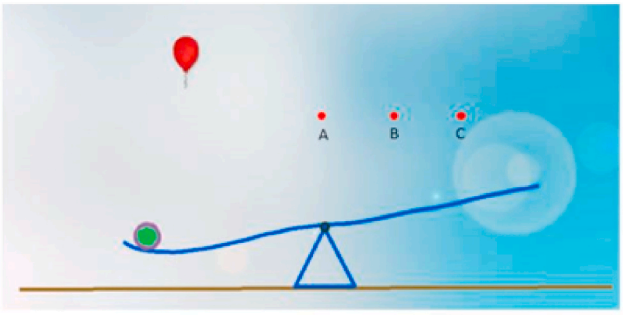


Fig. 2. Sample mathematics pre-test item with the correct response selected.



1. To launch the green ball up to the balloon, from which position should you drop a weight?

- Position A
- Position B
- Position C
- No difference between positions

Fig. 3. Sample physics pre-test item with the correct response selected.

3.4. Study procedure

Students were randomly assigned into teams of two and were seated at individual computer workstations in a computer lab. They first individually completed a series of pre-surveys (i.e., mathematics and physics pre-tests, questionnaire to obtain student demographics, teamwork survey). A researcher then set up Zoom to record students' screens, faces, and voices while students completed the two CPS tasks. The order of administration of the tasks was counterbalanced across teams such that half the teams received the mathematics task first whereas the others received the physics task first. Students were allotted around 30 min to complete each task, upon which they had to start on the second task. After students collaborated to complete the two tasks, they exited Zoom and individually completed a series of post-surveys on their respective computers; these data are not relevant for the purpose of the current paper and are not discussed further. Most students completed the study in a single session, but some students (with shorter class periods) completed the study in two sessions (i.e., Task 1 during session 1 and Task 2 during session 2).

3.5. Coding CPS skills

3.5.1. CPS ontology (framework)

We used a competency model represented as an ontology (similar to a concept map) (Andrews-Todd & Kerr, 2019) to code CPS skills. The ontology lays out concepts associated with the CPS competency, their relationships, and indicators or behaviors that demonstrate evidence of the skills in a principled way. The ontology was developed from discussions with subject matter experts and review of prior frameworks and literature in relevant areas such as computer-supported collaborative learning, linguistics, individual problem solving, and communication (Hesse et al., 2015; Liu et al., 2015; Meier et al., 2007; OECD, 2013a, 2013b; O'Neil et al., 1995). The ontology provides a comprehensive model of skills within the CPS competency that incorporate many of the components of existing frameworks while also incorporating additional skills or sub-skills hypothesized to be important for exploration of CPS, particularly in open environments with free flowing dialogue.

The ontology includes nine high-level CPS skills across social (collaboration, teamwork) and cognitive (problem solving, taskwork) dimensions. Each of the skills include behaviors, or indicators associated with what individuals do or say that provide evidence of the

skills. In the social dimension, *maintaining communication* corresponds to content-irrelevant social-oriented communication (Lipponen et al., 2003; Liu et al., 2015). *Sharing information* includes content-relevant information used to solve the problem. *Establishing shared understanding* refers to communication used to learn others' perspectives and ensure that what has been said is understood (Clark, 1996; Clark & Brennan, 1991). This corresponds to behaviors used to understand team members. *Negotiating* corresponds to language used to communicate if conflicts exist and resolve those conflicts. In the cognitive dimension, *exploring and understanding* corresponds to communication and actions used to explore the task environment and understand the problem (Frensch & Funke, 1995; OECD, 2013a). *Representing and formulating* includes communication used to build a mental representation of the problem and formulate hypotheses (Mayer & Wittrock, 1996; VanLehn, 1996). *Planning* refers to communication used to develop a strategy for solving the problem (Cohen, 1989; Hesse et al., 2015; Wirth & Klieme, 2003). *Executing* corresponds to actions (e.g., typing text) and communication used to carry out a plan (OECD, 2013a; Wirth & Klieme, 2003). Monitoring refers to actions and communication used to monitor progress toward the goal and team organization (OECD, 2013a, 2013b; O'Neil, 1999).

These nine high-level skills can be grouped into the following three facets. *Communicative participation* includes the behaviors that correspond to general participation in the communication among teammates (e.g., exchange of information) or activity that individuals engage in to remain part of the conversation (Isohätälä et al., 2020), namely the skills of maintaining communication, sharing information, and establishing shared understanding. *Social regulation* corresponds to behaviors individuals engage in to address the diversity of perspectives, experiences, and expertise among teammates as well as monitoring the organization of the team and evaluating progress (Hadwin et al., 2018; Hesse et al., 2015; Janssen et al., 2012; Järvelä et al., 2013; Lobczowski et al., 2020). This facet includes the skills negotiating and monitoring. *Task regulation and activity* is associated with the behaviors used to manage, coordinate, and enact task activities (Hesse et al., 2015; Janssen et al., 2012). It includes the skills exploring and understanding, representing and formulating, planning, and executing. Table 2 provides an overview of the CPS skills and facets with representative examples (or behaviors) from the mathematics and physics tasks.

3.5.2. Qualitative coding

Video recordings of students' collaboration on both tasks were segmented into turns, transcribed, and coded at the turn level by three trained raters. For each turn of talk or activity by an individual, the raters labeled the turn as one of the nine CPS skills from the ontology. Coding was carried out using Dedoose qualitative analysis software (Dedoose, 2018) which supported use of video and audio recordings for rater coding. To establish interrater reliability, all three raters coded 20% of the videos. For the T-Shirt Math Task across 1409 turns, the median intraclass correlations (ICCs) across the CPS skill ratings was 0.93 (range across skills = 0.68–1.00). For Physics Playground across 1374 turns, the median ICC across CPS skill ratings was 0.90 (range across skills = 0.43-0.99). Across both tasks (2783 turns), the median ICC across CPS skill ratings was 0.92 (range across skills = 0.61-0.99), indicating excellent agreement (Cicchetti, 1994). Once interrater reliability was established, the remaining videos were divided among the three raters for independent coding. A total of 10,239 turns were coded across 80 videos, with each video corresponding to a team completing one task (time range = 3.12–34.3 min). There was an average of 128 turns per video (SD = 70.5). We coded 41 T-Shirt Math Task videos (time range = 8.95–34.34 min) with an average of 133 turns per video (*SD* = 65.6) and 39 Physics Playground videos (time range = 3.12–29.21 min), with an average of 123 turns (*SD* = 76.7).

Once the segmented turns were coded, we calculated the frequency of each CPS skill across turns for each item (T-Shirt Math Task) or game level (Physics Playground) and for each task for later analysis. We also summed the frequencies of skills that comprised each facet. Finally, we calculated the number of words across turns of conversations each student generated within an item as an indicator of student verbosity during the task. A 99% winsorization was applied to the CPS skill frequency measures and the verbosity measure with the top 1% extreme values of each measure being replaced by the value at the 99th percentile to exclude extreme cases.

Table 2
CPS skills and facets with corresponding examples from the mathematics and physics tasks.

CPS Facet	CPS Skill	T-Shirt Math Task Example	Physics Playground Example
Communicative Participation	Maintaining Communication	"That's great"	"This game is funny"
	Sharing Information	"m will be the price for one shirt"	"The lever is the scoop"
	Establishing Shared Understanding	"Why do you think it's the third one and not the second one?"	"Alright, what do you think we should do here?"
Social Regulation	Negotiating	"No, no. 'Cause 350 is the number of shirts. Our number of shirts would be X. The cost per shirts would be the number before X."	"No don't make it yet it's going to start swinging"
	Monitoring	"we gotta hurry up"	"Are you ready?"
Task Regulation and Activity	Exploring and Understanding	Reading instructions or problem quietly to self	N/A (task affordances available make it difficult to distinguish from Executing during game play)
	Representing and Formulating	"We're basically just putting it into slope-intercept-form"	"This is the same thing of what we had with the ramp"
	Planning	"Let's plug in a number between 200 and 300 and solve for Y″	"Let's make a springboard with a mass under it and delete the mass"
	Executing	"Alright I'll rewrite the explanation"	"Put more support in the beginning so it doesn't snap"

4. Results

4.1. CPS skills exhibited within and across tasks (RQ1)

Table 3 provides descriptive information for the CPS skills and facets across individual students that emerged during the two tasks. For the T-Shirt Math Task, all skills in the CPS ontology were represented. The most frequently occurring skills for this task were Sharing Information (31.6%), Establishing Shared Understanding (19.6%), Negotiating (14.5%), and Executing actions (14.4%). With respect to the facets, Communicative Participation was displayed the most (55.8%) followed by Task Regulation and Activity (24.1%) and Social Regulation (20.0%). For Physics Playground, all possible CPS skills were represented. The most frequently occurring skills for this task were Establishing Sharing Understanding (25.6%), Executing communication (21.8%), Sharing Information (13.7%), and Negotiating (13.2%). Turning to the facets, Communicative Participation (46.4%) was displayed the most followed by Task Regulation and Activity (35.1%) and Social Regulation (18.6%).

We examined Spearman correlations to explore associations among the aggregate frequencies for skills/facets within each task (Table B1 in Appendix B) and critically across tasks (see Table 3). There was a wide range, such that Negotiating (0.56), Sharing Information (0.47), and Planning (0.46) exhibited the most consistent relationships between tasks, while Executing actions and Maintaining Communication were more divergent (0.08 and 0.10, respectively). Overall, the median correlation of 0.28 indicates moderate associations among the individual skills. At the facet level, correlations between tasks were stronger: Social Regulation (0.59); Task Regulation and Activity (0.46); Communicative Participation (0.36), resulting in a median correlation of 0.46.

4.2. CPS facets and performance outcomes (RQ2)

Student performance on the T-Shirt Math Task and Physics Playground (among teams who completed both tasks) was marginally significantly correlated ($\rho = 0.31$, $p = .07$). We constructed linear mixed effects models to explore the relationships between the CPS facets and CPS performance outcomes and whether they differed across tasks (RQ2). We did not focus on individual skills on performance outcomes because the frequencies of many skills were highly correlated with each other (see Table B1). Since performance was a team variable, the dependent variable in the models was item-level performance outcome for a team, specifically the correctness on each item in the T-Shirt Math Task, and whether a trophy was awarded in a game level in Physics Playground (see 3.3.1). We jointly entered the frequency of each CPS facet an individual displayed within an item as fixed factors, with random intercepts for items and teams; more complex random effects structures resulted in convergence errors. Teams' mean verbosity (with a square root transformation), prior content knowledge, self-reported teamwork, and team gender composition were included as covariates (see 3.3). Duration (i.e., time spent on each item/level) was not included as a covariate considering its correlation with verbosity (Spearman correlations were 0.45 and 0.67 for T-Shirt Math Task and Physics Playground, respectively). All predictors and covariates were standardized before being included in the models.

The descriptive statistics and correlations for the variables included in the models can be found in Appendix B. The model results, including the odds ratios (*OR*s) and *p*-values, are in Table 4. The odds ratios represent the ratios of the likelihood of receiving a trophy in a Physics Playground level or answering a T-Shirt Math Task item correctly associated with a one-unit increase in the corresponding predictors. Values greater than 1 indicate a positive relationship between the predictor and the outcome variable, while values less than 1 indicate a negative relationship.

Results indicated that a unit increase in Social Regulation significantly increased the odds of success on an item by 64% in the T-Shirt Math Task. A team's mean prior knowledge in math was also significantly associated with performance on the T-Shirt Math Task. For Physics Playground, a unit increase in the team's mean teamwork scale was significantly associated with a 101% higher likelihood

Table 3
Summary of reported aggregate individual CPS skills by facet and task across participants.

CPS Facets and Skills	T-Shirt Math Task (N = 5312)			Physics Playground (N = 4649)			ρ
	n (%)	Mean (SD)	Range	n (%)	Mean (SD)	Range	
Communicative Participation	**2966 (55.8)**	**36.2 (21.3)**	**0–84**	**2155 (46.4)**	**27.6 (21.4)**	**0–104**	**.36****
Maintaining Communication	246 (4.6)	3.0 (5.3)	0–32	329 (7.1)	4.2 (5.3)	0–33	.10
Sharing Information	1679 (31.6)	20.5 (12.8)	0–51	635 (13.7)	8.1 (6.5)	0–33	.47**
Establishing Shared Understanding	1041 (19.6)	12.7 (9.2)	0–55	1191 (25.6)	15.3 (15.3)	0–78	.31*
Social Regulation	**1065 (20.0)**	**13.0 (9.1)**	**0–37**	**863 (18.6)**	**11.1 (9.9)**	**0–43**	**.59****
Negotiating	770 (14.5)	9.4 (7.6)	0–33	613 (13.2)	7.9 (7.7)	0–36	.56**
Monitoring	295 (5.6)	3.6 (3.4)	0–14	250 (5.4)	3.2 (3.8)	0–18	.24*
Task Regulation and Activity	**1281 (24.1)**	**15.6 (9.7)**	**1–45**	**1631 (35.1)**	**20.9 (12.6)**	**0–62**	**.46****
Planning	227 (4.3)	2.8 (3.6)	0–18	189 (4.1)	2.4 (3.1)	0–12	.46**
Executing actions	767 (14.4)	9.4 (6.3)	0–24	405 (8.7)	5.2 (3.1)	0–14	.08
Executing communication	202 (3.8)	2.5 (3.0)	0–14	1014 (21.8)	13.0 (10.2)	0–47	.28*
Exploring and Understanding	61 (1.1)	0.7 (1.2)	0–5	–	–	–	–
Representing and Formulating	24 (0.5)	0.3 (0.7)	0–4	23 (0.5)	0.3 (0.6)	0–4	.19

Notes: **$p < .01$; *$p < .05$. Spearman correlations (ρ) of CPS facets and skills between tasks among students who completed both tasks (n = 70) were reported. Some utterances (T-Shirt Math Task = 140 or 2.6%; Physics Playground = 138 or 2.9%) were deemed not interpretable by human transcribers and are not reflected in the frequencies above.

Table 4
Using CPS facets to predict performance at the item level.

	T-Shirt Math Task		Physics Playground	
	Odds Ratio	*p*	*Odds Ratio*	*p*
Predictors				
Communicative Participation	1.09	0.785	0.73	0.553
Social Regulation	**1.64**	0.042 *	1.44	0.473
Task Regulation and Activity	0.84	0.436	0.49	0.273
Covariates				
Verbosity	1.09	0.830	1.15	0.865
Prior Knowledge	**1.71**	0.030 *	1.43	0.329
Teamwork Scale	0.91	0.697	**2.01**	0.046 *
Team Gender Composition	0.74	0.530	2.29	0.252
Random Effects				
σ^2	3.29		3.29	
τ_{00}	1.20_{Team}		1.88_{Team}	
	6.45_{Item}		2.94_{Item}	
ICC	0.70		0.59	
N	10_{Item}		6_{Item}	
	39_{Team}		35_{Team}	
Observations	376		167	
Marginal R^2/Conditional R^2	0.062/0.718		0.109/0.639	

Notes. $^{**}p < .01$; $^{*}p < .05$; . $P < .1$. Odds ratios (*OR*s) were reported.

of receiving a trophy on a level. The other CPS facets and other covariates were not significantly related to performance in these tasks. All variables in the T-Shirt Math Task model had variance inflation factor (VIF) values lower than 2; VIFs were <4 for Physics Playground. Both were well below the threshold of 10, indicating no multicollinearity issues (Neter et al., 1996).

5. Discussion

Collaborative problem solving is a critical competency for many contexts, as many of today's problems require groups of individuals to come together to find solutions. This has motivated increased interest in work dedicated to assessing and developing CPS skills. There has been limited attention in past CPS assessment work on how CPS skills are exhibited across different online tasks. Prior work has tended to focus on designing assessments that can capture broad, domain-general skills or skills relevant for a specific domain (e.g., science, reading, or mathematics), with little work exploring the use of multiple tasks that vary by academic domain or other task characteristics. Therefore, we have a limited understanding of how the display of individuals' CPS skills is associated across different tasks, the implications this might have for the design of digital assessment and learning environments, and how we might interpret individuals' skills across different task contexts.

In the current study, we explored the types of CPS skills students displayed in an online mathematics task and a physics educational game, how those skills were associated across the two online tasks (RQ1), and whether they differentially predicted task performance (RQ2). We found that students displayed a range of skills in both tasks, with students exhibiting all CPS skills represented in the framework on both tasks. Notably, two of the top three most frequently occurring CPS skills were the same in both tasks (i.e., Sharing Information and Establishing Shared Understanding). The overall results in terms of the specific CPS skills frequently observed in the tasks (Table 3) are in line with prior work reporting sharing information (Andrews-Todd & Forsyth, 2020; Hao et al., 2016), negotiating (Sun et al., 2022), establishing shared understanding (Andrews-Todd & Forsyth, 2020, Sun et al., 2022; Rosen, 2014), and executing communication (e.g., giving instructions to teammates for steps to take in solving the problem; Sun et al., 2022) as frequently occurring skills. This finding is notable because these skills occurred frequently across both tasks in our study (Table 3), despite intentionally choosing tasks that differ in multiple dimensions (see 3.2). This suggests that these skills may play an important role in CPS regardless of the task characteristics.

Turning to our first research question, we found moderate associations among CPS skills across the two tasks with Negotiating (r = 0.56), Sharing Information (r = 0.47), and Planning (r = 0.46) being more strongly associated. However, other CPS skills (e.g., Monitoring communication, Executing actions) were not significantly associated. This finding suggests that while some CPS skills may be important and occur frequently across a broad range of tasks, other skills may be differentially elicited in different tasks, and their occurrence may be dependent on specific task characteristics. Future research could more carefully control for differences between tasks in order to investigate the relationship between specific task characteristics and the CPS skills they elicit. Additionally, the correlations for the three facets across tasks were all significant at moderate magnitudes (median = .46, range = 0.36-0.59). Each of the three individual skills just mentioned were the most influential to driving those facet-level relationships (i.e., Negotiating under Social Regulation, Sharing Information under Communicative Participation, and Planning under Task Regulation and Activity). This finding is informative, as it suggests there may be stronger associations across tasks for less fine-grained skills (i.e., higher level facets of CPS) than for more specific, fine-grained skills.

For our second research question, we examined which CPS facets predict student performance in the tasks and how those results might differ between tasks. Results revealed that for Physics Playground, none of the CPS facets significantly predicted performance.

However, we did find a significant association between the team's self-reported teamwork and performance such that higher ratings on the teamwork scale were associated with better performance outcomes. Physics Playground likely requires a higher level of collaboration and is more conducive to collaboration than the T-Shirt Math Task partly because of its design features (e.g., an open-ended, game-based environment). In fact, we do find higher levels of verbosity in Physics Playground than in the T-Shirt Math Task (though the difference is not significant). The features of Physics Playground may make it more interesting or motivating than the T-Shirt Math Task. This is consistent with prior work suggesting that digital games can provide interesting and challenging environments to support learning (Kaimara et al., 2022), and they can enhance students' motivation (Chen & Hwang, 2014; Hwang & Wu, 2012). As we mentioned in section 3.2.3, although both tasks have the positive interdependence element, the rewards in Physics Playground are more salient, possibly leading to a higher level of positive interdependence in this task compared to the T-Shirt Math task. The higher positive interdependence could also explain the significant association between self-reported teamwork and performance and the higher levels of verbosity in the Physics Playground task.

For the T-Shirt Math Task, results revealed that increases in skills associated with the Social Regulation facet were positively associated with higher likelihood of task success. This facet includes behaviors such as resolving conflicts and making sure teammates are following established roles. This result is consistent with prior work that suggests the importance of particular negotiation behaviors in supporting positive performance outcomes (Sun et al., 2022; Hao et al., 2019; Zhang et al., 2022). Furthermore, prior work has shown the benefits of shared regulation behaviors in contributing to successful collaboration and promoting the development of CPS skills (Järvelä et al., 2013; Rojas et al., 2022). Our expectation was that more behaviors associated with Communicative Participation (e.g., sharing prior knowledge) would contribute to performance outcomes for this task. Along those lines, we did find that higher prior knowledge was associated with better performance on the task. The design features of the T-Shirt Math Task being more like a traditional assessment (e.g., multiple-choice and constructed response items) could have contributed to this relationship. Furthermore, previous research has shown that prior knowledge can positively influence knowledge development in collaborative contexts (Gijlers & De Jong, 2005; Nokes-Malach et al., 2012; Zambrano et al., 2019), and our work may contribute to this understanding. Importantly, the fact that different results were obtained for the two tasks is illuminating and suggests that there may not be a one-size-fits-all mapping between CPS facets and outcomes, but that the relationship is likely moderated by the task characteristics.

5.1. Limitations and future work

Our findings should be interpreted in light of some limitations. First, our sample size was relatively small (n = 100). This limitation was further exacerbated by challenges with in-classroom data collection (i.e., data quality issues, student attrition), which reduced the sample size of teams who completed both tasks to 70 students (35 teams). Future work will need larger sample sizes to further investigate the relationship between displayed CPS skills and performance across tasks that differ according to varying characteristics, enabling statistical comparisons with greater power.

Additionally, our study design made it difficult to disentangle academic domain and task design. In particular, the two tasks used in the present study were designed to be different on multiple dimensions, such as task structure (high in T-Shirt Math Task, low in Physics Playground) and task type and representation (mathematics items with text and static images versus physics game with dynamic visualizations). Some of these differences were representative of differences in the tasks typically used in secondary mathematics versus physics classrooms (e.g., highly structured content-based worksheets of problems are often used in mathematics classrooms while more open-ended games supporting creative exploration are more frequently used in physics classrooms). As discussed above, we intentionally selected tasks that were quite different in order to explore associations with CPS skills/facets across very different task contexts. However, this precludes more fine-grained investigations into each dimension. Future work should explore designs that control for various task affordances and academic domain across tasks to investigate the extent to which differences observed are due to the domain, task characteristics, or some combination. With such controls in place, it would also be beneficial to further explore the extent to which prior knowledge influences student interactions in such activities. The reason for this is that even with research showing relationships between specific prior knowledge and outcomes related to specific collaborative tasks (e.g., Forsyth et al., 2020; Gijlers & De Jong, 2005; Nokes-Malach et al., 2012; Uz-Bilgin et al., 2020; Zambrano et al., 2019), there is still the underlying notion that CPS in general may in fact be independent of prior knowledge (Sun et al., 2020). This may serve as a support for why an effect was observed for the mathematics task, but not the physics task.

We also only studied the relationship between CPS skills and objective outcome measures of task performance (i.e., levels solved, questions answered correctly). It is possible that other skills may differentially contribute to subjective outcomes (e.g., self-rated perceptions of collaboration quality), which should be investigated in future work. Given our research questions pertaining to task differences, we focused on examining relationships among the frequencies of CPS skills across tasks. However, because such a coding-and-counting approach does not support modeling temporal relationships among coded data (Zhu et al., 2020; Csanadi et al., 2018), future work could explore other approaches that account for temporality such as epistemic network analysis (Zhang et al., 2022) or multivariate vector autoregression (Zhou et al., 2022).

Future research should also include other populations in analyses. Although our sample was relatively diverse in terms of race and gender (see 3.1), all data were collected from students aged 12–15 in the Northeastern United States (most from an urban area). Future research should examine the generalizability of our findings to younger (e.g., elementary students) and older (e.g., undergraduate students) populations, as well as populations from different regions (e.g., rural areas) and countries. In the current study, participants did not show any statistically significant differences in prior knowledge as a function of their age (Physics: $F(3, 66) = 0.40$, $p = .757$; Math: $F(3, 74) = 0.37$, $p = .775$) or grade level (Physics: $F(2, 67) = 0.25$, $p = .783$; Math: $F(2, 75) = 1.67$, $p = .196$). However, it is possible that differences in age or maturity level could have an impact on how team members collaborate. Future work that explores

wide age ranges with larger sample sizes could incorporate variables in the models that account for students' maturity level or developmental stage to examine these possibilities.

Future work could also investigate the influence of topic/situational interest on CPS skills manifested by students in the two tasks. Aside from interest, other moderators such as group composition, psychological diversity, and other aspects of the task and domain could be considered.

6. Conclusions

Our study investigated associations among students' displayed CPS skills across two distinct online tasks. We found that the CPS skills represented in our framework were moderately associated across two rather different tasks suggesting a degree of convergence. Additionally, we studied how the frequency of CPS facets was predictive of task performance. We found a degree of divergence such that after controlling for several correlates, the same CPS facets did not predict CPS outcomes across tasks. Overall, our findings highlight the importance of incorporating multiple tasks during CPS assessments and can inform future research on CPS task design and computer-based CPS assessment.

Credit author statement

Jessica Andrews-Todd: Conceptualization, Investigation, Supervision, Funding acquisition, Data Curation, Writing - Original Draft, Writing - Review & Editing; Yang Jiang: Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing; Jonathan Steinberg: Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing; Samuel L. Pugh: Writing - Original Draft, Writing - Review & Editing; Sidney K. D'Mello: Methodology, Formal analysis, Funding acquisition, Writing - Original Draft, Writing - Review & Editing.

Author note

Jessica Andrews-Todd, Learning and Assessment Foundations and Innovation Center, Educational Testing Service; Yang Jiang, Learning and Assessment Foundations and Innovation Center, Educational Testing Service; Jonathan Steinberg, Foundational Psychometric and Statistical Research, Educational Testing Service; Samuel L. Pugh, Institute of Cognitive Science & Department of Computer Science, University of Colorado Boulder; Sidney K. D'Mello, Institute of Cognitive Science & Department of Computer Science, University of Colorado Boulder.

Data availability

Data will be made available on request.

Acknowledgements

This material is based upon work supported by the Institute of Education Sciences under Grant R305A170432 awarded to the first and last authors and by National Science Foundation awards DRL 2019805 and DUE 1745442/1660877 awarded to the last author. The opinions expressed are those of the authors and do not necessarily represent the views of the funding agencies.

Appendix A. Summary of Participant Sample Demographics

	Urban District (n = 82)		Open Recruitment (n = 18)		Total (n = 100)	
	N	%	N	%	N	%
Gender						
Female	51	62.2	9	50.0	60	60.0
Male	29	35.4	9	50.0	38	38.0
Missing	2	2.4	0	0.0	2	2.0
Race/Ethnicity						
American Indian/Alaskan Native	2	2.4	1	5.6	3	3.0
Asian/Asian American	2	2.4	10	55.6	12	12.0
Black/African American	24	29.3	0	0.0	24	24.0
Hispanic/Latino	23	28.0	0	0.0	23	23.0
White	11	13.4	1	5.6	12	12.0
Multi-Racial	12	14.6	3	16.7	15	15.0
Other/No Answer/Prefer Not to Answer	6	7.3	3	16.7	9	9.0
Missing	2	2.4	0	0.0	2	2.0
Grade Level						
7th	7	8.5	4	22.2	11	11.0

(*continued on next page*)

(*continued*)

	Urban District (n = 82)		Open Recruitment (n = 18)		Total (n = 100)	
	N	%	N	%	N	%
8th	45	54.9	5	27.8	50	30.0
9th	28	34.1	9	50.0	37	57.0
Missing	2	2.4	0	0.0	2	2.0

Appendix B. Descriptive Statistics for Model Variables

Table B1
Spearman correlations of reported aggregate individual CPS codes by facet and task across participants

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Communicative Participation	**.36****	.37**	.72**	.92**	.55**	.52**	.36**	.51**	.25*	.52**	.37**	–	.33**
2. Maintaining Communication	.45**	**.10**	.18	.16	.10	.12	.18	.07	.04	.33**	.07	–	−.13
3. Sharing Information	.89**	.21	**.47****	.56**	.64**	.57**	.62**	.50**	.46**	.25*	.48**	–	.27*
4. Establish Shared Understanding	.81**	.17	.56**	**.31***	.44**	.51**	.24*	.46**	.22	.50**	.40**	–	.42**
5. Social Regulation	.68**	.20	.59**	.65**	**.59****	.93**	.68**	.68**	.50**	.08	.63**	–	.15
6. Negotiating	.55**	.02	.48**	.60**	.94**	**.56****	.42**	.68**	.39**	.16	.71**	–	.15
7. Monitoring	.64**	.45**	.54**	.47**	.60**	.29**	**.24***	.37**	.59**	.00	.32	–	.18
8. Task Regulation and Activity	.60**	.09	.59**	.52**	.46**	.36**	.45**	**.46****	.58**	.29*	.93**	–	.26*
9. Planning	.21**	−.01	.59**	.51**	.52**	.52**	.28*	.76**	**.46****	.13	.39**	–	.26*
10. Executing actions	.19	.01	.21	.14	.10	.02	.24*	.73**	.34**	**.08**	.08	–	.13
11. Executing communication	.49**	.14	.40	.48**	.35**	.29**	.30**	.58**	.50**	.07	**.28***	–	.16
12. Exploring and Understanding	.14	.11	.12	.09	.22*	.11	.33**	.16	−.03	.00	.06	**-**	–
13. Representing and Formulating	.30**	.01	.28*	.31**	.25*	.21	.21	.19	.01	.07	.07	.20	**.19**

Notes: $^{**}p < .01$; $^{*}p < .05$. Values below the diagonal refer to the T-Shirt Math Task (n = 82); values above the diagonal refer to Physics Playground (n = 78); values on the diagonal (boldface) reflect correlations between the tasks (n = 70).

Table B2
Correlations across tasks and participants for CPS skills and model external variables

Code	Math T-Shirt Task				Physics Playground			
	Task Perf.	Pre-Test	Team-work	Team Gender	Task Perf.	Pre-Test	Team-work	Team Gender
Executing actions	.00	.12	.06	−.02	−.02	−.21	−.06	−.19
Executing (in chat)	.25*	.26*	.00	.15	.13	.02	.10	−.18
Exploring and Understanding	−.18	−.16	.09	.16	NA	NA	NA	NA
Monitoring (in chat)	.15	.02	.20	.24*	.39**	.20	.17	.10
Planning	.34**	.24*	.22	.31**	.46**	.08	.05	.04
Representing and Formulating	.10	−.03	−.01	−.07	.16	−.03	.03	.08
Establish Shared Understanding	.25*	.11	.16	.20	.17	−.07	.01	−.01
Maintaining Communication	−.10	.00	.05	.05	.06	.06	.04	.00
Negotiating	.40**	.34**	.25*	.18	.18	−.19	.07	−.03
Sharing Information	.39**	.19	.19	.34**	.45**	.00	.14	.08

Note: $^{**}p < .01$; $^{*}p < .05$. Pearson correlations are provided for all measures except gender which is dichotomous and for which a Spearman correlation is more appropriate.

Table B3a
Average display of CPS facets within each item for the T-Shirt Math Task across participants

Item	Verbosity	Communicative Participation	Social Regulation	Task Regulation and Activity	Task Performance (proportion Max Pts)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
1	129.62 (97.84)	6.92 (4.13)	2.58 (2.01)	3.07 (2.67)	0.27 (0.45)
2	58.73 (56.04)	4.11 (3.05)	1.54 (1.64)	1.57 (1.64)	0.61 (0.49)
3	27.67 (22.93)	2.35 (1.95)	0.78 (0.79)	1.08 (1.04)	0.92 (0.28)
4	58.90 (47.10)	3.96 (2.92)	1.21 (1.41)	1.30 (1.22)	0.46 (0.50)
5	14.17 (10.63)	1.38 (1.00)	0.45 (0.78)	0.84 (0.80)	1.00 (.00)
6	31.34 (30.38)	2.46 (2.04)	0.81 (0.86)	0.84 (0.86)	0.77 (0.42)
7	57.43 (85.71)	3.80 (4.35)	1.67 (1.69)	1.83 (1.89)	0.74 (0.44)
8	12.54 (10.31)	1.34 (0.99)	0.54 (0.66)	0.68 (0.70)	0.99 (0.12)
9	118.35 (114.68)	7.30 (5.70)	2.81 (2.89)	2.93 (2.98)	0.27 (0.45)
10	101.63 (97.72)	7.54 (5.13)	2.17 (2.45)	2.46 (2.43)	0.29 (0.46)

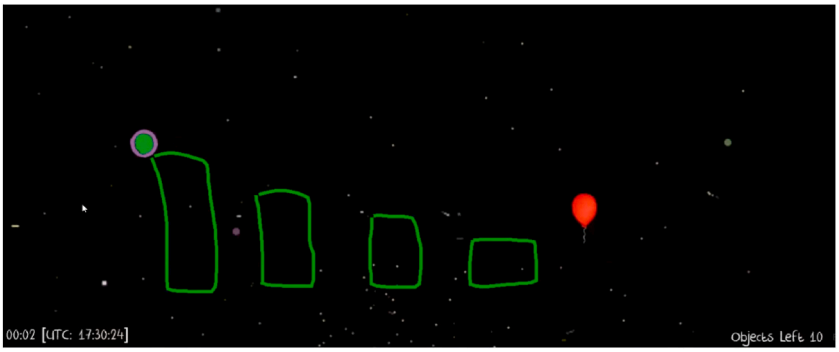
Table B3b
Average display of CPS facets within each level for Physics Playground across participants

	Verbosity	Communicative Participation	Social Regulation	Task Regulation and Activity	Task Performance (proportion Max Pts)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Down Hill	34.27 (35.11)	2.19 (3.15)	1.33 (1.59)	2.13 (1.85)	1.00 (.00)
Yippie	126.51 (129.85)	7.55 (8.15)	3.11 (3.44)	5.40 (4.05)	.68 (.47)
Scale	51.53 (59.60)	3.57 (3.39)	1.27 (1.30)	2.35 (2.09)	.90 (.31)
Cracks	64.38 (64.00)	3.52 (3.52)	1.00 (1.04)	3.16 (2.56)	.92 (.27)
Sunny Day	182.77 (177.55)	9.49 (8.84)	4.20 (3.67)	6.84 (4.93)	.49 (.50)
Little Mermaid	157.02 (142.71)	10.98 (8.38)	3.33 (3.27)	6.58 (4.22)	.52 (.51)

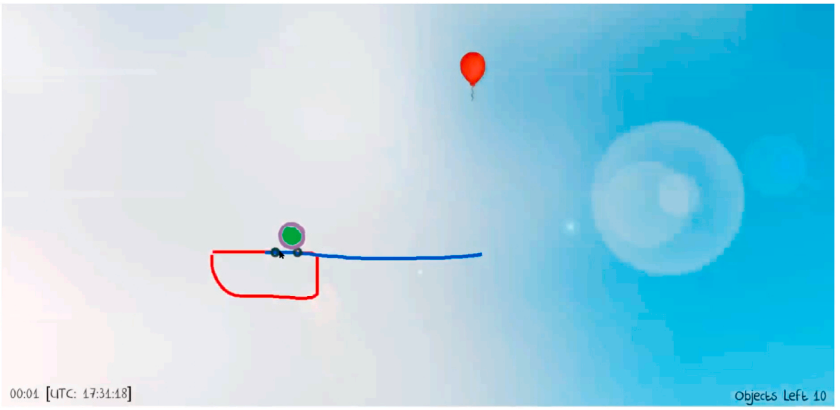
Appendix C. Screenshots of Physics Playground Levels

Block 1 Levels

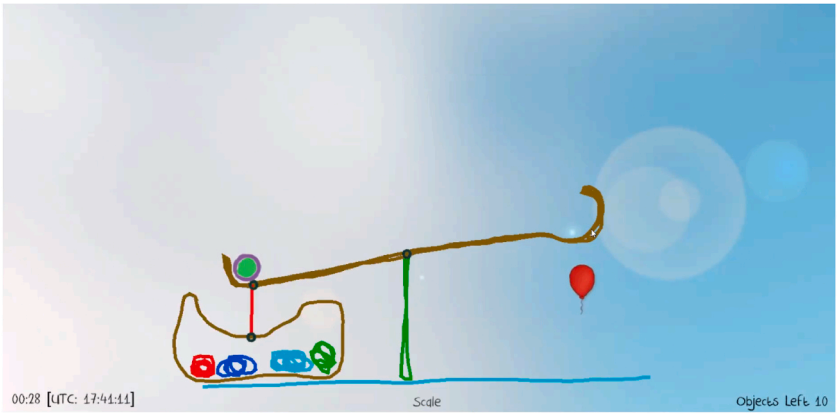
Downhill Level.



Yippie Level.

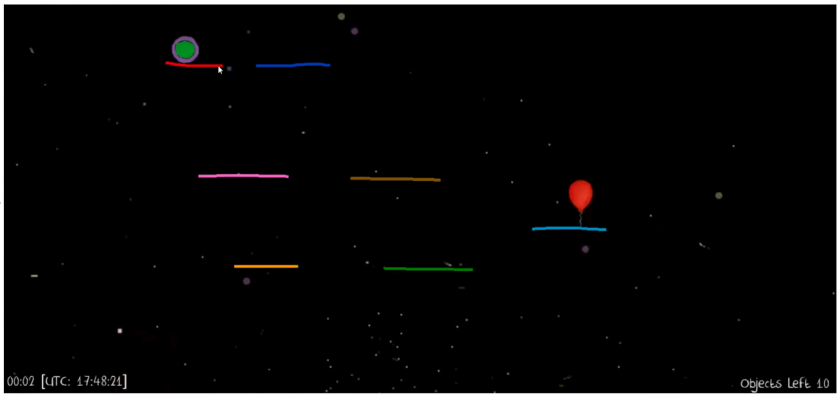


Scale Level.

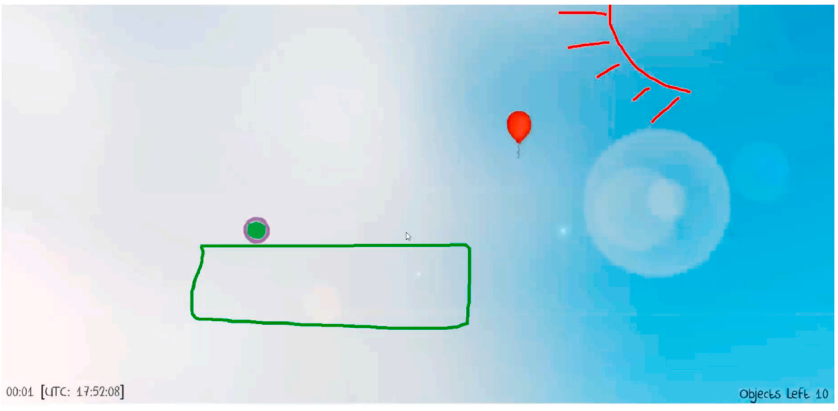


Block 2 Levels

Through the Cracks Level.



Sunny Day Level.



Little Mermaid Level.

References

Aguado, D., Rico, R., Sánchez-Manzanares, M., & Salas, E. (2014). Teamwork competency test (TWCT): A step forward on measuring teamwork competencies. *Group Dynamics: Theory, Research, and Practice, 18*(2), 101–121.

Ahonen, A., & Harding, S.-M. (2018). Assessing online collaborative problem solving among school children in Finland: A case study using ATC21S TM in a national context. *International Journal of Learning, Teaching and Educational Research, 17*(2), 138–158. https://doi.org/10.26803/ijlter.17.2.9

Andrews, J. J., Kerr, D., Mislevy, R. J., von Davier, Hao, J., & Liu, L. (2017). Modeling collaborative interaction patterns in a simulation-based task. *Journal of Educational Measurement, 54*(1), 54–69.

Andrews, J. J., & Rapp, D. N. (2015). Benefits, costs, and challenges of collaboration for learning and memory. *Translational Issues in Psychological Science, 1*(2), 182–191.

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*, 105759. https://doi.org/10.1016/j.chb.2018.10.025.

Andrews-Todd, J., & Forsyth, C. M. (2022). Assessment of collaborative problem solving skills. In R. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International Encyclopedia of Education* (pp. 494–503). Elsevier.

Andrews-Todd, J., Jackson, G. T., & Kurzum, C. (2019). *Collaborative problem solving assessment in an online mathematics task (Research Report RR-19-24)*. Educational Testing Service.
Andrews-Todd, J., & Kerr, D. (2019). Application of ontologies for assessing collaborative problem solving skills. *International Journal of Testing, 19*(2), 172–187.
Andrews-Todd, J., & Toscano, M. (2020). *Effects of task design on collaboration patterns in an online task*. In M. Gresalfi, & I. S. Horn (Eds.) (pp. 1723–1724). International Society of the Learning Sciences.
Barrett, E., & Lally, V. (1999). Gender differences in an on-line learning environment. *Journal of Computer Assisted Learning, 15*(1), 48–60.
Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *The Journal of the Learning Sciences, 9*(4), 403–436.
Britton, E., Simper, N., Leger, A., & Stephenson, J. (2017). Assessing teamwork in undergraduate education: A measurement tool to evaluate individual teamwork skills. *Assessment & Evaluation in Higher Education, 42*(3), 378–397.
Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready for work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. In *Corporate Voices for Working Families, and Society for Human Resource Management* (pp. 1–64). The Conference Board, Partnership for 21st Century Skills.
Chen, N.-S., & Hwang, G.-J. (2014). Transforming the classrooms: Innovative digital game-based learning designs and applications. *Educational Technology Research & Development, 62*(2), 125–128.
Chung, G. K. W. K., O'Neil, H. F., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior, 15*(3), 463–493.
Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.
Clark, H. H. (1996). *Using language*. Cambridge University Press.
Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on Socially Shared Cognition, 13*(1991), 127–149.
Cohen, P. R. (1989). Planning and problem solving. In P. R. Cohen, & E. A. Feigenbaum (Eds.), *The handbook of artificial intelligence* (pp. 513–562). Addison-Wesley.
Crippen, K. J., & Antonenko, P. D. (2018). Designing for collaborative problem solving in STEM cyberlearning. In *Cognition, metacognition, and culture in stem education* (pp. 89–116). Springer.
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*, 419–438.
Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment*. Princeton, NJ: Educational Testing Service.
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.
Dedoose. (2018). *Dedoose: Web application for managing, analyzing, and presenting qualitative and mixed method research data*. SocioCultural Research Consultants, LLC version 8.0.35. .
Fenwick, G. D., & Neal, D. J. (2001). Effect of gender composition on group performance. *Gender, Work and Organization, 8*(2), 205–225.
Fernández, M., Wegerif, R., Mercer, N., & Rojas-Drummond, S. (2001). Re-conceptualizing" scaffolding" and the zone of proximal development in the context of symmetrical collaborative learning. *Journal of Classroom Interaction, 36*(2), 40–54.
Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., … A, A. (2017). Collaborative problem solving: Considerations for the national assessment of educational progress (pp.1-83). *National Center for Education Statistics*.
Flor, M., & Andrews-Todd, J. (2022). Towards automatic annotation of collaborative problem-solving skills in technology-enhanced environments. *Journal of Computer Assisted Learning, 38*(5), 1434–1447. https://doi.org/10.1111/jcal.12689.
Flor, M., Yoon, S.-Y., Hao, J., Liu, L., & von Davier, A. (2016). Automated classification of collaborative problem solving interactions in simulated science tasks. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 31–41). Association for Computational Linguistics.
Forsyth, C., Andrews-Todd, J., & Steinberg, J. (2020). Are you really a team player? Profiling of collaborative problem solvers in an online environment. In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.), *International Educational Data Mining Society* (pp. 403–408). Proceedings of The 13th International Conference on Educational Data Mining.
Frensch, P. A., & Funke, J. (1995). *Complex problem solving: The European perspective*. Lawrence Erlbaum.
Gao, Q., Zhang, S., Cai, Z., Liu, K., Hui, N., & Tong, M. (2022). Understanding student teachers' collaborative problem solving competency: Insights from process data and multidimensional item response theory. *Thinking Skills and Creativity, 45*, 101097.
Gijlers, H., & De Jong, T. (2005). The relation between prior knowledge and students' collaborative discovery learning processes. *Journal of Research in Science Teaching, 42*(3), 264–282.
Gilabert, R., Barón, J., & Llanes, À. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance. *International Review of Applied Linguistics in Language Teaching, 47*, 367–395.
Gillies, R. M. (2004). The effects of cooperative learning on junior high school students during small group learning. *Learning and Instruction, 14*(2), 197–213.
Gobert, J. D., Pedro, S., M, A., Baker, R. S. J. D., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining, 4*, 111–143.
Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest, 19*(2), 59–92.
Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C., & Germany, M.-L. (2018). Challenges of assessing collaborative problem solving. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills* (pp. 75–91). Springer.
Graesser, A. C., Kuo, B.-C., & Liao, C.-H. (2017). Complex problem solving in assessments of collaborative problem solving. *Journal of Intelligence, 5(2), 10*. https://doi.org/10.3390/jintelligence5020010
Greiff, S. (2012). From interactive to collaborative problem solving: Current issues in the Programme for International Student Assessment. *Review of Psychology, 19*(2), 111–121.
Griffin, P., & Care, E. (2014). *Assessment and teaching of 21st century skills: Methods and approach*. Springer.
Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*. Springer.
Gu, X., Chen, S., Zhu, W., & Lin, L. (2015). An intervention framework designed to develop the collaborative problem-solving skills of primary school students. *Educational Technology Research & Development, 63*(1), 143–159.
Hadwin, A., Järvelä, S., & Miller, M. (2018). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In *Handbook of self-regulation of learning and performance* (pp. 83–106). Routledge.
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 RR–19-41; ETS Research Report Series*, 1–17. Educational Testing Service.
Hao, J., Liu, L., von Davier, A, A., & Kyllonen, P. C. (2017). Initial steps towards a standardized assessment for collaborative problem solving (CPS): Practical challenges and strategies. In *Innovative assessment of collaboration* (pp. 135–156). Springer.
Hao, J., Liu, L., von Davier, A, A., Kyllonen, P., & Kitchen, C. (2016). Collaborative problem solving skills versus collaboration outcomes: Findings from statistical analysis and data mining. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th international conference on educational data mining* (pp. 382–387). International Educational Data Mining Society.
Hastie, C., Fahy, K., & Parratt, J. (2014). The development of a rubric for peer assessment of individual teamwork skills in undergraduate midwifery students. *Women and Birth, 27*(3), 220–226.
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*, 105624.

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). Springer.
Honey, M. A., & Hilton, M. L. (2011). *Learning science through computer games and simulations*. National Academies Press.
Hwang, G.-J., & Wu, P.-H. (2012). Advancements and trends in digital game-based learning research: A review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology, 43*(1), E6–E10.
Isohätälä, J., Näykki, P., & Järvelä, S. (2020). Cognitive and socio-emotional interaction in collaborative learning: Exploring fluctuations in students' participation. *Scandinavian Journal of Educational Research, 64*(6), 831–851.
Janssen, J., Erkens, G., Kirschner, P. A., & Kanselaar, G. (2012). Task-related and social regulation during online collaborative learning. *Metacognition and Learning, 7*(1), 25–43.
Järvelä, S., Järvenoja, H., Malmberg, J., & Hadwin, A. F. (2013). Exploring socially shared regulation in the context of collaboration. *Journal of Cognitive Education and Psychology, 12*(3), 267–286.
Järvenoja, H., Järvelä, S., & Malmberg, J. (2020). Supporting groups' emotion and motivation regulation during collaborative learning. *Learning and Instruction, 70*, 101090.
Jeong, H., Hmelo-Silver, C. E., & Jo, K. (2019). Ten years of computer-supported collaborative learning: A meta-analysis of CSCL in STEM education during 2005–2014. *Educational Research Review, 28*, 100284.
Kaimara, P., Fokides, E., Oikonomou, A., & Deliyannis, I. (2022). Pre-service teachers' views about the use of digital educational games for collaborative learning. *Education and Information Technologies, 27*(4), 5397–5416.
Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*(4), 681–706.
Kerr, D., Andrews, J. J., & Mislevy, R. J. (2016). The in-task assessment framework for behavioral data. In A. A. Rupp, & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 472–507). Wiley-Blackwell.
Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55*, 623–655.
Kuo, B.-C., Liao, C.-H., Pai, K.-C., Shih, S.-C., Li, C.-H., & Mok, M. M. C. (2020). Computer-based collaborative problem-solving assessment in Taiwan. *Educational Psychology, 40*(9), 1164–1185.
Kyllonen, P. C., Zhu, M., von Davier, & A, A. (2017). Introduction: Innovative assessment of collaboration. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 1–18). Springer.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310
Laughlin, P. R., Zander, M. L., Knievel, E. M., & Tan, T. K. (2003). Groups perform better than the best individuals on letters-to-numbers problems: Informative equations and effective strategies. *Journal of Personality and Social Psychology, 85*(4), 684–694.
LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology, 61*(2), 273–307.
Lipponen, L., Rahikainen, M., Lallimo, J., & Hakkarainen, K. (2003). Patterns of participation and discourse in elementary students' computer-supported collaborative learning. *Learning and Instruction, 13*(5), 487–509.
Liu, L., von Davier, A, A., Hao, J., Kyllonen, P., & Zapata-Rivera, J.-D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development* (pp. 344–359). IGI-Global.
Lobczowski, N. G., Allen, E. M., Firetto, C. M., Greene, J. A., & Murphy, P. K. (2020). An exploration of social regulation of learning during scientific argumentation discourse. *Contemporary Educational Psychology, 63*, 101925.
Lou, Y., Abrami, P. C., & d'Apollonia, S. (2001). Small group and individual learning with technology: A meta-analysis. *Review of Educational Research, 71*(3), 449–521.
Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). Lawrence Erlbaum.
McGunagle, D., & Zizka, L. (2020). Employability skills for 21st-century STEM students: The employers' perspective. *Higher Education, Skills and Work-based Learning, 10*(3), 591–606.
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.
National Research Council. (2008). *Research on future skill demands: A workshop summary*. National Academies Press.
Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). McGraw-Hill.
Nokes-Malach, T. J., Meade, M. L., & Morrow, D. G. (2012). The effect of expertise on collaborative problem solving. *Thinking & Reasoning, 18*(1), 32–58.
Nouri, J., Åkerfeldt, A., Fors, U., & Selander, S. (2017). Assessing collaborative problem solving skills in technology-enhanced learning environments–The PISA Framework and modes of communication. *International Journal of Emerging Technologies in Learning, 12*(4), 163–174.
OECD.. (2013a). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing.
OECD.. (2013b). *PISA 2015 collaborative problem solving framework*. OECD Publishing.
OECD.. (2017). Collaborative problem solving. *PISA 2015 results, V*. https://doi.org/10.1787/9789264285521-en. OECD Publishing.
Oliveri, M. E., Lawless, R., & Molloy, H. (2017). A literature review on collaborative problem solving for college and workforce readiness. *ETS Research Report Series, 2017*(1), 1–27.
O'Neil, H. F. (1999). Perspectives on computer-based performance assessment of problem solving. *Computers in Human Behavior, 15*(3/4), 225–268.
O'Neil, H. F., Chung, G. K. W. K., & Brown, R. S. (1995). Measurement of teamwork processes using computer simulation (CSE Tech. Rep. No. 399; pp. 1-80). *National Center for Research on Evaluation, Standards, and Student Testing*.
Partnership of 21st Century Learning. (2016). Framework for 21st century learning. http://www.p21.org/storage/documents/docs/P21_framework_0816.pdf.
Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press.
Prinsen, F. R., Volman, M. L. L., & Terwel, J. (2007). Gender-related differences in computer-mediated communication and computer-supported collaborative learning. *Journal of Computer Assisted Learning, 23*(5), 393–409.
Pugh, S. L., Subburaj, S. K., Rao, A. R., Stewart, A. E., Andrews-Todd, J., & D'Mello, S. K. (2021). *Say What? Automatic modeling of collaborative problem solving skills from student speech in the wild* (pp. 55–67). International Educational Data Mining Society.
Pugh, S. L., Rao, A., Stewart, A. E., & 'Mello, S. K. (2022). *Do speech-based collaboration analytics generalize across task contexts?*. In (pp. 208–218) LAK22: 12th International Learning Analytics and Knowledge Conference.
Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science, 323*, 75–79.
Resta, P., & Laferrière, T. (2007). Technology in support of collaborative learning. *Educational Psychology Review, 19*(1), 65–83.
Rojas, M., Nussbaum, M., Chiuminatto, P., Guerrero, O., Greiff, S., Krieger, F., & Van Der Westhuizen, L. (2021). Assessing collaborative problem-solving skills among elementary school students. *Computers & Education, 175*, 104313.
Rojas, M., Nussbaum, M., Guerrero, O., Chiuminatto, P., Greiff, S., Del Rio, R., & Alvares, D. (2022). Integrating a collaboration script and group awareness to support group regulation and emotions towards collaborative problem solving. *International Journal of Computer-Supported Collaborative Learning, 17*(1), 135–168.
Romero, C., González, P., Ventura, S., del Jesús, M, J., & Herrera, F. (2009). Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. *Expert Systems with Applications, 36*, 1632–1644.
Rosen, Y. (2014). Comparability of conflict opportunities in human-to-human and human-to-agent online collaborative problem solving. *Technology, Knowledge and Learning, 19*(1–2), 147–164.
Rosen, Y., Wolf, I., & Stoeffler, K. (2020). Fostering collaborative problem solving skills in science: The Animalia project. *Computers in Human Behavior, 104*, 105922.
Schriver, A. T., Morrow, D. G., Wickens, C. D., & Talleur, D. A. (2008). Expertise differences in attentional strategies related to pilot decision making. *Human Factors, 50*(6), 864–878.

Shute, V. J., & Becker, B. J. (2010). Prelude: Assessment for the 21st century. In V. J. Shute, & B. J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 1–11). Springer.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research, 106*, 423–430.

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education, 157*, 103964.

Stewart, A. E., Vrzakova, H., Sun, C., Yonehiro, J., Stone, C. A., Duran, N. D., … D'Mello, S. K. (2019). *I say, you say, we say: Using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving*. In (pp. 1–19). Proceedings of the ACM on Human-Computer Interaction, 3 (CSCW).

Stoeffler, K., Rosen, Y., Bolsinova, M., von Davier, & A, A. (2020). Gamified performance assessment of collaborative problem solving skills. *Computers in Human Behavior, 104*, 106036.

Sun, C., Shute, V. J., Stewart, A. E., Beck-White, Q., Reinhardt, C. R., Zhou, G., … D'Mello, S. K. (2022). The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Computers in Human Behavior, 128*, 107120.

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*, 103672.

Swiecki, Z. (2021). Measuring the impact of interdependence on individuals during collaborative problem-solving. *Journal of Learning Analytics, 8*(1), 75–94.

Szewkis, E., Nussbaum, M., Rosen, T., Abalos, J., Denardin, F., Caballero, D., Tagle, A., & Alcoholado, C. (2011). Collaboration within large groups in the classroom. *International Journal of Computer-Supported Collaborative Learning, 6*(4), 561–575.

Tang, P., Liu, H., & Wen, H. (2021). Factors predicting collaborative problem solving: Based on the data from PISA 2015. *Frontiers in Education, 6*, 619450.

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. https://doi.org/10.1177/1046496414525478

Trivedi, A., Kaur, E. K., Choudhary, C., & Barnwal, P. (2023). Should AI technologies replace the human jobs? 2nd International conference for innovation in technology (INOCON), 1–6. https://doi.org/10.1109/INOCON57975.2023.10101202.

Uz-Bilgin, C., Thompson, M., & Anteneh, M. (2020). Exploring how role and background influence through analysis of spatial dialogue in collaborative problem-solving games. *Journal of Science Education and Technology, 29*(6), 813–826.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47*(1), 513–539.

Whorton, R., Casillas, A., Oswald, F. L., & Shaw, A. (2017). Critical skills for the 21st century workforce. In J. Burrus, K. D. Mattern, B. Naemi, & R. D. Roberts (Eds.), *Building better students: Preparation for the workforce* (pp. 47–72). Oxford University Press.

Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice, 10*(3), 329–345.

Zambrano, J., Kirschner, F., Sweller, J., & Kirschner, P. A. (2019). Effects of prior knowledge on collaborative and individual learning. *Learning and Instruction, 63*, 101214.

Zhang, S., Gao, Q., Sun, M., Cai, Z., Li, H., Tang, Y., & Liu, Q. (2022). Understanding student teachers' collaborative problem solving: Insights from an epistemic network analysis (ENA). *Computers & Education, 183*, 104485.

Zhou, G., Moulder, R. G., Sun, C., & D'Mello, S. K. (2022). nvestigating temporal dynamics underlying successful collaborative problem solving behaviors with multilevel vector autoregression. In A. Mitrovic, & N. Bosch (Eds.), *International Educational Data Mining Society* (pp. 290–301). Proceedings of the 15th International Conference on Educational Data Mining. https://doi.org/10.5281/zenodo.6853137.

Zhu, M., Andrews-Todd, J., & Zhang, M. (2020). Application of network analysis in understanding collaborative problem solving processes and skills. In J. Hong, & R. W. Lissitz (Eds.), *Information Age. Innovative Psychometric Modeling and Methods*.