

Computer-based Assessment of Collaborative Problem Solving: Exploring the Feasibility of Human-to-Agent Approach

Yigal Rosen¹

Published online: 12 May 2015

© International Artificial Intelligence in Education Society 2015

Abstract How can activities in which collaborative skills of an individual are measured be standardized? In order to understand how students perform on collaborative problem solving (CPS) computer-based assessment, it is necessary to examine empirically the multi-faceted performance that may be distributed across collaboration methods. The aim of this study was to explore possible differences in student performance in human-to-agent (H-A), compared to human-to-human (H-H) CPS assessment tasks. One hundred seventy nine 14 years-old students from the United States, Singapore and Israel participated in the study. Students in both H-H and H-A modes were able to collaborate and communicate by using identical methods and resources. However, while in the H-A mode, students collaborated with a simulated computer-driven partner, and in the H-H mode students collaborated with another student to solve a problem. Overall, the findings showed that CPS with a computer agent involved significantly higher levels of shared understanding, progress monitoring, and feedback. However, no significant difference was found in a student's ability to solve the problem or in student motivation with a computer agent or a human partner. One major implication of CPS score difference in collaboration measures between the two modes is that in H-A mode one can program a wider range of interaction possibilities than would be available with a human partner. Thus, H-A approach offers more opportunities for students to demonstrate their CPS skills. This study is among the first of its kind to investigate systematically the effect of collaborative problem solving in standardized assessment settings.

Keywords Collaborative problem solving · Computer agent · Performance assessment

✉ Yigal Rosen
yigal.rosen@pearson.com

¹ Pearson, Brookline, MA, USA

Introduction

Collaborative problem solving (CPS) is a critical competency for college and career readiness. Students emerging from schools into the workforce and public life will be expected to have CPS skills as well as the ability to perform that collaboration in various group compositions and environments (Griffin et al. 2012; OECD 2013; O’Neil and Chuang 2008; Rosen and Rimor 2012). Recent curriculum and instruction reforms have focused to a greater extent on teaching and learning CPS (National Research Council 2011; U.S. Department of Education 2010). However, structuring standardized computer-based assessment of CPS skills, specifically for large-scale assessment programs, is challenging. In a standardized assessment situation, a student should be matched with various types of group members that will represent different CPS skills and contexts. In addition, the discourse between the group members should be manageable and predictable. The two major questions thus are: *Can partners for CPS be simulated but still maintain authentic human aspects of collaboration? And, how can manageable and predictable group discourse spaces be created within the assessment?* This paper explores the feasibility of the use of computer-agent methodology for scalable computer-based assessment of CPS, providing findings from an empirical pilot study conducted in three countries, as well as discussing implication of the findings on further research and development.

Defining Collaborative Problem Solving

Currently, the terms “collaborative problem solving”, “cooperative work” and “group work” are used interchangeably in the education research literature to mean similar constructs. Collaborative problem solving thus refers to problem-solving activities that involve collaboration among a group of individuals (O’Neil et al. 2010; Zhang 1998). CPS is a conjoint construct consisting of collaboration, or: “coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem” (Roschelle and Teasley 1995, p. 70), and problem solving, or: “cognitive processing directed at achieving a goal when no solution method is obvious to the problem solver” (Mayer and Wittrock 1996). According to Griffin et al. (2012), CPS refers to the ability to recognize the points of view of other persons in a group; to contribute knowledge, experience, and expertise in a constructive way; to identify the need for contributions and how to manage them; to recognize structure and procedure involved in resolving a problem; and as a member of the group, to build and develop group knowledge and understanding. CPS is one of the two major areas that the Organisation for Economic Co-operation and Development (OECD) selected in 2015 for primary development in Programme for International Student Assessment (PISA). In PISA 2015, CPS competency is defined as “the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills, and efforts to reach that solution” (OECD 2013). An agent could be considered either a human agent or a computer agent that interacts with the student. The competency is assessed by evaluating how well the individual collaborates with agents during the problem-solving process. This includes establishing and maintaining shared understanding, taking appropriate actions to solve the problem, and establishing and maintaining group organization.

In our research, an operational definition of CPS refers to “*the capacity of an individual to effectively engage in a group process whereby two or more agents attempt to solve a problem by sharing knowledge and understanding, organizing the group work and monitoring the progress, taking actions to solve the problem, and providing constructive feedback to group members.*”

First, CPS requires students to be able to establish, monitor, and maintain the shared understanding throughout the problem-solving task by responding to requests for information, sending important information to agents about tasks completed, establishing or negotiating shared meanings, verifying what each other knows, and taking actions to repair deficits in shared knowledge. Shared understanding can be viewed as an effect, if the goal is that a group builds the common ground necessary to perform well together, or as a process by which peers perform conceptual change (Dillenbourg 1999). CPS is a coordinated joint dynamic process that requires periodic communication between group members. Communication is a primary means of constructing a shared understanding, as modeled in Common Ground Theory (Clark 1996). An “optimal collaborative effort” is required of all of the participants in order to achieve adequate performance in a collaborative environment (Dillenbourg and Traum 2006).

Second, collaboration requires the capability to identify the type of activities that are needed to solve the problem and to follow the appropriate steps to achieve a solution. This process involves exploring and interacting with the problem situation. It includes understanding both the information initially presented in the problem and any information that is uncovered during interactions with the problem. The accumulated information is selected, organized, and integrated in a fashion that is relevant and helpful to solving the particular problem and that is integrated with prior knowledge. Setting sub-goals, developing a plan to reach the goal state, and executing the plan that was created are also a part of this process. Overcoming the barriers of reaching the problem solution may involve not only cognition, but motivational and affective means (Funke 2010; Mayer and Wittrock 2006).

Third, students must be able to help organize the group to solve the problem; consider the talents and resources of group members; understand their own role and the roles of the other agents; follow the rules of engagement for their role; monitor the group organization; reflect on the success of the group organization, and help handle communication breakdowns, conflicts, and obstacles (Rosen and Rimor 2012).

Assessing Collaborative Problem Solving Skills

Student performance in CPS can be assessed through a number of different methods. These include measures of the quality of the solutions and the objects generated during the collaboration (Avouris et al. 2003); analyses of log files, intermediate results, paths to the solutions (Adejumo et al. 2008), team processes and structure of interactions (O’Neil et al. 1997a, b); and quality and type of collaborative communication (Cooke et al. 2003; Foltz and Martin 2008; Graesser et al. 2008). There are distinct tradeoffs between the large amount of information that can be collected in a collaborative activity and what can be measured. For example, while the content of spoken communications is quite rich, analyses by human markers can be quite time consuming and difficult to automate. Nevertheless, much of the problem-solving process data as well other

communication information (turn taking, amount of information conveyed) can be analyzed by automatic methods.

To ensure valid measurement on the individual level, each student should be paired with partners displaying various ranges of CPS characteristics (Graesser et al. 2015). This way each individual student will be situated fairly similarly to be able to show his or her proficiency in CPS. Educators are urged to carefully consider group composition when creating collaborative groups or teams (Fall et al. 1997; Rosen and Rimor 2009; Webb 1995; Wildman et al. 2012). Additionally, students should act in different roles (e.g., team leader) and be able to work collaboratively in various types of environments.

Among other factors that may influence student CPS are gender, race, status, perceived cognitive or collaborative abilities, motivation and attractiveness. According to Dillenbourg (1999), effective collaboration is characterized by a relatively symmetrical structure. Symmetry of knowledge occurs when all participants have roughly the same level of knowledge, although they may have different perspectives. Symmetry of status involves collaboration among peers rather than interactions involving facilitator relationships. Finally, symmetry of goals involves common group goals rather than individual goals that may conflict. The degree of negotiability is additional indicators of collaboration (Dillenbourg 1999). For example, trivial, obvious, and unambiguous tasks provide few opportunities to observe negotiation because there is nothing about which to disagree.

Thus, in a standardized assessment situation, it is possible that a student should be matched with various types of group members that will represent different collaboration and problem-solving skills, while controlling for other factors that may influence student performance (e.g., asymmetry of roles).

Another challenge in CPS assessment refers to the need for synthesizing information from individuals and teams along with actions and communication dimensions (Laurillard 2009; O'Neil et al. 2008; Rimor et al. 2010). Communication among the group members is central in the CPS assessment and it is considered a major factor that contributes to the success of CPS (Fiore and Schooler 2004; Dillenbourg and Traum 2006; Fiore et al. 2010). Various techniques were developed to address the challenge of providing a tractable way to communicate in CPS assessment context. One interesting technique that has been tested is communication through predefined messages (Chung et al. 1999; Hsieh and O'Neil 2002; O'Neil et al. 1997a, b). In these studies, participants were able to communicate using the predefined messages and to successfully complete a task (a simulated negotiation or a knowledge map), and the team processes and outcomes were measurable. Team members used the predefined messages to communicate with each other, and measures of CPS processes were computed based on the quantity and type of messages used (i.e., each message was coded a priori as representing adaptability, coordination, decision making, interpersonal skill, or leadership). The use of messages provides a manageable way of measuring CPS skills and allows real-time scoring and reporting.

A key consideration in CPS assessment is the development of situations where collaboration is critical to performing successfully on the task. Such situations require interdependency between the students where information must be shared. For example, dynamic problem situations can be developed where each team member has a piece of information and only together can they solve the problem (called hidden-profile or jigsaw problems, Aronson and Patnoe 2011). Similarly, ill-defined tasks can be developed using

such tasks as group bargaining where there are limited resources but a group must converge on a solution that satisfies the needs of different stakeholders. Finally, information between participants may also be conflicting, requiring sharing of the information and then resolution in order to determine what information best solves the problem. CPS tasks may include one or more of these situations, while the common factors in all these collaborative tasks are handling discord, disequilibrium, and group think. Usually, a group member cannot complete the task without taking actions to ensure that a shared understanding is established. Thus, a key element in CPS tasks is interdependency between the group members so that grounding is both required and observable.

Human-to-Human and Human-to-Agent Approach in CPS Assessment

Collaboration can take many forms, ranging from two individuals to large teams with predefined roles. Thus, there are a number of dimensions that can affect the type of collaboration and the processes used in problem solving. For example, there can be different-sized teams (two equal team members vs. three or more team members working together), different types of social hierarchies within the collaboration (all team members equal vs. team members with different levels of authority), and, for assessment purposes, different agents – whether all team members are human or some are computer agents. There are advantages and limitations for each method. The Human-to-Human (H-H) approach provides an authentic human-human interaction which is a highly familiar situation for students. Students may be more engaged and motivated to collaborate with their peers. Additionally, the H-H situation is closer to the CPS situations students will encounter in their personal, educational, professional and civic activities. However, pairing can be problematic because of individual differences that can significantly affect the CPS process and its outcome. Therefore, the H-H assessment approach of CPS may not provide enough opportunity to cover variations in group composition, diversity of perspectives and different team member characteristics in controlled manners, which are all essential for assessment on an individual level. Simulated team members for collaboration with a preprogrammed profile, actions and communication would potentially provide the coverage of the full range of collaboration skills with sufficient control. In the Human-to-Agent (H-A) approach, CPS skills are measured by pairing each individual student with a computer agent or agents that can be programmed to act as team members with varying characteristics relevant to different CPS situations. Group processes are often different depending on the task and could even be competitive. Use of computer agents provides an essential component of non-competitiveness to the CPS situation, as it is experienced by a student. Additionally, if the time-on-task is limited, taking the time to explain to each other may lower group productivity. As a result of these perceived constraints, a student collaborating in H-H mode may limit significantly the extent to which CPS dimensions, such as shared understanding, are externalized through communication with the partner. The agents in H-A communication can be developed with a full range of capabilities, such as text-to-speech, facial actions, and optionally rudimentary gestures. In its minimal level, a conventional communication media, such as text via emails, chat, or graphic organizer with lists of named agents can be used for H-A CPS purposes. However, CPS in H-A settings

deviate from natural human communication delivery and can cause distraction and sometimes irritation. The dynamics of H-H interaction (timing, conditional branching) cannot be perfectly captured with agents, and agents cannot adjust to idiosyncratic characteristics of humans. For example, human collaborators can propose unusual, exceptional solutions; the characteristic of such a process is that it cannot be included in a system following an algorithm, such as H-A interaction. If educators rely on CPS teaching of students in H-A interactions exclusively, there may be the risk that these students will build up expectations that do exactly follow such an algorithm.

It should be noted that in both settings, CPS skills should be measured through a number of tasks, where each task represents a phase in the problem solving and collaborative process and can contain several steps. CPS tasks may include: consensus building, writing a joint document, making a presentation, or Jigsaw problems. In each task, the student is expected to work with one or more team members to solve a problem, while the team members should represent different roles, attitudes and levels of competence in order to vary the collaborative problem solving situation the student is confronted with.

Research shows that computer agents can be successfully used for tutoring, collaborative learning, co-construction of knowledge, and CPS (e.g., Biswas et al. 2010; Graesser et al. 2008; Millis et al. 2011). A computer agent can be capable of generating goals, performing actions, communicating messages, sensing its environment, adapting to changing environments, and learning (Franklin and Graesser 1996). One of the examples for computer agent use in education is a teachable agent system called Betty's Brain (Biswas et al. 2005; Leelawong and Biswas 2008). In this system, students teach a computer agent using a causal map, which is a visual representation of knowledge structured as a set of concepts and their relationships. Using their agent's performance as motivation and a guide, students study the available resources so that they can remediate the agent's knowledge and, in the process, learn the domain material themselves. Operation ARIES (Cai et al. 2011; Millis et al. 2011) uses animated pedagogical agents that converse with the student in a game-based environment for helping students learn critical-thinking skills and scientific reasoning within scientific inquiry. The system dynamically adapts the tutorial conversations to the learner's prior-knowledge. These conversations, referred to as "triads" are between the human learner and two computer agents (student and teacher). The student learns vicariously by observing the agents, gets tutored by the teacher agent, and teaches the student agent.

In summary, CPS assessment must take into account the types of technology, tasks and assessment contexts in which it will be applied. The assessment will need to consider the kinds of constructs that can be reliably measured and also provide valid inferences about the collaborative skills being measured. Technology offers opportunities for assessment in domains and contexts where assessment would otherwise not be possible or would not be scalable. One of the important improvements brought by technology to educational assessment is the capacity to embed system responses and behaviors into the instrument, enabling it to change its state in response to student's manipulations. These can be designed in such a way that the student will be exposed to an expected scenario and set of interactions, while the student's interactions as well as the explicit responses are captured and scored automatically. Computer-based assessment of CPS involves the need for advancements in educational assessment methodologies and technology. Group composition, discourse management, and the use of

computer agents, are considered as the major challenges in designing valid, reliable, and scalable assessment of CPS skills (Graesser et al. 2015). The paper addresses these challenges by studying student CPS performance in two modes of CPS assessment.

Research Questions

The study addressed empirically the following primary question regarding students' CPS performance in H-A, compared to H-H CPS settings:

What are the differences in student CPS performance between H-A and H-H mode of assessment, as reflected in shared understanding, problem solving, progress monitoring and providing feedback measures?

In order to better understand possible factors that differentiate student performance in H-A and H-H settings, the following research questions were examined:

What are the differences in student motivation while collaborating with a computer agent or a human partner on CPS assessment tasks?

What are the differences in student CPS performance between H-A and H-H modes of assessment, as reflected in time-on-task, and number of attempts to solve the problem?

It should be noted that the H-A condition represents student performance with a medium-level CPS behavior of a computer agent.

Method

Study participants included 179 students age 14, from the United States, Singapore and Israel. The results presented in the current article came from a larger study in which students from six countries were recruited to participate in a 21st Century Skills Assessment project investigating the innovative ways to develop computer-based assessment of critical-thinking, and CPS. The researchers collected data between November 2012 and January 2013. Recruitment of participating schools was achieved through collaboration with local educational organizations based on the following criteria: (a) the school is public, (b) the school is actively involved in various 21st Century Skills projects, (c) the population is 14 years-old students proficient in English, and (d) there is sufficient technology infrastructure (e.g., computers per student, high-speed Internet). In all, 136 students participated in the H-A group and 43 participated in the H-H group (43 additional students participated in the H-H setting, acting as 'collaborators' for the major H-H group). Specifically in H-H assessment mode, students were randomly assigned into pairs to work on the CPS task. Because the H-H approach required pairs of students working together in a synchronized manner, the number of pairs was limited. This is due to the characteristics of technology infrastructures in participating schools.

The students were informed prior to their participation in the study whether they collaborate with a computer agent or a classmate. In a case of H-H setting, the students were able to see the true name of their partner.

Of the total students who participated, 88 were boys (49.2 %) and 91 were girls (50.8 %). Table 1 summarizes the country and gender distribution of participating students between the H-A and H-H groups.

No significant differences were found in Grade Point Average (GPA), English Language Arts (ELA), and Math average scores between participants in H-A and H-H mode within the countries. This similar student background allowed comparability of student results in CPS assessment task between the two modes of collaboration.

Collaborative Problem Solving Assessment

In this CPS computer-based assessment task, the student was asked to collaborate with a partner (computer-driven agent or a classmate) to find the optimal conditions for an animal at the zoo (26 years of life expectancy). The task was designed to measure the CPS skills of the student to establish and maintain shared understanding, take appropriate actions to solve the problem, monitor progress, and provide feedback to the partner. The task demands were developed in line with PISA (OECD 2013) design considerations for CPS assessment tasks. The assessment is focused on how well the individual interacts with agents during the course of problem solving; this includes achieving a shared understanding of the goals and activities as well as efforts to solve the problem and pooling resources. An agent could be considered either a human or a computer agent that interacts with the student. In both cases, an agent has the capability of generating goals, performing actions, communicating messages, sensing its environment, and adapting to changing environments (Franklin and Graesser 1996). Some of the CPS skills are reflected in actions that the student performs, such as making a decision by choosing an item on the screen, or selecting values of parameters in a simulation. Other skills require acts of communication, such as asking the partner questions, answering questions, making claims, issuing requests, giving feedback on other agents' actions, and so on.

The Zoo Quest Task

In this task, the student was able to select different types of food, life environments, and extra features, while both partners were able to see the selections made and communicate through a phrase-chat (selections from predefined 4–5 options). An animal's life

Table 1 Research population by mode of assessment and country

Group	H-A		H-H	
	Female	Male	Female	Male
United States	22	19	6	4
Singapore	12	31	8	7
Israel	30	22	13	5
Overall	64	72	27	16

expectancy under the given conditions was presented after each trial of the conditions. The student and the partner were prompted to discuss how to reach better conditions for an animal at the beginning of the task. By the end of the task, the student was asked to rate the partner (1–3 stars) and provide written feedback on the partner's performance. It should be noted that due to the centrality of the collaboration dimension in CPS as it was defined in this study, the difficulty level of the problem was relatively low and served primarily as a platform for the overall assessment of CPS skills. Additionally, due to the exploratory nature of the study, the students were not limited either in a number of attempts to reach optimal solution or in the time-on-task. However, the task was programmed in such a way that at least two attempts for problem solving and at least one communication act with a partner were required to be able to complete the assessment task.

The task was checked with ten teachers from the three participating countries to ensure that students would be able to work on the task, that the task could differentiate between high and low levels of CPS ability, and that the task was free of cultural biases. Interviews were conducted with eight students representing the target population to validate various CPS actions and communication programmed for the computer agent and to establish automatic scoring of student responses. Through a series of one-on-one interviews, researchers asked teachers and students to “think out loud” when looking at the CPS task. General questions to the participants included items such as: Would it matter if the partner for the task was a human or a computer player? What if the open chat box was a phrase chat box instead? Which parts were hard/easy? What do you think it really tests? Some of the screen-specific questions included: Which parts are like things you've done before in the classroom? Which parts aren't? What would you change? The structure of the analyses was exploratory in nature and aimed at gaining insight as to teachers' and students' perspective on H-H and H-A conditions of the CPS task. Results suggest that the CPS task included in this study aligns well to the intended construct and provides an engaging and authentic task that allow for the assessment of CPS. The teachers and students expressed a high comfort level in working on a CPS assessment with a computer agent and with a human partner. Students were wary of being paired with a less effective partner, and liked the idea of a computer agent. Additionally, the students emphasized the added-value of real-life scenarios embedded into a game-like assessment environment. The results of the interviews should be interpreted with a consideration to the relatively small samples of teachers and students. Despite this major limitation, the interviews provided rich data and shed light on potential avenues for development of CPS assessment tasks.

The following information was presented interactively to the student during the Zoo Quest task in this study:

Episode #1: It was a normal zoo... But suddenly disaster struck! A rare animal called an Artani was found dead! You and a friend have been tasked with saving the other Artani by finding the most suitable conditions for them within the zoo.

Episode #2: Collaborate with your partner, help the Artani: In this task you will be in control of the selections made by you and your friend. Work with your partner to determine the best living conditions for Artani. You must change the three elements of Environment, Food, and Extras to find the best living conditions.

Your friend can help you plan a strategy to improve the conditions for Artani, before you make your selection.

Episode #3: Figures 1 and 2 show examples of the task screens. The major area of the screen allows the partners to view the options available for the Environment, Food, and Extras. Both partners can see what variables were selected. However, the selections of the variables were made by one partner only (i.e., by the student in H-A mode or by one of the students in H-H mode) as well as the ability to try out the variables selected (by clicking on “Go”). On the right side of the screen, the partners were able to communicate by using a phrase-chat. The phrases presented at the chat were based on a pre-programmed decision-tree and situated in order to allow the student to authentically communicate with a partner and to be able to cover the CPS measures defined for the task. The computer agent was programmed to act with varying characteristics relevant to different CPS situations (e.g., agree or disagree with the student, contributing to solving the problem or proposing misleading strategies, etc.). Each student in H-A setting (regardless of student’s actions) was provided with the same amount of help and misleading strategies proposed by the agent. This approach provided each individual student with similar optimal chances to show his or her CPS skills.

While in the H-H mode, the partners were provided with exactly the same set of possible phrases for each CPS situation. There was no control over the selections made by the students. In the H-H setting, the chat starts with the “leader” who asks questions and leads conversation. The other person takes the role of “collaborator” and generally replies to questions asked by the leader or initiating statements. Certain questions or responses, however can lead to different sets of available responses during conversation (e.g., asking a question), and so the person initially asking questions may not do so for

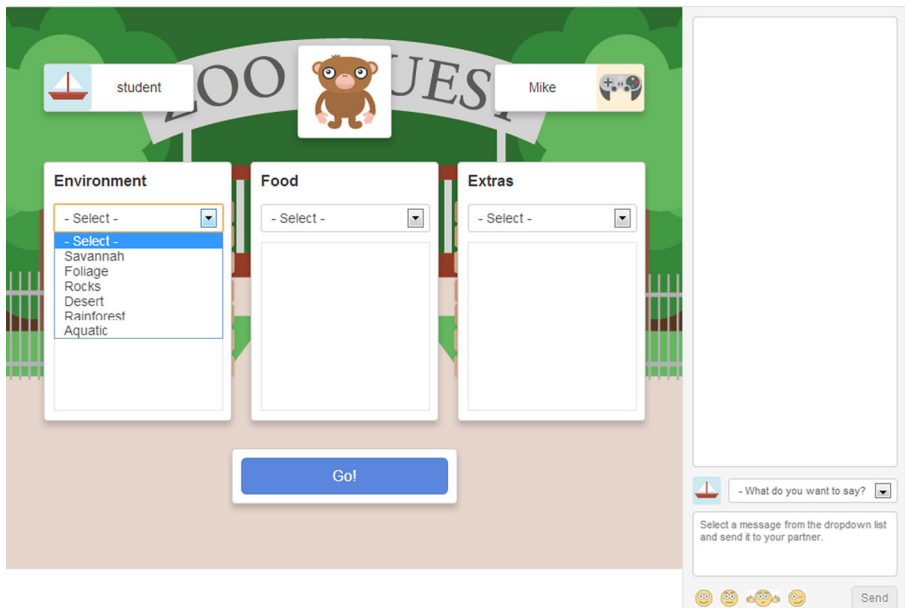


Fig. 1 Selecting variables in a CPS assessment task

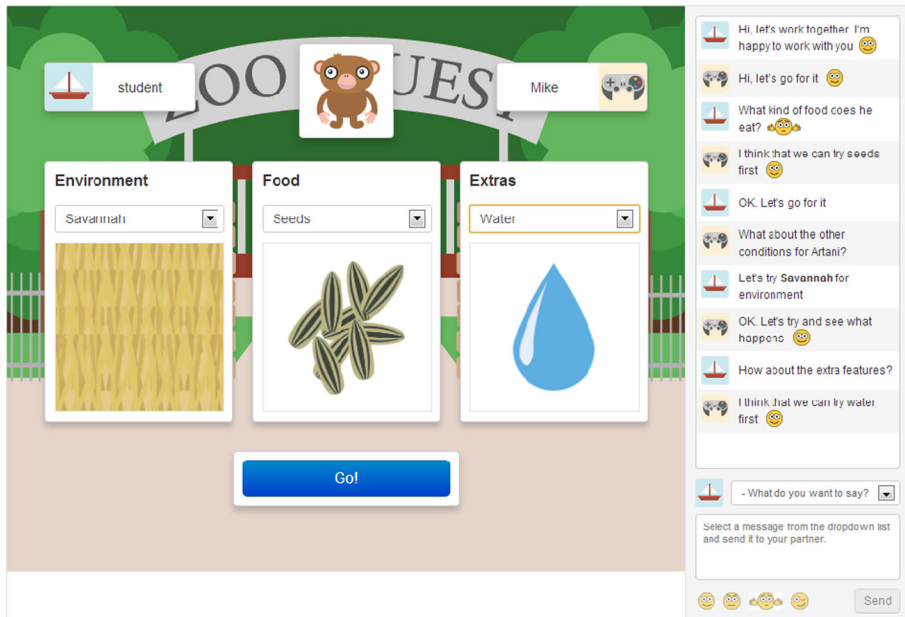


Fig. 2 Communication through a phrase-chat in a CPS assessment task

the entire conversation. Only the leader can submit guesses for the conditions and complete the task though. As the task progresses, some of the possible replies may be marked as “hidden”, which will mean they do not show again until the leader has submitted the attempt at solving the problem. If no replies are available then a secondary list of replies are available – this will be the case when all conditions have been agreed upon, and the focus will change to submitting. If the students have submitted their attempt at solving the problem, then some additional statements may become available to be said. There are various types of sentences that can be used through the phrase-chat. First, accepting statements that indicate agreement with the previous statement, such as “I think we should use ...” and “Yes, I agree.”. Both students need to agree on all three variables the first time they submit their attempt at solving the problem, but following this they only need to agree on changing one variable before they can progress. Second, some statements are linked, so that if one should be hidden as it has been accepted, then its linked statement should be as well. For example, “What kind of food does he eat?” and ‘Let’s try ... for food’ refer to the same subject; if the students have agreed on the kind of food, then there is no need to display these or similar questions again this time round. Last, some questions have options associated with them, such as ideas for food or environment. These are highlighted as “options” and can be selected by the students.

In the H-A setting, the student leads the chat that gets an automated response to simulate a two-way chat with a collaborator (i.e., a computer agent). The sentences a student can use are limited to what is relevant at that time and change based on their progress through the task, similarly to the H-H setting. Certain questions or replies, however can lead to a flipping of these roles during conversation, and so the student may not continue in his or her original role for the entire conversation. If the team fails

to improve the results for a few try outs, the computer agent provides a student with helpful advice, such as “Let’s try to change one condition per trial”, “I think we can try water”, or “I don’t think that ‘seeds’ is the best choice”. Only the student can submit the guesses for the conditions and complete the task though.

Clicking on “Go” provided the partners with the possibility to see the life expectancy of the animal under the variables selected (0–26 years) and to read textual information regarding the result achieved (see Fig. 3). At this stage the partners were allowed to communicate about the result and about ways to reach the optimal solution, and then decide whether to keep the selections or try again (i.e., change the variables). Figure 4 presents a simplified sample log file for student performance in the task.

Episode #4: Give feedback to your partner: Having saved the Artani, you need to provide feedback on your partner. Give you partner a star rating (1 to 3) and add written feedback below.

Episode #5: Partner’s feedback: Thanks for your feedback. You have done a great job on your part! Hope to work with you again sometime.

Designing Human and Computer Agent Interactions

Designing CPS assessment task requires detailed planning and prediction of students’ behaviors both in H-H and H-A modes of collaboration. Sample JavaScript Object Notation (JSON) code that led the interactions in H-H and H-A mode are presented below. It should be noted that the possible options to questions which have space in them include values for environment (‘Savannah’, ‘Foliage’, ‘Rocks’, ‘Desert’, ‘Rainforest’, ‘Aquatic’), food (‘Plants’, ‘Seeds’, ‘Vegetables’, ‘Fruits’, ‘Meat’, ‘Fish’), extras (‘Stones’,

The screenshot displays a user interface for a CPS assessment task. At the top, a banner reads "ZOO QUEST" with icons for a student, a monkey (Artani), and Mike. The main content area shows a vertical bar representing "Artani Life Expectancy" with a green bar reaching the number "18 years". Below this, text states: "Using your selection, the Artani are expected to live... 18 years". A cartoon character of a zookeeper is shown holding a shovel. A blue button labeled "Try again" is at the bottom right of the main area. On the right side, a chat window shows a conversation between the student and Mike. The chat messages are:

- Student: Hi, let's work together. I'm happy to work with you 😊
- Mike: Hi, let's go for it 😊
- Student: What kind of food does he eat? 🐵
- Mike: I think that we can try seeds first 😊
- Student: OK. Let's go for it
- Mike: What about the other conditions for Artani?
- Student: Let's try Savannah for environment
- Mike: OK. Let's try and see what happens 😊
- Student: How about the extra features?
- Mike: I think that we can try water first 😊

At the bottom of the chat window, there is a text input field with a dropdown menu and a "Send" button.

Fig. 3 Viewing the results of variables selected in a CPS assessment task

ATTEMPT	1
Student	<i>Hi, let's work together. I'm happy to work with you. (Emoticon = smile)</i>
Agent	Hi, let's go for it! (Emoticon=smile)
Student	<i>What kind of food does he eat? (Emoticon = question)</i>
Agent	I think that we can try seeds first. (Emoticon=smile)
Student	<i>How about the extra features?</i>
Agent	I think that we can try water first.
Student	<i>OK. Let's go for it.</i>
Agent	What about the other conditions for Artani?
Student	<i>Let's try [Aquatic] for environment. (Emoticon=smile)</i>
Agent	OK. Let's try and see what happens (Emoticon=smile)
Student	<i>Are you ready to go ahead with our plan?</i>
Agent	Click Go if you are ready to tryout the conditions
CHOICES	Aquatic, Seeds, Water
RESULT	10 [animal life expectancy in years]
PROMPT	Artani will be unhappy with the current conditions. I'd suggest you to reconsider some of your selections and provide Artani with conditions more suitable to its needs.
ATTEMPT	2
Student	<i>Let's try [Rainforest] for environment</i>
Agent	I don't think that it is the best choice.
Student	<i>What can we do to reach better conditions for the animal?</i>
Agent	Let's try to change one condition per trial.
CHOICES	Rainforest, Seeds, Water
RESULT	18 [animal life expectancy in years]
PROMPT	Artani will be very happy with the current conditions. Still, I would suggest you to reconsider some of your selections and provide Artani with the best conditions to its needs.
ATTEMPT	3
Student	<i>Should we keep this selection or try again?</i>
Agent	The target is 26 years, let's get as close as we can to that!
Student	<i>Let's try [Plants] for food.</i>
CHOICES	Rainforest, Plants, Water
RESULT	20 [animal life expectancy in years]
PROMPT	Artani will be very happy with the current conditions. Still, I would suggest you reconsider some of your selections and provide Artani with the best conditions to meet its needs.
ATTEMPT	4
Student	<i>Let's try [Tire swing] for extra features. (Emoticon=question)</i>
Agent	OK. Let's try and see what happens.
Student	<i>Let's go ahead with our plan?</i>
Agent	Click Go if you are ready to tryout the conditions.
CHOICES	Rainforest, Plants, Tire swing
RESULT	26 [animal life expectancy in years]
PROMPT	A job well done! You've chosen the best possible conditions for Artani.

Fig. 4 Communication through a phrase-chat in a CPS assessment task

‘Water’, ‘Tree house’, ‘Weed’, ‘Tire swing’), and types (‘Environment’, ‘Food’, ‘Extras’). Where the term ‘FirstPass’ is used in JSON statements, it refers to the first time the students have tried to guess the three variables, after submitting their first attempt the first pass is over. After the first three conditions have been agreed on, using accepting statements as discussed above, additional statements will be available. Also, some statements are linked, so that if one should be hidden as it's been accepted, then its linked statement should be as well. For example ‘What kind of food does he eat?’ and ‘Let's try ___ for food’ refer to the same subject and if the students have agreed

on the kind of food, then there's no need to display these or similar questions again, this time round. In H-H mode, sometimes questions need phrasing differently for the leader or the collaborator; this will be visible in a question when the properties 'altTextForLeader' or 'altTextForCollaborator' are in the JSON statements.

Each question within the H-A JSON has an associated "answerId", this is the ID of the answer to the question asked. Some question have the setting "secondAnswerId", this is the answer given if the student asks the same question twice. The students' progress through the task limits the questions they can ask; this is partially set using "optionGroup". Asking certain questions can add additional questions to their question list, this is set using "addQuestionAfterAsked". Some questions may be removed from their list after being asked, so they cannot ask the same question twice this is set using "removeAfterAsked". Some question and answers contain a variable; this is set on a question using "questionOptionListId" and on an answer using "answerOptionListId". The student can select one of these variables when asking the question.

Sample JSON code in H-H mode:

```
var responses={
  '1': {question: "Hi, let's work together. I'm happy to work with you", possibleReplies: [8, 2, 3, 4, 5, 6, 7],
  hideWhenAsked: true, hidden: false},
  '2': {question: "What will be the best environment for Artani?", linkWith: [5, 11], hideWhenAccepted: true,
  possibleReplies: [11, 10]},
  '3': {question: "What kind of food does he eat?", linkWith: [6, 12],
  hideWhenAccepted: true, possibleReplies: [12, 10]},
  '5': {question: "Let's try _____ for environment", linkWith: [2, 11],
  used: false, options: responseOptions.environment, possibleReplies: [9, 18]},
  '6': {question: "Let's try _____ for food", linkWith: [3, 12],
  used: false, options: responseOptions.food, possibleReplies: [9, 18]},
  '9': {question: "OK. Let's go for it", acceptingQuestion: true,
  possibleReplies: [10, 2, 3, 4, 5, 6, 7], whenNoAvailableReplies: [14, 15, 10, 5, 6, 7],
  additionalRepliesAfterFirstPass: [14, 15]},
  '11': {question: "I think that we can try _____ first",
  showOtherQuestionAfterFirstPass: 5, linkWith: [2, 5],
  hideWhenAccepted: true, options: responseOptions.environment,
  possibleReplies: [9, 18]},
  '14': {question: "Are you ready to go ahead with our plan?",
  possibleReplies: [16, 17]},
  '17': {question: "Click Go if you are ready to tryout the conditions",
  altTextForLeader: "I think we're ready to go ahead with our plan",
  possibleReplies: [14, 15, 2, 3, 4, 5, 6, 7]},
  '18': {question: "I don't think we are on the right track",
  possibleReplies: [2, 3, 4, 5, 6, 7], whenNoAvailableReplies: [14, 15, 5, 6, 7]},
```

```

'19': {question: "What can we do to reach better conditions for
the animal?", possibleReplies: [21, 22]},
'21': {question: "Let's try to change one condition per trial",
possibleReplies: [2, 3, 4, 5, 6, 7]},
'22': {question: "Maybe we can try change the _____",
possibleReplies: [2, 3, 4, 5, 6, 7], options:
responseOptions.types},
'23': {question: "Do you think we should try to improve the
conditions for the Artani?", possibleReplies: [25, 26]},
'24': {question: "Should we keep our selection?",
possibleReplies: [25, 26]},
'25': {question: "I think we should try to improve the
conditions", possibleReplies: [2, 3, 4, 5, 6, 7]},
'28': {question: "Nice working with you. Goodbye.",
possibleReplies: [27], disableChatAfter: true}

```

The following is sample JSON code in H-A mode.

```

questions: {
'1': {question: "Hi, let's work together. I'm happy to work
with you", answerId: 1, removeAfterAsked: true,
removeAfterAnyAsked: true},
'2': {question: "What will be the best environment for Artani?",
answerId: 2, optionGroup: 1, addQuestionAfterAsked: 8,
removeQuestionAfterAsked: 2, markQuestionAsAnswered: 2},
'3': {question: "What kind of food does he eat?", answerId:
3, optionGroup: 2, addQuestionAfterAsked: 8,
removeQuestionAfterAsked: 3, markQuestionAsAnswered: 3},
'5': {question: "Let's try _____ for environment",
questionOptionListId: 1, optionGroup: 1, answerId: 5,
secondAnswerId: 11, dontAddOrRemoveOthersIfEvenAnswer: true,
removeQuestionAfterAsked: 2, markQuestionAsAnswered: 2},
'6': {question: "Let's try _____ for food",
questionOptionListId: 2, optionGroup: 2, answerId: 5,
secondAnswerId: 11, dontAddOrRemoveOthersIfEvenAnswer: true,
removeQuestionAfterAsked: 3, markQuestionAsAnswered: 3},
'9': {question: "Are you ready to go ahead with our plan?",
answerId: 7},
'11': {question: "What can we do to reach better conditions for
the animal?", answerId: 9, secondAnswerId: 10,
removeThisAndQuestionAfterAsked: 12, removeAfterAsked: true},
'14': {question: "Nice working with you. Goodbye.", answerId:
15,}

```

Sample of possible answers the computer and human agent can respond with:

```

answers: {
'1': {answer: "Hi, let's go for it :)"},
'2': {answer: "I think that we can try rocks first :)"},

```

```
'3': {answer: "I think that we can try seeds first :)"},
'5': {answer: "OK. Let's try and see what happens :)"},
'6': {answer: "What about the other conditions for Artani?"},
'7': {answer: "Yeah! Make our selections and click Go"},
'9': {answer: "Let's try to change one condition per trial"},
'10': {answer: "Maybe we can try change the ",
answerOptionListId: 4},
'11': {answer: "I don't think that it is the best choice"},
'14': {answer: "The target is 26 years, let's get as close as we
can to that!"},
'15': {answer: "Nice working with you. Goodbye :)"}

```

The computer agent was programmed to act with varying characteristics relevant to CPS dimensions: problem solving, shared understanding, and group organization. The problem solving dimension was represented by the extent to which the agent helped the student to reach the optimal conditions for the animal (i.e., environment, food, and extra features). Shared understanding interaction was focused on agent's responses to grounding questions from the student. Group organization dimension was represented by agent's responses to student's questions related to progress monitoring. The agent was programmed to agree with student's selections in the first stage. Later in the process the agent occasionally disagreed with the student (every even selection per variable made by student excluding optimal selection), confirmed selections made by the student (every odd selection per variable made by student or in case of optimal selection), or proposed alternative solution (each two tryouts of all selected variables, excluding optimal selection). This standardized medium-level collaborative behavior of computer agent provided each individual student with similar optimal chances to show his or her CPS skills.

Scoring Considerations

The target CPS skills for this assessment consisted of shared understanding, taking appropriate actions to solve the problem, establishing and maintaining group organization, and providing feedback (e.g., OECD 2013; Graesser et al. 2015). In line with CPS dimensions, scores for student performance in the assessment task broken down as being: shared understanding (40 points), problem solving (26 points), and group organization represented by monitoring progress (26 points) and providing feedback (8 points). The different scales for scores across CPS measures result from different range of scorable situations or data points in each category. It should be noted that each student there were equal opportunities for full points available regardless of initial conversational paths chosen both in H-A and H-H conditions. First, students in this task must be able to establish, monitor, and maintain the shared understanding throughout the problem solving task by responding to requests for information, sending important information to the partner about tasks completed, establishing or negotiating shared meanings, verifying what each other knows, and taking actions to repair deficits in shared knowledge. Second, students must be able to identify the type of activities that are needed to solve the problem and to follow the appropriate steps to achieve a solution. Moreover, in this task students must be able to help organize the group to

solve the problem, consider the resources of group members, understand their own role and the roles of the other agents, follow the rules of engagement for their role, monitor the group organisation, and help handle communication breakdowns, conflicts, and obstacles. Finally, the students must be able to reflect on the success of the CPS process and the outcome.

Both in H-H and H-A settings, student scores in the first three CPS dimensions were generated automatically based on a predefined programmed sequence of possible optimal actions and communication that was embedded into the assessment task. It should be noted that the data from ‘leader’ student was used in H-H condition for CPS scoring purposes. The data from ‘collaborator’ was used for qualitative purposes only. This is due to the fact that only the performance of the ‘leader’ student was comparable to the role of the student in H-A settings both in terms of control over the selections of the living conditions and the similarity in available chat options. The problem-solving dimension was scored as one point per each year of the animal’s life expectancy that was achieved by selecting the variables. Students that were not fully engaged in an effort to reach the optimal life expectancy were scored lower in problem solving score, than the students that reached the solution. Shared understanding score consisted of a number of grounding questions that were initiated by a student in appropriate situations (e.g., explaining the reason for a variable selection, questioning “What can we do to reach better conditions for the animal?”) and appropriate responses to the grounding questions made by the partner. For example, students that made their own selections with low or no communication with the partner were scored low in shared understanding measure. Monitoring progress score was created based on communication initiated by the student prior to the submission of the selected variables (e.g., questioning “Are you ready to go ahead with our plan?” before clicking on “Go”) and the statements made by the student based on the life expectancy results that were achieved (e.g., “Should we keep this selection or try again?”). For example, students that clicked on “Go” without being engaged in confirmation of the selection made before submission were scored low in progress monitoring score.

Scoring of student feedback was provided independently by two teachers from participating schools in the United States. The teachers were trained through a 1-day workshop to consistently evaluate whether student’s feedback indicated both successful and challenging aspects of working with the partner on the task, and acknowledged the contributions the partner made toward reaching a solution. Spelling and grammar issues did not affect student score. Overall, the scoring strategy was discussed with a group of ten teachers from participating countries in order to achieve consensus on CPS scoring strategy and reduce cultural biases as much as possible. Inter-coded agreement of feedback scoring was 92 %.

It should be noted that the ‘collaborator’ student’s performance in H-H setting was not scored because of the non-comparability of this performance to the full CPS actions performed by the “leader” student.

Questionnaire

The questionnaire included four items to assess the extent to which students were motivated to work on the task. Participants reported the degree of their agreement with each item on a four-point Likert scale (1=strongly disagree, 4=strongly agree). The

items were adopted from motivation questionnaires used in previous studies, and included: “I felt interested in the task”; “The task was fun”; “The task was attractive”; “I continued to work on this task out of curiosity” (Rosen 2009; Rosen and Beck-Hill 2012). The internal reliability (Cronbach’s Alpha) of the questionnaire was .85.

Students were also asked to indicate their background information, including gender, GPA, and Math and ELA average score, as measured by school assessments.

Results

All results are presented on an aggregative level beyond the countries, since no interaction with student-related country was found. First, the results of student performance in a CPS assessment are presented to determine whether there is a difference in student CPS score as a function of collaborating with a computer agent versus a classmate. Next, student motivation results are presented to indicate possible differences in H-A and H-H modes. Last, time-on-task and number of attempts to solve the problem in both modes of collaboration are demonstrated.

Comparing Student CPS Performance in H-H and H-A Settings

In order to explore possible differences in students’ CPS scores analysis of variance was performed. First, MANOVA results showed significant difference between H-H and H-A groups (Wilks’ Lambda=0.904, $F(df=4174)=4.6$, $p<.01$). Hence, we proceed to perform t-tests and calculate Effect Size (Cohen’s d) for each CPS measure. The results indicated that students who collaborated with a computer agent showed significantly higher level of performance in establishing and maintaining shared understanding (ES=.4, $t(df=177)=2.5$, $p<.05$), monitoring progress of solving the problem (ES=.6, $t(df=177)=4.0$, $p<.01$), and in the quality of the feedback (ES=.5, $t(df=177)=3.2$, $p<.01$). The findings showed non-significant difference in the ability to solve the problem in the H-A and H-H mode of collaboration (ES=-.3, $t(df=177)=-1.9$, $p=.06$). Table 2 shows the results of student CPS scores in both modes.

Further qualitative analysis of the data revealed an interesting pattern (Rosen 2014). A process analysis of the chats and actions of the students showed that in the H-A group the students encountered significantly more situations of disagreement than in the H-H group (selecting a phrase such as, “I don’t think that [...] is the best choice for

Table 2 Student CPS scores in H-A and H-H modes of the assessment

CPS measure	H-A M(SD)	H-H M(SD)	ES	t(df)
Shared understanding	18.4(10.8)	14.6(8.3)	0.4	2.4(177)*
Problem solving	24.6(3.2)	25.3(1.7)	-.3	-1.9(177)
Monitoring progress	5.0(7.5)	1.3(4.2)	0.6	4.0(177)**
Providing feedback	4.1(2.2)	3.0(1.9)	0.5	3.2(177)**

** $p<.01$, * $p<.05$

environment.”, or submitting an unconfirmed solution). Students engaged in H-A mode of collaboration encountered 19.2 % more disagreement situations, compared with students in H-H setting. However, it was found that students in H-H setting were engaged in 11.3 % more situations in which the partner proposed different solutions for a problem (selecting a phrase such as, “let’s try [...] for food.”, or making a different choice of values for the life conditions).

Student Motivation

In attempting to determine possible differences in student motivation of being engaged in CPS with a computer agent versus a classmate, data on student motivation was analyzed. The result demonstrated that it is a matter of indifference in student’s motivation whether collaborating with a computer agent or a classmate ($M=3.1$, $SD=0.7$ in H-A mode, compared to $M=3.1$, $SD=0.4$ in H-H mode; $ES=0.1$, $t(df=177)=0.5$, $p=.64$).

Attempts to Solve a Problem and Time-on-Task

In order to examine possible differences in the number of attempts for problem-solving as well as time-on-task, a comparison of these measures was conducted between H-A and H-H modes of collaboration. In practice, the average number of attempts for problem solving in H-A mode was 8.4 ($SD=7.3$), compared to 6.1 ($SD=5.7$) in a H-H mode ($ES=0.3$, $t(df=177)=2.1$, $p<.05$). No significant difference was found in time-on-task ($t(df=177)=-1.6$, $p=.11$). On average, time-on-task in H-A mode was 7.9 min ($SD=3.6$), while student in the H-H mode spent 1.1 more minutes on a task ($M=9.0$, $SD=4.5$).

Discussion

Policymakers, researchers, and educators are engaged in vigorous debate about assessing CPS skills on an individual level in valid, reliable and scalable ways. Analyses of the list of challenges facing CPS in large-scale assessment programs suggests that both H-H and H-A approaches in CPS assessment should be further explored. The goal of this study was to explore differences in student CPS performance in H-A and H-H modes. Students in each of these modes were exposed to identical assessment tasks and were able to collaborate and communicate by using identical methods and resources. However, while in the H-A mode students collaborated with a simulated computer-driven partner, and in the H-H mode students collaborated with another student to solve a problem. The findings showed that students assessed in H-A mode outperformed their peers in H-H mode in their collaborative skills. CPS with a computer agent involved significantly higher levels of shared understanding, progress monitoring, and feedback. The design of an agent-based assessment was flexibly adaptive to the point where no two conversations are ever the same, just as is the case of collaborative interactions among humans. Although students in both H-H and H-A modes were able to collaborate and communicate by using identical methods and resources, full comparability was not expected. This is due to the fact that each student in H-H mode represented a specific set of CPS skills, while in the H-A mode each

individual student collaborated with a computer agent with a predetermined large spectrum of CPS skills. Differences across H-H groups could be affected by a given performance of the collaborator. Additionally, because of the relatively low difficulty of the problem that was represented by the CPS task, and much larger emphasis on collaboration, students in H-A were faced with more opportunities to show their collaboration skills. The qualitative findings suggest that students that acted as collaborators in H-H settings involve themselves less in disagreement situations, compared with a computer agent setting. This pattern partially explains the higher scores in ‘shared understanding’ and ‘monitoring progress’ scores in the H-A condition. However, human collaborators provided more insights into possible ways to solve the problem. The agent was programmed to partially disagree with the student, and occasionally misinterpret the results, while also propose different solutions. Conflicts are essential to enhance collaborative skills, therefore conflict situations are vital to collaborative problem solving performance (Mitchell and Nicholas 2006; Scardamalia 2002; Stahl 2006; Weinberger and Fischer 2006). It includes handling disagreements, obstacles to goals, and potential negative emotions (Barth and Funke 2010; Dillenbourg 1999). For example, the tendency to avoid disagreements can often lead collaborative groups towards a rapid consensus (Rimor et al. 2010), while students accept the opinions of their group members because it is a way to quickly advance with the task. This does not allow the measurement of a full range of collaborative problem solving skills. Collaborative problem solving assessment tasks may have disagreements between team members, involving highly proficient team members in collaborative problem solving (e.g., initiates ideas, supports and praises other team members), as well as team member with low collaborative skills (e.g., interrupts, comments negatively about work of others). By contrast, with integrative consensus process, students reach a consensus through an integration of their various perspectives and engage in convergent thinking to optimize the performance. An integrative approach for collaborative problem solving is also aligned with what is expected from a high proficiency student in collaborative problem solving (OECD 2013), that is expected to be able to “initiate requests to clarify problem goals, common goals, problem constraints and task requirements when contextually appropriate (p. 29)”, as well as to “detect deficits in shared understanding when needed and take the initiative to perform actions and communication to solve the deficits (p. 29)”. The differences in interaction style between H-A and H-H mode are in line with research on Intelligent Tutoring Systems (Rosé and Torrey 2005). This strand of research suggests that designers of computer-mediated dialogue systems should pay attention to negative social and metacognitive statements that students make to human and computer agents. For example, one of the key capabilities of human agents is the ability to engage in natural language dialogue with students, providing scaffolding, feedback and motivational prompts. Given the large space of possibilities and tradeoffs between them, further research is needed to better understand which of the techniques used by human agents will be effective in H-A interaction.

Challenges in Comparing Collaborative Approaches

Establishing absolute equivalence between H-A and H-H conditions may be impossible to achieve. There are three fundamental challenges. First, in H-H condition there are no

guarantees that the test-takers will end up interacting with an agent-like combination of other students to establish a convincing comparison. Second the time-on-task and number of attempts to solve the problem with humans may vary across conditions, so any direct comparison would be beset with a confound of differential task completion times. Third, measures would need to be established on what are defensible metrics of comparability. While a set of specific measures is explored in this study, different approaches are possible. Nevertheless, the argument can be made that such equivalence is not necessary for the overarching goal of measuring CPS skills. Many characteristics of H-H CPS can be captured by a computer agent approach so that the fidelity is sufficient for meaningful assessment of CPS skills.

One major possible implication of CPS score difference in collaboration measures between the H-A and H-H modes is that in H-A setting one can program a wider range of interaction possibilities than would be available with a human collaborator. Thus, H-A approach offers more opportunities for the human leaders to demonstrate their CPS skills. As such, this supports a validity argument for using the H-A approach in tasks that are intended to provide greater opportunities for evidence to support ones claims about students' proficiency in CPS skills. However, it should be acknowledged that each mode of CPS assessment can be differently effective for different educational purposes. For example, a formative assessment program which has adopted rich training on the communication and collaboration construct for its teachers may consider the H-H approach for CPS assessment as a more powerful tool to inform teaching and learning, while H-A may be implemented as a formative scalable tool across a large district or in standardized summative settings. Non-availability of students with a certain CPS level in a class may limit the fulfilment of assessment needs, but technology with computer agents can fill the gaps. In many cases, using simulated computer agents instead of relying on peers is not merely a replacement with limitations, but an enhancement of the capabilities that makes independent assessment possible. Furthermore, a phrase-chat used in this study can be replaced by an open-chat in cases where automated scoring of student responses is not needed.

In contrast, we found that it is a matter of indifference in student ability to solve the problem with a computer agent or a human partner, although on average students in H-A mode applied more attempts to solve the problem, compared to the H-H mode. Student performance studied here was in the context of well-structured problem-solving, while primarily targeting collaborative dimensions of CPS. The problem-solving performance in this task was strongly influenced by the ability of the students to apply a vary-one-thing-at-a-time strategy (Vollmeyer and Rheinberg 1999), which is also known as control of variables strategy (Chen and Klahr 1999). This is a method for creating experiments in which a single contrast is made between experimental conditions. This strategy is suggested by the agent in the logfile example in Table 2 (“Let’s try to change one condition per trial”) and was part of a phrase-chat menu that was available for a student “collaborator” in H-H settings. While the computer agent was programmed to suggest this strategy to each participant in a standardized way (before the second submission), there was no control over the suggestions made by the human partner. However, as shown in this study, participants in H-A mode do not outperform participants in H-H mode in their problem-solving score, while the major difference between the students’ performance in H-H and H-A settings are the collaboration-related skills. Interdependency is a central property of tasks that are desired for

assessing collaborative problem solving, as opposed to a collection of independent individual problem solvers. A task has higher interdependency to the extent that student A cannot solve a problem without actions of student B. Although, interdependency between the group members was required and observable in the studied CPS task, the collaboration in both settings was characterized by asymmetry of roles. A “leader” student in the H-H setting and the student in the H-A setting were in charge of selecting the variables and submitting the solutions in addition to the ability to communicate with the partner. According to Dillenbourg (1999), asymmetry of roles in collaborative tasks could affect each team member’s performance. Thus, a possible explanation for these results is the asymmetry in roles between the “leader” student and the “collaborator” in the H-H setting and the student and the computer agent in the H-A setting. In a more controlled setting (i.e., H-A) the asymmetrical nature of collaboration was associated with no relationship to the quality of collaborative skills that were observed during the task. It is possible, that while in the H-H setting, in which the human “collaborator” was functioning with no system control over the suggestions that he or she made, the asymmetry in roles was associated with the quality of collaborative skills that were observed during the task. The qualitative results of the study on differences in disagreement situations provide some initial insights into this possible pattern. A major factor that contributes to the success of CPS and differentiates it from individual problem solving is the role of communication among team members (Fiore and Schooler 2004; Dillenbourg and Traum 2006; Fiore et al. 2010). Communication is essential for organizing the team, establishing a common ground and vision, assigning tasks, tracking progress, building consensus, managing conflict, and a host of other activities in CPS.

The difficulty of the problems varies in addition to challenges in collaboration. Problem difficulty increases for ill-defined problems over well-defined problems, dynamic problems (that change during the course of problem solving) over static problems, problems that are a long distance versus a short distance from the given state to the goal state, a large problem space over a small space, the novelty of the solution, and so on. It is conceivable that the need for effective collaboration would increase as a function of the problem difficulty. More cognitively challenging problem-solving space in CPS tasks can possibly lead to differential results in H-A and H-H settings. So an important question to raise is: “How well would the results found here generalize to ill-structured problem-solving assessment context?” Specifically, do the similarities and differences between H-A and H-H group performance found in this study overestimate what would be found in other assessment contexts? Future studies could consider exploring differences in student performance in a wide range of problem-solving complexity and ill-structured tasks that cannot be solved by a single, competent group member. Such tasks require knowledge, information, skills, and strategies that no single individual is likely to possess. When ill-structured tasks are used, all group members are more likely to participate actively, even in groups featuring a range of student ability (Webb et al. 1998).

Concerning the level of motivation and time-on-task in collaborating with a computer agent or a human partner on CPS assessment task, we found no evidence for differences between the two modes. In other words, students felt motivated and efficient in collaborative work with computer agents just at the same level as when collaborating with their peers. Previous research found that examinee motivation

tended to predict test performance among students in situations in which the tests had low or no stakes for the examinees (Sundre 1999; Sundre and Kitsantas 2004; Wise and DeMars 2005). To the degree to which students do not give full effort to an assessment test, the resulting test scores will tend to underestimate their levels of proficiency (Eklöf 2006; Wise and DeMars 2006). We believe that two major factors in computer agent implementation contributed to student motivation in CPS assessment tasks. On the one hand, the student and the agent shared the responsibility to collaborate in order to solve the problem. A computer agent was capable to generate suggestions to solve the problem (e.g., “Let’s change one condition per trial”) and communicate with the student in a contextual and realistic manner. On the other hand, a shared representation of the problem-solving space was implemented to provide a concrete representation of the problem state (i.e., life expectancy) and the selections made (e.g., selection of the conditions).

The Role of Computer agents in Collaborative Assessment

Collaboration on tasks permits problem solving to occur by shared understandings, coordinated actions, and reflections on progress as the group comes to a solution. This includes an ability to identify the mutual knowledge (what each other knows about the problem), to identify the perspectives of other agents in the collaboration, and to establish a shared vision of the problem states and activities. An agent has the capability of generating goals, performing actions, communicating messages, sensing its environment, adapting to changing environments, and learning. The design of some agent-based technologies is flexibly adaptive to the point where no two conversations are ever the same, just as is the case of some interactions among humans. The space of alternative conversations can be extremely large even when there are a limited number of fixed actions or discourse moves at each point in a conversation. For those problems that change over time, the agents may adapt to these changes, just as some humans do. Agent-based systems often implement context-sensitive rules, processes, and strategies that are routinely adopted in models of human cognition and social interaction (Anderson 1990; Newell and Simon 1972). These capabilities are incorporated in the development of CPS assessment tasks with agents.

The focus of this research is an assessment of individuals in a collaborative problem solving context. An adequate assessment of a student’s CPS skills requires interactions with multiple groups in order to cover the constructs critical for assessment. If a student is matched with a problematic group of peers, then there will be no valid measurement of the constructs. However, computer environments can be orchestrated for students to interact with different agents, groups, and problem constraints to cover the range situated items required to have meaningful assessment (for instance, a situation in which a student supervises the work of agents, where there is an asymmetry in roles and status). The assessment tasks will include simulations of disagreements between agents and the student, collaboratively-orientated team members (e.g., initiates ideas, consensus-builder, and supports and praises other team members), as well as team members with low collaborative orientation (e.g., interrupts other members of the team, comments negatively about work of others) to provide a sufficient dataset for CPS assessment. It would be impractical to organize several groups of humans to meet the required standardized measurement of student CPS proficiency levels in the diverse set of situations that are needed for an appropriate coverage of the CPS constructs.

In summary, a computer agent approach satisfies the definitions of collaboration, and allows adequate control, standardization, efficiency, and precision in CPS assessment.

Research Limitations and Directions for Future Studies

The current study had several limitations. First, it is based on a relatively small and non-representative sample of 14-years-old students in three countries. However, due to lack of empirical research in the field of computer-based assessment of CPS skills, it is necessary to conduct small-scale pilot studies in order to inform a more comprehensive approach of CPS assessment. Further studies could consider including a representative sample of randomly assigned students in a wider range of ages and backgrounds. Second, the study operationalized the communication between the partners in CPS through a phrase-chat to ensure standardization and automatic scoring, while other approaches could be considered, including verbal conversations and open-chat. Third, it is possible that the comparability findings between H-A and H-H performance in other problem-solving and collaboration contexts will be different. Future studies could consider exploring differences in student performance in a wide range of problems and collaboration methods. Finally, the study implemented a certain set of measures and techniques to assess CPS. Various research methodologies and measures developed in previous studies of CPS, collaborative learning, and teamwork processes potentially can be adapted to CPS assessment (e.g., Biswas et al. 2005; Hsieh and O'Neil 2002; O'Neil and Chuang 2008; Rosen and Rimor 2012; Weinberger and Fischer 2006).

Conclusions

CPS assessment methods described in this article offer one of the few examples today of a direct, large-scale assessment targeting social and collaboration competencies. CPS brings new challenges and considerations for the design of effective assessment approaches because it moves the field beyond standard item design tasks. The assessment must incorporate concepts of how humans solve problems in situations where information must be shared and considerations of how to control the collaborative environment in ways sufficient for valid measurement of individual and team skills. The quality and practical feasibility of these measures are not yet fully documented. However, these measures rely on the abilities of technology to engage students in interaction, to simulate others with whom students can interact, to track students' ongoing responses, and to draw inferences from those responses. Group composition is one of the important issues in large-scale assessments of collaborative skills (Webb 1995; Wildman et al. 2012). Overcoming possible bias of differences across groups by using computer agents or other methods becomes even more important within international large-scale assessments where cultural boundaries are crossed. The results of this study suggest that computer agent technique is one of the promising directions for CPS assessment tasks that should be further explored. However, as discussed in this article, each mode of collaboration involves limitations and challenges. Further research is needed in order to establish comprehensive validity evidence and generalization of findings both in H-A and H-H CPS settings.

References

- Adejumo, G., Duimering, R. P., & Zhong, Z. (2008). A balance theory approach to group problem solving. *Social Networks*, *30*(1), 83–99.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale: Erlbaum.
- Aronson, E., & Patnoe, S. (2011). *Cooperation in the classroom: The jigsaw method*. London: Pinter & Martin, Ltd.
- Avouris, N., Dimitracopoulou, A., & Komis, V. (2003). On analysis of collaborative problem solving: an object-oriented approach. *Computers in Human Behavior*, *19*(2), 147–167.
- Barth, C. M., & Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion*, *24*(7), 1259–1268.
- Biswas, G., Leelawong, K., Schwartz, D., & Vye, N. (2005). Learning by teaching: a new agent paradigm for educational software. *Applied Artificial Intelligence*, *19*, 363–392.
- Biswas, G., Jeong, H., Kinnebrew, J. S., Sulcer, B., & Roscoe, A. R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology-Enhanced Learning*, *5*(2), 123–152.
- Cai, Z., Graesser, A. C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., Halpern, D., & Butler, H. (2011). Trialog in ARIES: User input assessment in an intelligent tutoring system. In W. Chen & S. Li (Eds.), *Proceedings of the 3rd IEEE international conference on intelligent computing and intelligent systems* (pp. 429–433). Guangzhou: IEEE Press.
- Chen, Z., & Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120.
- Chung, G. K. W. K., O'Neil, H. F., Jr., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior*, *15*, 463–494.
- Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.
- Cooke, N. J., Kiekel, P. A., Salas, E., Stout, R., Bowers, C., & Cannon-Bowers, J. (2003). Measuring team knowledge: a window to the cognitive underpinnings of team performance. *Group Dynamics: Theory, Research and Practice*, *7*(3), 179–219.
- Dillenbourg, P. (Ed.). (1999). *Collaborative learning: Cognitive and computational approaches*. Amsterdam, NL: Pergamon, Elsevier Science.
- Dillenbourg, P., & Traum, D. (2006). Sharing solutions: persistence and grounding in multi-modal collaborative problem solving. *The Journal of the Learning Sciences*, *15*(1), 121–151.
- Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, *66*(4), 643–656.
- Fall, R., Webb, N., & Chudowsky, N. (1997). *Group discussion and large-scale language arts assessment: Effects on students' comprehension. CSE Technical Report 445*. Los Angeles: CRESST.
- Fiore, S., & Schooler, J. W. (2004). Process mapping and shared cognition: Teamwork and the development of shared problem models. In E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 133–152). Washington DC: American Psychological Association.
- Fiore, S., Rosen, M., Smith-Jentsch, K., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: predicting process in complex collaborative contexts. *The Journal of the Human Factors and Ergonomics Society*, *53*(2), 203–224.
- Foltz, P. W., & Martin, M. J. (2008). Automated communication analysis of teams. In E. Salas, G. F. Goodwin, & S. Burke (Eds.), *Team effectiveness in complex organizations and systems: Cross-disciplinary perspectives and approaches* (pp. 411–431). New York: Routledge.
- Franklin, S., & Graesser, A. C. (1996). Is it an agent or just a program? A taxonomy for autonomous agents. In *Proceedings of the Agent Theories, Architectures, and Languages Workshop* (pp. 21–35). Berlin: Springer-Verlag.
- Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cognitive Processing*, *11*, 133–142.
- Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes*, *45*(4), 298–322.
- Graesser, A. C., Foltz, P., Rosen, Y., Forsyth, C., & Germany, M. (2015). Challenges of assessing collaborative problem solving. In B. Csapo, J. Funke, & A. Schleicher (eds.), *The nature of problem solving*. OECD Series.
- Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching 21st century skills* (pp. 1–15). Heidelberg: Springer.

- Hsieh, I.-L., & O'Neil, H. F., Jr. (2002). Types of feedback in a computer-based collaborative problem solving group task. *Computers in Human Behavior*, *18*, 699–715.
- Laurillard, D. (2009). The pedagogical challenges to collaborative technologies. *International Journal of Computer-Supported Collaborative Learning*, *4*(1), 5–20.
- Leelawong, K., & Biswas, G. (2008). Designing learning by teaching systems: the betty's brain system. *International Journal of Artificial Intelligence in Education*, *18*(3), 181–208.
- 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). New York: Macmillan Library Reference USA, Simon & Schuster Macmillan.
- Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 287–304). Mahwah: Lawrence Erlbaum Associates.
- Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C., & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & J. Lakhmi (Eds.), *Serious games and edutainment applications* (pp. 169–195). London: Springer.
- Mitchell, R., & Nicholas, S. (2006). Knowledge creation in groups: the value of cognitive diversity, transactive memory and open-mindedness norms. *The Electronic Journal of Knowledge Management*, *4*(1), 64–74.
- National Research Council. (2011). *Assessing 21st century skills*. Washington: National Academies Press.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs: Prentice-Hall.
- O'Neil, H. F., Jr., & Chuang, S. H. (2008). Measuring collaborative problem solving in low-stakes tests. In E. L. Baker, J. Dickieson, W. Wulfecck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 177–199). Mahwah: Lawrence Erlbaum Associates.
- O'Neil, H. F., Jr., Chung, G. K. W. K., & Brown, R. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O'Neil Jr. (Ed.), *Workforce readiness: Competencies and assessment* (pp. 411–452). Mahwah: Lawrence Erlbaum Associates.
- O'Neil, H. F., Jr., Chen, H. H., Wainess, R., & Shen, C. Y. (2008). Assessing problem solving in simulation games. In E. L. Baker, J. Dickieson, W. Wulfecck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 157–176). Mahwah: Lawrence Erlbaum Associates.
- O'Neil, H. F., Jr., Chuang, S. H., & Baker, E. L. (2010). Computer-based feedback for computer-based collaborative problem solving. In D. Ifenthaler, P. Pimay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 261–279). New York: Springer.
- OECD (2013). *PISA 2015 Collaborative problem solving framework*. OECD Publishing.
- Rimor, R., Rosen, Y., & Naser, K. (2010). Complexity of social interactions in collaborative learning: the case of online database environment. *Interdisciplinary Journal of E-Learning and Learning Objects*, *6*, 355–365.
- Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem-solving. In C. E. O'Malley (Ed.), *Computer-supported collaborative learning* (pp. 69–97). Berlin: Springer.
- Rosé, C. P., & Torrey, C. (2005). Interactivity and expectation: Eliciting learning oriented behavior with tutorial dialogue systems. In *Human-computer interaction-INTERACT 2005* (pp. 323–336). Springer Berlin Heidelberg.
- Rosen, Y. (2009). Effects of an animation-based on-line learning environment on transfer of knowledge and on motivation for science and technology learning. *Journal of Educational Computing Research*, *40*(4), 451–467.
- 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), 147–174.
- Rosen, Y., & Beck-Hill, D. (2012). Intertwining digital content and a one-to-one laptop environment in teaching and learning: lessons from the Time To Know program. *Journal of Research on Technology in Education*, *44*(3), 225–241.
- Rosen, Y., & Rimor, R. (2009). Using collaborative database to enhance students' knowledge construction. *Interdisciplinary Journal of E-Learning and Learning Objects*, *5*, 187–195.
- Rosen, Y., & Rimor, R. (2012). Teaching and assessing problem solving in online collaborative environment. In R. Hartshorne, T. Heafner, & T. Petty (Eds.), *Teacher education programs and online learning tools: Innovations in teacher preparation* (pp. 82–97). Hershey: Information Science Reference, IGI Global.
- Scardamalia, M. (Ed.). (2002). *Collective cognitive responsibility for the advancement of knowledge*. Chicago: Open Court.
- Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. Cambridge: MIT Press.
- Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

- Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: can examinee self-regulation and test-taking motivation predict consequential and nonconsequential test performance? *Contemporary Educational Psychology*, 29(1), 6–26.
- U.S. Department of Education. (2010). *Transforming American education – learning powered by technology: National Education Technology Plan 2010*. Washington: Office of Educational Technology, U.S. Department of Education.
- Vollmeyer, R., & Rheinberg, F. (1999). Motivation and metacognition when learning a complex problem. *European Journal of Psychology of Education*, 14, 541–554.
- Webb, N. M. (1995). Group collaboration in assessment: multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis*, 17(2), 239–261.
- Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: group composition and performance. *American Educational Research Journal*, 35(4), 607–651.
- Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46, 71–95.
- Wildman, J. L., Shuffler, M. L., Lazzara, E. H., Fiore, S. M., Burke, C. S., Salas, E., & Garven, S. (2012). Trust development in swift starting action teams: a multilevel framework. *Group & Organization Management*, 37(2), 138–170.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment*, 10(1), 1–17.
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: the effort-moderated IRT model. *Journal of Educational Measurement*, 43(1), 19–38.
- Zhang, J. (1998). A distributed representation approach to group problem solving. *Journal of the American Society for Information Science*, 49, 801–809.