



Measuring collaborative problem solving: research agenda and assessment instrument

Anita Pásztor-Kovács ^a, Attila Pásztor ^{a,b} and Gyöngyvér Molnár ^a

^aInstitute of Education, University of Szeged, Szeged, Hungary; ^bMTA-SZTE Research Group on the Development of Competencies, University of Szeged, Szeged, Hungary

ABSTRACT

In this paper, we present an agenda for the research directions we recommend in addressing the issues of realizing and evaluating communication in CPS instruments. We outline our ideas on potential ways to improve (1) generalizability in Human–Human assessment tools and ecological validity in Human–Agent ones; (2) flexible and convenient use of restricted communication options; and (3) an evaluation system of both Human–Human and Human–Agent instruments. Furthermore, in order to demonstrate possible routes for realizing some of our suggestions, we provide examples through an introduction of the features of our own CPS instrument. It is a Human–Human pre-version of a future Human–Agent instrument and a promising diagnostic and research tool in its own right, as well as the first example of transforming the so-called MicroDYN approach so that it is suitable for Human–Human collaboration. We offer new alternatives for communication in addition to pre-defined messages within the test, which are also suitable for automated coding. For example, participants can send or request visual information in addition to verbal messages. As regards evaluation as a hybrid solution, not only are the pre-defined messages proposed as indicators of different CPS skills, but so are a number of behavioural patterns.

ARTICLE HISTORY

Received 22 October 2020
Accepted 22 October 2021

KEYWORDS

Collaborative problem solving; Human–Human design; Human–Agent design; automated coding; constrained communication; MicroDYN approach

Teamwork offers numerous advantages in the case of solving complex problems: increased coverage of knowledge, skills and ideas becomes available through the members of a team (Graesser et al., 2018a; Rosen et al., 2020). The potential of teamwork has been widely recognized and used in the workforce in recent decades; therefore, the ability to effectively solve problems in collaboration with others represents a continuously growing value (Binkley et al., 2012; Fiore et al., 2017; Fiore & Wiltshire, 2016). Consequently, it is a highly significant aim for a school-leaver to be competent in working in groups and in creating solutions to particular problems collaboratively (Fiore et al., 2018). Thus, it is necessary to develop collaborative problem solving in an educational context. To be able to monitor the development of this skill, we need effective instruments (Fiore & Kapalo, 2017).

Technology-based assessment both in the case of large-scale measurements and everyday school practice is an obvious choice due to the number of advantages (e.g. a higher level of objectivity, the possibility of using innovative item types, such as audio and video files or interactive elements, and the reduced need for human resources to register and code data; Csapó et al., 2012). However, we

CONTACT Anita Pásztor-Kovács  pasztor-kovacs@edpsy.u-szeged.hu

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

are faced with a number of methodological challenges if we want to assess CPS skills on an individual level with a technology-based instrument (Graesser et al., 2018b; Liu et al., 2016).

This paper presents a research agenda with possible solutions for dealing with the issues of realizing and evaluating communication in CPS instruments. We outline ideas for potential ways to improve (1) generalizability in Human–Human assessment tools and ecological validity in Human–Agent ones; (2) flexible and convenient use of restricted communication options; and (3) an evaluation system for both Human–Human and Human–Agent instruments. Furthermore, in order to demonstrate possible routes for realizing certain suggestions, we provide examples through an introduction to the features of a new CPS instrument. It is a Human–Human pre-version of a future Human–Agent instrument and a promising diagnostic and research tool in its own right, as well as the first example of a transformation of the so-called MicroDYN approach suitable for Human–Human collaboration. To provide test-takers with the widest interaction space possible and also to reduce their frustration at the lack of free chat to a minimum, we offer new alternatives for communication in addition to pre-defined messages within the test, which are also suitable for automated coding. For example, participants can send or request visual information in addition to verbal messages. As regards evaluation, as a hybrid solution, not only are the pre-defined messages proposed as indicators of different CPS skills, but so are a number of behaviour patterns.

Defining collaborative problem solving

Collaborative problem solving refers to “problem-solving activities that involve interactions among a group of individuals” (O’Neil et al., 2003, p. 4; Zhang, 1998, p. 1). In a more detailed definition, “CPS in educational setting is a process in which two or more collaborative parties interact with each other to share and negotiate ideas and prior experiences, jointly regulate and coordinate behaviours and learning activities, and apply social strategies to sustain the interpersonal exchanges to solve a shared problem” (Dingler et al., 2017, p. 9). The theoretical models of CPS name different skills and subskills. These subskills also differ in their arrangement; they are ordered in a hierarchy as well as in a matrix (Graesser et al., 2018a; Sun et al., 2020). There is one aspect, however, which seems to be common in the rest of the models (Dingler et al., 2017; Hesse et al., 2015; Liu et al., 2016; OECD, 2017; O’Neil et al., 2003): they contain two major elements representing a social or collaborative and a cognitive or problem-solving aspect of the construct. In our research we accept the two-dimensional concept of the skill and base our evaluation model on a cognitive and a social component.

Recent research on complex problem solving has defined two empirically distinct stages in problem-solving processes (Fischer et al., 2012; Fischer et al., 2015). In the first, knowledge acquisition phase, the problem solver systematically generates information, integrates this information into a viable mental model of the situation and selectively focuses on the most relevant, central and urgent aspects of the problem. In the second, knowledge application part he/she makes a set of interdependent decisions based on the explicit and implicit knowledge acquired, and monitors the prerequisites and consequences of these decisions continuously in order to systematically solve the problem at hand (Fischer et al., 2017, p. 112). Problem solving is described with these two phases in our instrument.

As regards the collaborative dimension, the CRESST teamwork model has been used to define the construct (O’Neil et al., 1997). The model consists of six skills: (1) *adaptability* refers to the skill of “monitoring the source and nature of a problem through an awareness of team activities and factors bearing on the task”; (2) *coordination* is required for the “process by which team resources, activities and responses are organized to ensure that tasks are integrated, synchronized, and completed with established temporal constraints”; (3) *decision making* represents “the ability to integrate information, use logical and sound judgment, identify possible alternatives, select the best solution, and evaluate the consequences”; (4) *interpersonal* ability is for improving “the quality of team member interactions through the resolution of team members’ dissent, or the use of cooperative behaviour”; (5) *leadership* means “the ability to direct and coordinate the activities of other team

members, assess team performance, assign tasks, plan and organize, and establish a positive atmosphere”; (6) *communication* is “the process by which information is clearly and accurately exchanged between two or more team members in the prescribed manner and by using proper terminology, and the ability to clarify or acknowledge the receipt of information” (Hsieh & O’Neil, 2002, p. 703).

Issues in realizing and evaluating interactions in collaborative problem solving instruments

The Human–Agent vs. Human–Human discussion

After the individual interactive problem-solving assessment in 2012, the OECD decided that problem-solving skills would be assessed again in 2015 (Csapó & Funke, 2017). However, this time the focus of the assessment was the individual’s capacity for solving problems collaboratively instead of on his or her own. This choice involved serious methodological issues. One of the biggest questions was the way in which comparable data should be produced. To be able to produce such data, every student should be tested in the exact same context: working on the same tasks with the same team members.

This design may seem impossible to achieve at first sight. Technology, however, offers a creative and heretofore unique solution: the application of computer agents as collaborators. In a technology-based assessment context, where the collaborating peer is not another person, but a conversational agent, it becomes possible to develop a standardized test environment, as agents can generate their reactions from the same pre-programmed set of responses to every test taker. The OECD decided to accept this choice, the so-called Human–Agent (H–A) approach, in the PISA survey, as well as several other CPS instruments later on (He et al., 2017; OECD, 2017; for further examples of H–A CPS instruments, see Krkovic et al., 2016; Rosen & Foltz, 2014; Stoeffler et al., 2020).

While the option of providing a standardized test environment in the H–A condition is quite crucial, the ecological validity of these instruments has been debated. Those who support the Human–Human (H–H) line of assessment, in which collaborators are actual humans, point out that the H–A solution is far from being realistic. One could hardly expect a computer agent to show the sort of broad range of feelings or sometimes rather irrational thinking that may characterize a human participant (Andrews-Todd & Forsyth, 2020; Care & Griffin, 2017; Griffin & Care, 2015; Scoular & Care, 2020; Yuan et al., 2019).

Recent CPS assessments, including the PISA survey, have used so-called minimalist agents, which only provided an opportunity for restricted written communication with pre-defined messages (Graesser, Dowell, et al., 2017; Rosen & Mosharraf, 2016). This strongly controlled design apparently does not contribute to creating a markedly realistic digital collaborator. Moreover, they employed pre-determined chat, which means the agent tightly restricted the conversation by dynamically changing the set of 3–5 pre-defined messages offered for exchange (He et al., 2017; OECD, 2017; Rosen & Foltz, 2014; Rosen & Mosharraf, 2016). This kind of conversation seems even farther from an authentic H–H interaction.

To investigate the question of whether collaboration with computer agents can be considered equivalent to the joint problem-solving process used by real students, two validity studies have been carried out so far. In the first research, the achievement of H–A and H–H dyads on the same CPS tasks was compared (Rosen & Foltz, 2014). In the second, PISA validation study (Herborn et al., 2020), PISA problems were used which students had originally solved together with two or three computer agents. In the validation study, one agent was replaced by a human student in half of the cases.

Neither of these studies found sizeable differences in achievement between groups in which students were collaborating with agents exclusively compared to those in which students had a human collaborator. This result seems promising at first glance. It should be noted, however, that students were only able to collaborate through a very limited list of pre-defined messages in H–A mode. Ironically, validating H–A CPS instruments itself is a big challenge as it involves a paradox. The most

informative method would be to compare the results of H–A groups to those of H–H groups in which students are allowed to have an open-chat discussion. However, if the type of communication is changed, the comparability of the two conditions becomes compromised, as we are no longer dealing with the same tasks. Stadler et al. (2020) followed a different line for validation. They correlated the students' test results on PISA tasks with self-rated and teacher-rated scales on their collaborative skills and found moderate correlation. This result again provides a reason for optimism, although here too we have to have some reservations because of the classic objectivity issues with questionnaires.

Automated data coding

Communication between collaborators contains core information about the problem-solving process and participants' CPS skills, so how it is realized and what solutions can be found to evaluate it represent a key issue. One of the greatest advantages of technology-based assessment is the option of automated coding (Csapó et al., 2012). With regard to CPS, both large-scale assessments and everyday educational practice would require instruments which generate results that can be coded automatically, as teachers are not necessarily experts on methods of analysing human discourse.

Open-ended communication can undoubtedly be considered as the most valid way of exchanging ideas. However, evaluating open-ended discussions, especially in the case of large-scale assessments, can be extremely resource- and time-consuming. The reason is that content analysis, the traditional method for analysing interactions, cannot be implemented at the current stage of technology with the complete elimination of human rating (Care et al., 2015).

To handle this case, a possible option is to apply natural language processing (NLP) to evaluate interactions (Hao et al., 2017; Landauer et al., 2013; Liu et al., 2016). Recent analytical software is capable of processing a discourse based on syntactic characteristics; furthermore, with an embedded vocabulary, it can search for predicted keywords and phrases in the discourse (Dowell et al., 2019; Graesser, Dowell, et al., 2017; Graesser, Forsyth, et al., 2017; Reilly & Schneider, 2019; Rosen & Mosharaf, 2016; Rosé et al., 2017). These technologies represent significant steps toward the future aim of understanding the semantics of the interactions using artificial intelligence tools. Nevertheless, processing the content with NLP methods is still a current problem to be solved, especially in the case of agglutinative languages, such as Hungarian, Turkish and Japanese, where the large number of possible word forms obtained from one root makes syntactic analysis extremely challenging.

Another alternative to automated data coding of free chat has been to develop behavioural indicators based on an analysis of a large number of H–H problem-solving interactions and create algorithms which search for these in the log stream data (Adams et al., 2015; Griffin & Care, 2015; Scoular & Care, 2020; Yuan et al., 2019). The indicators have been related to the presence or absence of specific actions, for instance, asking a question or sending a message before entering a final answer to a problem. While this analytical method represents another hopeful automated coding approach for future CPS measurement tools, it is not appropriate for capturing the content of the communication satisfactorily.

The third solution for handling the complicated case of automated coding in CPS instruments was to eliminate the option of conversing freely. More specifically, group members can only talk by exchanging a set of pre-defined messages, which are previously assigned to different skills, so automated data coding can be developed and implemented with this pre-assignment. This option can lead to more valid results than the application of behavioural indicators, as it opens the door to taking the actual content of the interaction into account in the evaluation. However, in addition to this great advantage, this alternative also has its shortcomings.

Pre-defined message exchange

Research in the last twenty years has demonstrated that pre-defined message exchange has proved to be an effective way to interact with the aim of problem solving (Chung et al., 1999; Hsieh & O'Neil,

2002; Krkovic et al., 2016; OECD, 2017; O’Neil et al., 1997; Rosen & Foltz, 2014; Stoeffler et al., 2020). Students were capable of completing the tasks under this condition, in some ways actually more effectively than in the case of chatting freely, as a significant amount of off-task discussion was excluded with this restriction (Chung et al., 1999). However, in studies where they had the chance to provide feedback on changing pre-defined messages, they continuously stressed how much they missed the option of formulating their own messages (Chung et al., 1999). Also, if they had the opportunity to send both types of messages, they started to ignore the chance to send pre-defined messages quickly and shifted to typing in their own (O’Neil et al., 1997). Therefore, besides its effectiveness for problem solving and automated coding, pre-defined message exchange also proved to have its limits: it may lead to frustration, as participants may be disturbed at not being able to express themselves if the messages fail to cover every possible scenario for talk (Krkovic et al., 2014).

Pre-defined messages are basically provided for test-takers in two ways at present. The first way is the already mentioned pre-determined way in H–A approaches, when the set changes turn by turn based on the script the agent follows (OECD, 2017; Rosen & Foltz, 2014). This design is a highly feasible choice from the perspective of scoring. It is obviously much less difficult to create a coding scheme if the human participant has only 3–5 messages to choose from in every conversational turn. One solution may be to evaluate every possible message within a turn using different scores, or, as in the case of the PISA survey, to technically implement a multiple-choice design and only give a score for one, “right” message (OECD, 2017; Rosen & Mosharraf, 2016; Scoular et al., 2017).

The second way is to provide the complete pre-defined message set constantly (Chung et al., 1999; Hsieh & O’Neil, 2002; Krkovic et al., 2016; O’Neil et al., 1997). If we choose the latter way, we are faced with a much more complex issue vis-à-vis the development of the coding system. It is not possible to make a well-substantiated decision on which CPS skill a pre-defined message mostly belongs to within a given framework without being aware of the context of the discussion. For example, the answer “No” should be evaluated differently after the question “Do you understand?” than it is after the request “Wait, please”. Consequently, if we assign the pre-defined messages to different CPS skills and base the automated evaluation on this pre-assignment without taking into account the line of the message exchange, the results may be completely invalid.

Proposed research priorities

Key directions in improving CPS instruments vis-à-vis the Human–Human vs. Human–Agent discussion

After almost a decade of discussion, it seems time to move beyond the question of which is the “right” assessment line to be followed. Clearly, both conditions have their advantages and disadvantages for different assessment situations. The H–H approach has greater potential for creating a detailed profile of one’s CPS skills, as these instruments can be a very rich source of data with a sophisticated quality of describing students’ CPS behaviour (Andrews-Todd & Forsyth, 2020; Scoular et al., 2017; Scoular & Care, 2020; Yuan et al., 2019). Nevertheless, such tools will not be able to offer a standardized test environment, so the data produced by them will never be impeccable in terms of generalizability. H–A approaches, on the other hand, while they can ensure the latter feature perfectly, are unable to reach the ecological validity of H–H instruments. Both assessments have their own advantages depending on the aim of the assessment, so we believe both instrument types are worth investing in. There is much potential for improvement in the case of both approaches: increasing the generalizability level of H–H instruments and the ecological validity level of H–A instruments is a realizable aim.

To increase generalizability to a sensible level in H–H assessment tools, a potential solution would be to have the test-takers solve problems in groups with multiple team members. The more students

with different abilities with whom to collaborate, the greater space they have to manifest their CPS skills (Hao et al., 2017; Rosen, 2017). The problem with this solution is that in the case of written communication, which has been the choice of every CPS instrument, it is much more difficult for the student to follow the chat conversation if he/she needs to collaborate with more than one person. Interdependence can demand closely collaborative work, as team members cannot solve the problem without the others' contribution. Thus, it is very important to follow what kind of information has been shared and who has shared it. Experiences tied to group problem solving in written conversation show that as the conversation proceeds, it becomes increasingly difficult to search back in the chat window for the important moves, and this grows even more complex with an increasing number of team members (Fuks et al., 2006). This may explain why several H–H and even H–A approaches use the smallest unit of a group in their assessments, which is a dyad (Griffin & Care, 2015; Krkovic et al., 2016; Rosen & Foltz, 2014; Stoeffler et al., 2020; Yuan et al., 2019). We understand the practical advantage of involving pairs in H–H assessment tools; however, it is still possible to have a student collaborate with multiple partners within one CPS test. The solution may be that the test-taker works with a different collaborator on every problem.

In the case of H–A approaches, the biggest challenge at this point would be to improve the ecological validity of these instruments. The most significant step toward this aim should be to analyse prior H–H interactions on CPS tasks and then carefully base the agents' reactions on these interactions. What is more, the ideal route to maximizing ecological validity in an H–A instrument would actually be to create a H–H pre-version of that instrument first: we need to set up a productive H–H assessment tool initially and then develop an agent based on an analysis of the human interactions that have emerged. Following this solution, in the H–A version, both the student and the agent can use the conversational moves that were used in the H–H version. The openings and reactions of the agent in a conversational turn should track the most typical message exchanges identified in the human interactions.

Despite its feasibility, this stage of analysing human interactions on the given CPS tasks has been omitted from the developmental process for H–A CPS instruments. Notably, in case of those learning environments in which the agent serves as a tutor to the student, the relevance of human–human interaction analysis has been recognized. Some of these tutoring environments were based on an analysis of hundreds of hours of face-to-face tutor–student interactions and interactions between student groups and a mentor (Graesser, Dowell, et al., 2017). However, it is necessary to underline the great difference between the dynamics of peer–peer interaction and that of the tutor–student kind. The improvement of authenticity demands an analysis of peer–peer collaboration.

Key directions in improving constrained communication within CPS instruments

As we outlined above, the advantage of automated coding is so essential that it should be exploited in every CPS instrument, whether it is H–H or H–A. At the current stage of technology, we find constrained communication to be the most feasible for this aim. One of the main tasks at this point is to maximize the flexibility and convenience of constrained communication and thus increase the ecological validity of interactions realized in a constrained way.

In raising the validity level of restricted communication, the key role of H–H interaction analysis should be stressed again, this time by specifically highlighting the condition of open-ended communication. In terms of flexibility and convenience it would be fundamental to base the pre-defined message set on an analysis of previous open-ended interactions on the problem-solving tasks within a CPS test. While this step may seem quite obvious, hardly any CPS instruments have implemented it in the development process (for exceptions, see Chung et al., 1999; O'Neil et al., 1997).

We see great potential in some alternative ways of constrained communication beyond pre-defined messages, which participants find convenient to use and satisfying in expressing themselves and which are still suited to automated coding. Consequently, it would be worth addressing the discovery of new constrained communication options in future research.

Furthermore, test-takers would need more of an opportunity to lead and initiate, a much bigger interaction space in general than they have in the case of the pre-determined chat design. If we eliminate free chat, it would be essential to provide as many options as possible for communication within a test environment without cognitive overload. To ensure this, we find the second way of presenting pre-defined messages, in which the complete message set is constantly available, more compelling.

Key directions in improving evaluation within CPS instruments

As we have outlined, the pre-determined chat design should be retained, in our view. This step, however, implies serious challenges for evaluation. In the case of a wider interaction space, possible options for interacting within a turn greatly increase. The multiple-choice design obviously cannot be implemented this way; moreover, the “context problem”, which means one should take into account the line of pre-defined messages, also needs to be addressed.

We noted that if we base the scoring on pre-defined message exchange or behavioural indicators exclusively, the results can be misleading. However, a combination of these two methods can lead to more valid results. In H–H assessment tools, we recommend a hybrid evaluation system: pre-defined messages complemented with behavioural indicators. An analysis of H–H discussions realized by pre-defined messages can enable us to identify meaningful sequences in the texts which should be taken into account in the coding scheme as behavioural indicators, either with a positive or a negative weight. This design can cover up the content of the communication on a greatly advanced level.

In the case of H–A approaches, the evaluation of the problem-solving process should lean strongly on the H–H versions of the instruments. In agent-based assessment tools, the conversation line should be segmented into different parts by significant milestones in the problem-solving process, and the agents would have a specific protocol to follow with reference to every different segment. These protocols would contain the agent’s script with the pre-programmed reactions to each possible human conversational move turn by turn. As the coding scheme of the H–H version would contain indicators referring to the content of the pre-defined messages as well as the specific line of some interactional moves, we could easily decide in the case of the H–A version which moves we will evaluate in a conversational turn and with what kind of weight.

Developmental stages in creating collaborative problem-solving instruments

Our recommendations list the necessary developmental stages of the CPS instruments, which can be summarized as follows:

In the case of H–H instruments,

- (1) the first version of the assessment tool should permit open-ended discussion;
- (2) pre-defined messages and further restricted communication options should be based on an analysis of data gathered via this first version, which permits open-ended discussion;
- (3) after eliminating free chat, restricted communication options should be tested in several further H–H tests;
- (4) if the restricted ways of communicating can be considered well-established, large-scale data collection will be necessary; next,
- (5) we can create the evaluation system by having created the behavioural indicators based on an analysis of the interactions;
- (6) the instrument, with the well-grounded, user-friendly restricted communication options which are suitable for automated coding and with a solid evaluation system, is ready to use, even with multiple members within a test.

The initial developmental stages of H–A instruments can be considered the same, as we believe they should be built on the H–H versions of themselves. After this is done,

- (7) we should define the problem-solving segments of the CPS tasks with specific milestones and identify the typical conversational turns in the different segments based on the interactions evolved in the H–H version; then,
- (8) the agent’s protocol should be created for each segment based on typical openings and reactions. The protocols should contain every possible route (the agent’s reactions to every potential human move);
- (9) to make sure students cannot run into dead-end discussions with the agent, the emerging H–A version should be tested a number of times;
- (10) after the necessary troubleshooting, we should create the evaluation system, strictly based on the H–H evaluation system;
- (11) as the final step in perfecting the H–A version, large-scale data collection is again recommended, in which students solve half of the problems using the H–H version and the other half with the H–A one. If the data gathered with the two versions correlate satisfactorily, we can consider our H–A instrument sound and ecologically valid.

In the next half of the paper, we demonstrate exemplary routes to realizing some of the proposed ideas by providing the CPS assessment tool we are currently developing.

The Human–Human version of a new CPS instrument based on the MicroDYN approach

In our research, we are developing a new online CPS measurement tool in the eDia electronic diagnostic assessment system (Csapó & Molnár, 2019; Molnár & Csapó, 2019). According to our research agenda, we are in the third stage of the developmental line of creating CPS instruments. In the initial steps of the development, we permitted open-ended discussions on the tasks so that the restricted communication options of the H–H instrument could be based on an analysis of data gathered via this first version. Currently, we are about to consolidate the new version, which uses restricted communication options exclusively, through small-scale trials (Pásztor-Kovács, 2018; Pásztor-Kovács et al., 2018). The instrument is suitable for later computer agent embedding; moreover, it aims to become a valuable H–H assessment tool in its own right.

Collaborators can converse through pre-defined messages, which are constantly available. Furthermore, to handle the case of inflexible communication, new alternatives have been developed for interaction within the platform. In addition to verbal messages, participants can send or request visual information during the problem-solving process. We thus aim to create a user-friendly test environment, which can reduce students’ potential frustration at the lack of free chat to a minimum.

The assessment system can assign students to groups either randomly, or, given a specific pedagogical aim, the composition of groups can also be pre-defined within a given sample. It is also possible to change group compositions task by task within one test. For example, on a four-task test, students have the chance to work with four different partners selected either randomly or in a pre-defined way. This solution is expected to create far more generalizable results than the design of having students collaborate in the same team throughout a test. Furthermore, for instance, in the case of a classroom assessment, where the teacher may have some presumptions of students’ CPS levels, it can be even more informative for him/her to combine the compositions of the dyads by himself/herself and choose the pre-defined way.

In the following, the constrained communication options of the instrument will be introduced; furthermore, we present the foundations of the evaluation system, involving pre-defined message exchange and several other activities as behavioural indicators.

The problem type – the MicroDYN approach

As we chose to make the complete pre-defined message set constantly available, a sort of message list was necessary which accurately covers possible interactions yet is still perspicacious and requires no advanced cognitive capacities to process. Toward this aim, we sought problem types in which the problem space is of a reasonable size and the possible stages and outcomes of the problem-solving process are relatively predictable. The problem-solving task types of the both theoretically and empirically well-grounded MicroDYN approach was found to be the best choice for this goal (Greiff et al., 2012; Wüstenberg et al., 2012).

The MicroDYN approach aims to assess individuals' interactive problem solving, which refers to their "ability to explore and identify the structure of (mostly technical) devices in dynamic environments by means of interacting and to reach specific goals" (Greiff & Funke, 2017, p. 95). The problems contain at least one, but as many as three input variables and also at least one, but as many as three interrelated output variables. They are content-general, with no prior knowledge required to solve them. The tests consisting of MicroDYN-based problems have proved to be reliable and valid in a number of studies, including the PISA 2012 problem-solving assessment (Fischer et al., 2017; Greiff & Funke, 2017).

The problems involve (a) the acquisition of knowledge of relevant aspects of the problem to ascertain the problem structure by interacting with a simulation and (b) the application of this knowledge to reach certain stages, more specifically, the goal values of the output variables in the simulation (Greiff & Funke, 2017). In the knowledge acquisition phase, students need to discover the relations by systematically manipulating the input variables. By moving the sliders linked to them and pressing the Apply button, they can observe the impact of the manipulation on the output variables by looking at the diagrams and graphs (Figure 1). After exploring the system, they need to build a model for the relations of the variables by drawing the suitable arrows between them (Figure 1). The right model earns a score. In the knowledge application phase, the correct model is already shown to the students. The task is to reach the target values of the output variables in four steps (e.g. by pressing the Apply button no more than four times; Figure 2). The score is earned if all the target values have been successfully reached. Both phases have a time limit within which to work.

A pioneer example of making MicroDYN tasks collaborative can already be found in the literature (Krkovic et al., 2016). In the COLBAS (computer-assisted assessment for collaborative behaviour) instrument, participants need to collaborate with a computer agent. COLBAS has been an inspiring model for us on how to transform the MicroDYN items for collaborative work: there are unavailable input and output variables for students in both instruments. To learn about their features, participants need to contact their partner.

Nevertheless, there are some fundamental differences between the two assessment tools. First of all, in COLBAS only the first, knowledge acquisition phase has been transformed, the knowledge application phase remained original. Furthermore, COLBAS gives the chance for open-ended besides restricted communication. Students can send questions and requests to the agent with pre-defined messages and use free chat to make assertions. However, while the pre-defined messages get a pre-defined response, the assertions in the open chat are not followed by a reaction from the agent. The content of the sent messages also stays unprocessed. The collaborative dimensions are manifested in the three speech acts: the scores represent the frequency of questioning, requesting and asserting (informing) during the problem-solving process. Thus, as its name says, it is more suitable for assessing collaborative behaviour besides problem solving than collaborative skills, which is the aim of our instrument. The most important difference is that COLBAS is not based on a human pre-version, which would be in our view a core requirement to raise the ecological validity of agent-based CPS assessment tools.

Consequently, while the two assessment tools are both built on the MicroDYN approach, our ideas about the necessary steps for transforming it for a collaborative environment are entirely

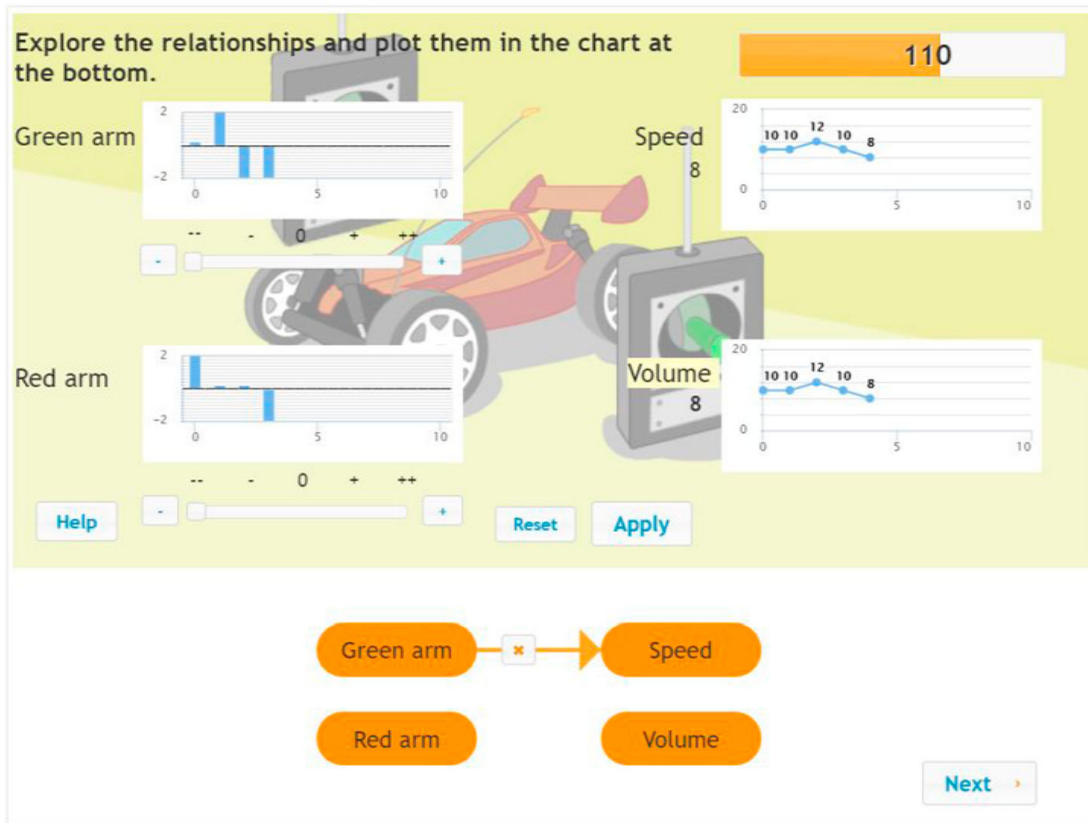


Figure 1. The knowledge acquisition phase of a MicroDYN problem in eDia. Students need to explore the effect of a green-armed and a red-armed remote control on the speed and volume of the racing car.

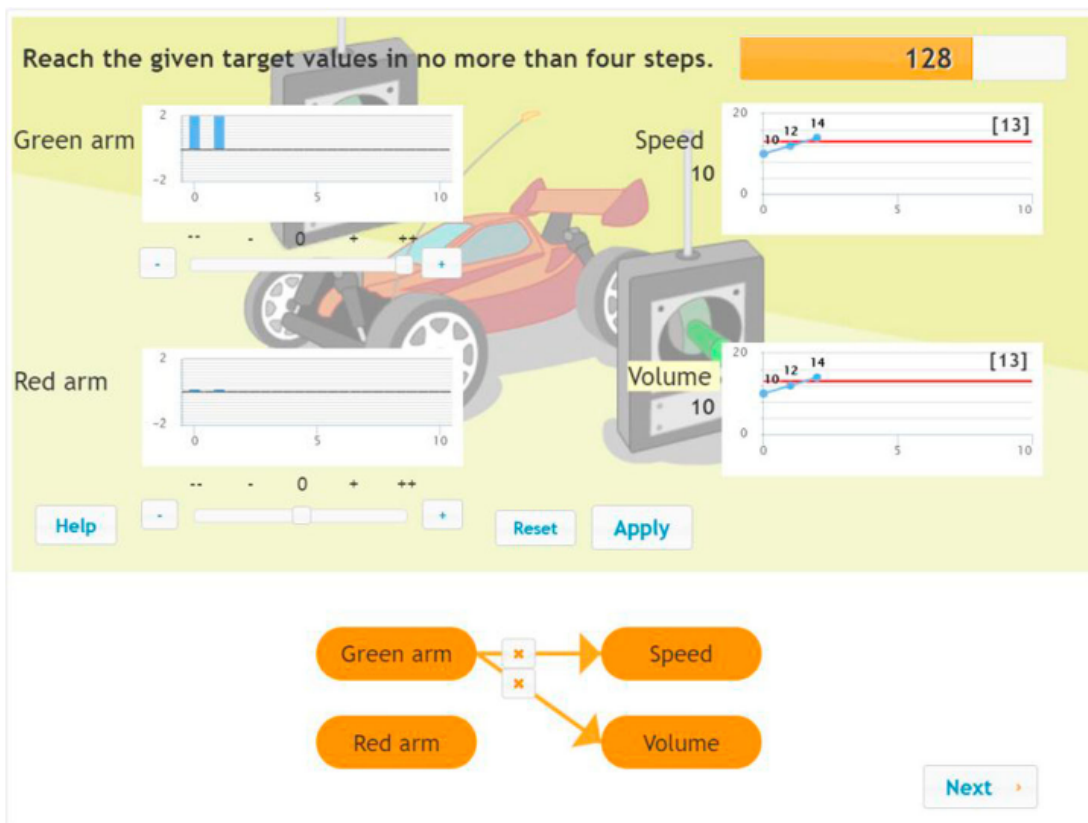


Figure 2. The knowledge application phase. The target values are placed next to the output graphs and represented by the red lines.

divergent. The H–H collaboration required a very different design with innovative restricted communication options. In the following sections, we outline the modifications implemented in the eDia platform, with a special focus on these options and the potential for their evaluation. As the knowledge acquisition and knowledge application phases required quite different modifications, they will be introduced in separate sections.

Restricted communication options in the knowledge acquisition phase

In the knowledge acquisition tasks, some input and output variables are unavailable for the team members. In the sample racing car problem in the figures, half of the variables are hidden from Student 1 and the other half from Student 2 (Figure 3). Let us assume that Student 1 is a boy and Student 2 is a girl. Student 1 is not able to see the change of the green arm diagram and the speed graph, and although he can move the slider for the green arm variable, pressing Apply will have no effect. Nevertheless, he has access to the red arm slider, the diagram and the volume graph. Student 2 is in the exact opposite situation. The available variables are always indicated with red frames for Student 1 and blue frames for Student 2, with the “frozen” ones being light grey. If, for example, Student 1 moves the red arm slider and presses Apply, he has no information on whether the speed graph has changed or not. Also, if he experiences a change on the volume graph, he cannot know whether it was the outcome of his manipulation on the red arm or if his peer manipulated the green arm slider and this was the reason for the change. To obtain this information, he needs to communicate with Student 2. To make it easier to learn who pressed Apply and when, Student 1’s applications are always indicated with a red spot on the active graphs and a red bar on the active diagrams for both students, while Student 2’s applications are indicated in blue (Figure 3).

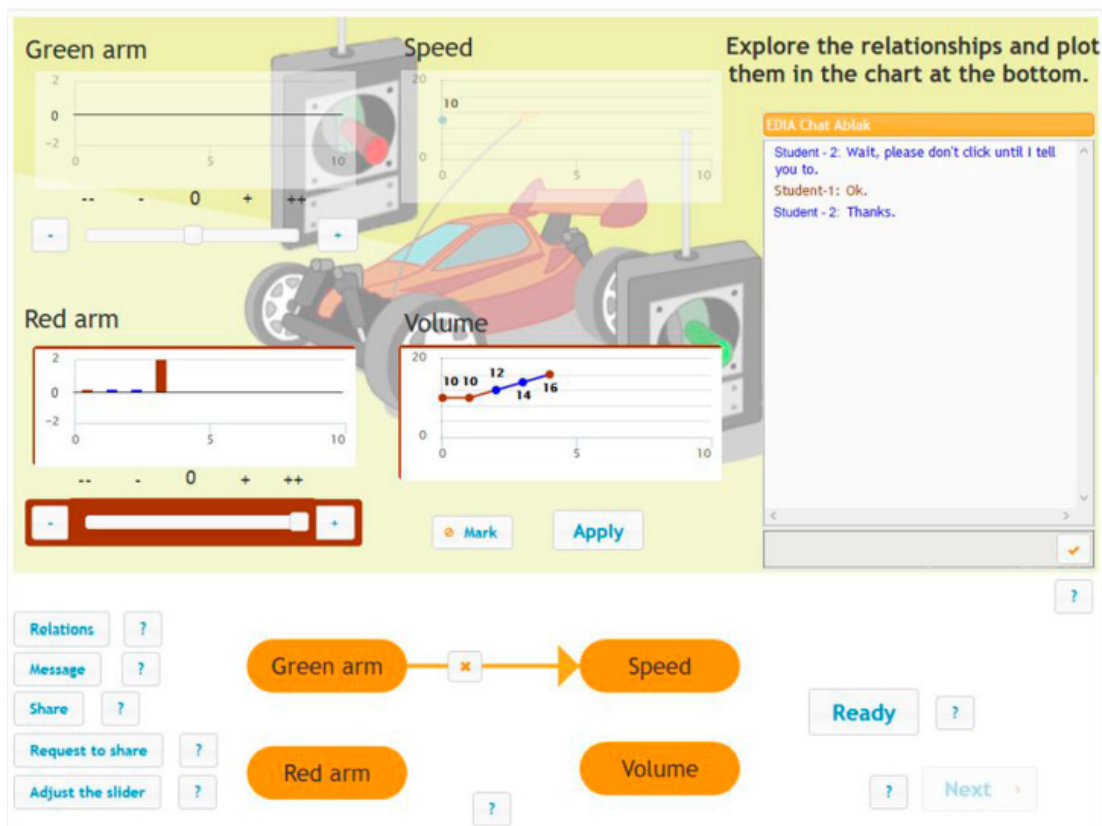


Figure 3. Student 1’s platform. The input and output variables at the top are unavailable. The two colours on the diagram and the graph show that both team members have already pressed Apply. The chat window shows that some pre-defined messages have been exchanged. (This is a translated version of the original task).

There are five buttons in the bottom left-hand corner of the surface, three of them for visual information exchange (Share, Request to share and Adjust the slider) and two of them for verbal information exchange (Message and Relations). The Message button provides the commonly used option of sending a pre-defined message. The messages are not specific to the problems; thus, the current twenty-five messages are the same in all problems in the knowledge acquisition phases. When the Message button is pressed, a pop-up window opens, containing the potential messages to be sent in two columns (Figure 4).

For verbal communication, we have created another, innovative option, in the form of the Relations button. This option was established to avoid the necessity of an extremely long, unprocessable list of pre-defined messages. Instead, through the alternative of the Relations button, peers can discuss the relations, manipulations and changes of the variables in numerous combinations. The solution of offering a short list (1–3) of optional elements of a statement has already been implemented in some studies (Chung et al., 1999; Hsieh & O’Neil, 2002). We have improved this method by creating a way of building the whole statement out of optional elements. If one presses the Relations button, a pop-up window offers 37–40 elements (depending on the number of variables) ordered in seven columns following the line of their supposed places in a potential statement (Figure 5). The elements chosen appear in the chat window from left to right in the order of their places in the columns (see Figure 6 for a seven-element and a three-element message). As the content behind this button is strongly related to the variables in a given task, it is different in every problem; however, it is the same in the knowledge acquisition and application phases within a problem.

The remaining three buttons, Share, Request to share and Adjust the slider, are again part of our latest innovation for constrained communication based on visual information exchange. For the two students to be able to jointly build the model, they need to share the current state of the active diagrams, graphs and sliders. However, talking about these states would again require a much too long

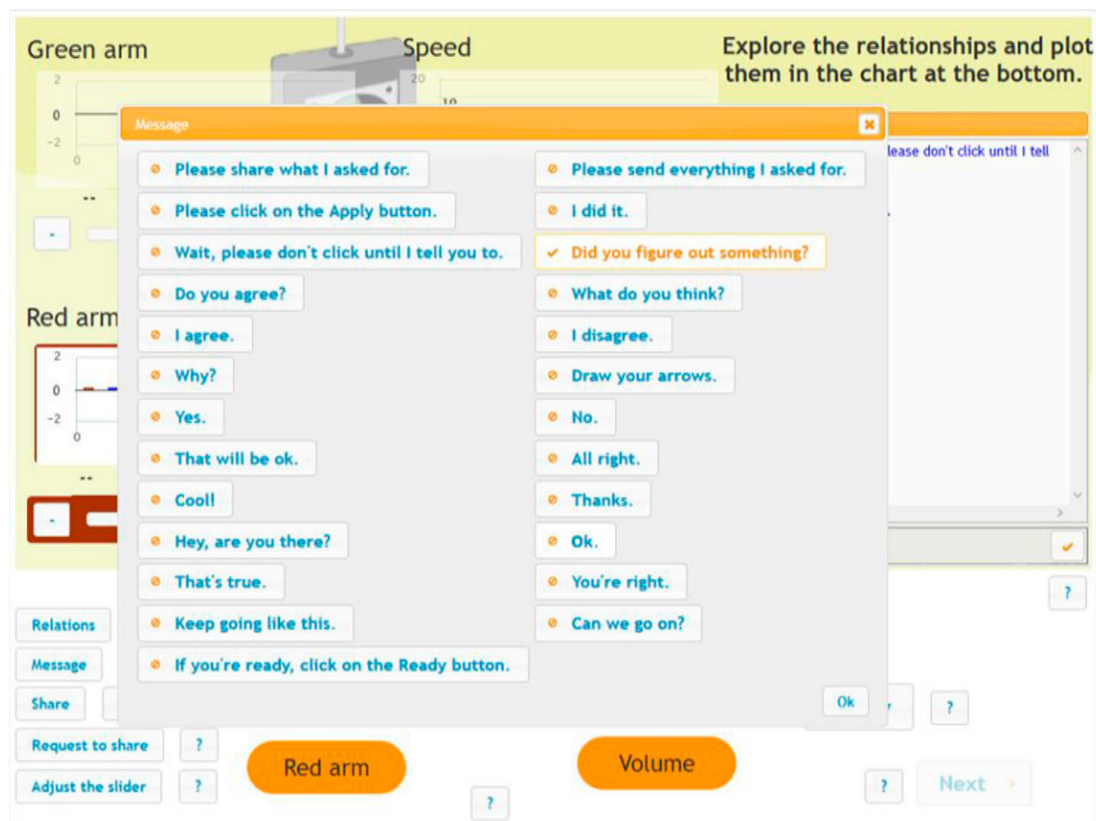


Figure 4. The pop-up window for the Message button in the knowledge acquisition phase. The tick indicates that Student 1 has selected a message to share, which will appear in the chat window after he presses OK (see Figure 6).

Figure 5. The pop-up window for the Relations button. The ticks indicate the elements already selected for sharing (see Figure 6).

Figure 6. Student 2's platform. After exchanging some pre-defined messages and a statement composed of pre-set elements with the Relations button in the chat window, Student 2 has requested information about the states of the Volume graph, the Red arm diagram and the Red arm slider and now views what Student 1 has shared as a reply.

list of pre-defined messages because of the numerous possibilities. This is why we established the option of allowing students to share an image of the present state of their elements. To do so, they click on the Mark button, then on the elements to be shared and finally on the Share button. If someone, for example, Student 1, has pressed Share (Figure 6), a message appears in the chat window informing Student 2 that Student 1 has shared information. Clicking on the names of the shared elements in the chat window, Student 2 can view the current state of the diagrams, graphs and sliders on Student 1's side (see Figure 6).

Students are not only able to share, but also to make a request to have information shared with them. When students click on the Mark button, then on the elements they want to be informed about and finally on the Request to share button, a message appears again in the chat window telling them that, for example, Student 2 has asked for information about particular elements (Figure 6).

The fifth button on the left, labelled "Adjust the slider", represents the third option for visual information exchange. If Student 2 wants to ask Student 1 to put the green arm slider in a particular position, she first needs to set it in the desired way on her own platform, and then press Mark, the slider and finally Move the slider. A message about Student 2's suggestion appears in the chat window (see Figure 6). Clicking on it, Student 1 sees the green arm slider moving to the spot where Student 2 had suggested it should be, so technically he can see an image of the state of Student 2's slider.

If a member discovers a relation between the variables and draws an arrow in the model, he/she has the chance to share it as well by clicking on the Mark button, then on the arrow and then on the Share button. A message again reports the information that has been shared. Clicking on it, the other member can see the arrows that have been sent appearing in his/her own model with a fairly distinct shade that is lighter than that of the arrows that he/she has already drawn.

If a member builds his/her model and considers it to be the final one, he/she needs to share it by pressing the Ready button. A message that says Final model appears in the chat window. Clicking on it, peers can review each other's models in the ways described above. The Next button only becomes active if both members have shared their respective versions of the final model. These models are not expected to be similar, however. This design aims to increase the chances of the answer actually being the student's and not merely a copy of his/her partner's model.

Activities to be evaluated in the knowledge acquisition phase

The final coding scheme for the H-H version can only be created after large-scale data collection in the fifth developmental step in our research agenda; however, many activities during the testing can already be assumed at this stage to be relevant indicators of different CPS skills. In Table 1, some activities are presented which are good candidates as indicators. The rest of them are linked to the use of the pre-defined messages. In addition, we name many behaviours that refer to the use of innovative options. The behaviours enumerated will be presented in line with our pre-conception of their relevance to the problem-solving and collaboration dimensions, including the specific CPS skills.

The elements of the cognitive dimension are given. The MicroDYN problems originally observe the knowledge acquisition and knowledge application part of the problem-solving process (Greiff & Funke, 2017). It is possible to retain this division for evaluation in the collaborative version. The problem-solving achievement in the knowledge acquisition phase, just like in the individual version, can be assessed with the discrete variable of the model (correct or not). Another option for evaluation can be a correct statement sent about the relations of the variables. The combinations of the relevant words can easily be pre-programmed.

The collaborative dimension is evaluated using the CRESST teamwork model, which, as noted above, contains the following skills: adaptability, coordination, decision making, interpersonal ability, leadership and communication. We have gathered inspiration for the assignment of the six

Table 1. Examples of activities to be extracted as indicators assigned to the different CPS skills in the knowledge acquisition phase.

	<i>Observed skill</i>	<i>Activity to be evaluated</i>
<i>Cognitive dimension</i>	Knowledge acquisition	- Providing the correct model with the <i>Ready</i> button - Sending a correct statement with the <i>Relations</i> button
<i>Social dimension</i>	Adaptability	- Sending the pre-defined message "Did you figure something out?"/"What do you think?"
	Coordination	- Sending the pre-defined message "If you're ready, click on the Ready button."/ "Draw your arrows." - Pressing the <i>Share</i> or <i>Next</i> button
	Decision making	- Sending the pre-defined message "Do you agree?"/"I agree."/"I disagree."/"Why?"/ "That will be ok." - Sending any statement (whether correct or not) with the <i>Relations</i> button - Sharing an arrow or arrows with the <i>Share</i> button - Sending the pre-defined message "Yes." or "No." after receiving the message "Do you agree?"
	Interpersonal	- Sending the pre-defined message "Cool!"/"Thanks."/"That's true."/"You're right."/"Keep going like this." - Pressing the <ul style="list-style-type: none"> • <i>Share</i> button after one's partner has made a request with the <i>Request to share</i> button • <i>Share</i> button after receiving the pre-defined message "Please share what I asked for." • <i>Apply</i> button after receiving the pre-defined message "Please click on the Apply button." • <i>Ready</i> button after receiving the pre-defined message "If you're ready, click on the Ready button." - Moving the slider to the spot requested with the <i>Adjust the slider</i> button - Drawing arrows after the pre-defined message "Draw your arrows." - Staying inactive after the message "Wait, please don't click until I tell you to." - Sending the pre-defined message "No" after receiving the pre-defined message "Please share what I asked for."/"Please click on the Apply button."/"If you're ready, click on the Ready button."/"Draw your arrows."/"Wait, please don't click until I tell you to." - Sending the pre-defined message "No." after one's partner has pressed the <i>Adjust the slider</i> button
	Leadership	- Sending the pre-defined message "Please share what I asked for."/"Please send everything I asked for."/"Please click on the Apply button."/"Wait, please don't click until I tell you to."/"Hey, are you there?"/"Can we go on?" - Pressing the <i>Request to share</i> or <i>Adjust the slider</i> button
	Communication	- Sending the pre-defined message "I did it."/"Yes."/"No."/"All right."/"Ok."

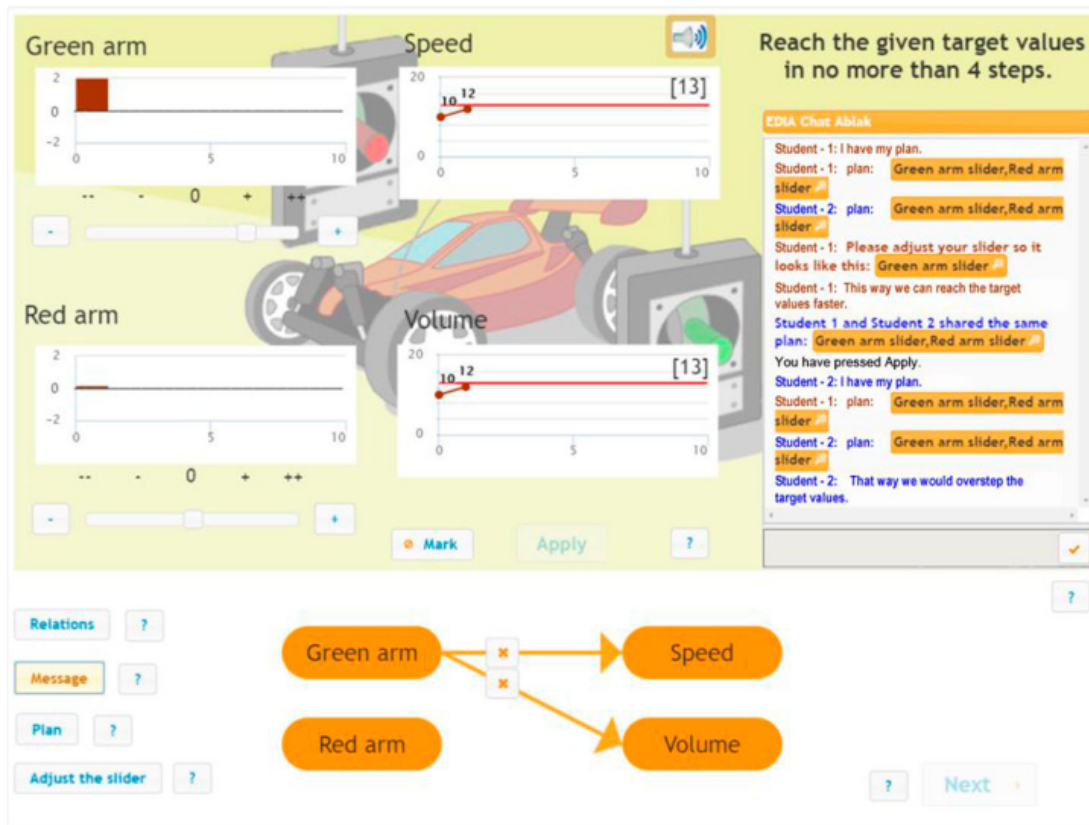
Note: The activities to be evaluated in reverse are indicated with a grey background.

skills with pre-defined messages and other activities from studies by Chung et al. (1999), Hsieh and O'Neil (2002), and O'Neil et al. (1997) (Table 1).

Restricted communication options in the knowledge application phase

The students are required to build a consensus in this phase: they need to agree on every step before implementing it. This condition also ensures that no member can solve a problem alone by rapidly pressing Apply four times without any discussion.

To enable dyads to agree on every step, we installed a new button for visual information exchange called Plan. If a student presses Plan, the current state of all his/her sliders is shared. A message informs his/her collaborator of this action in the chat window (Figure 7). If the other



The screenshot displays a learning interface for a car simulation. It features four sliders: 'Green arm', 'Red arm', 'Speed', and 'Volume'. The 'Speed' and 'Volume' sliders have numerical values of 10 and 12, and a target value of 13. A chat window on the right shows a conversation between two students discussing their plans. At the bottom, there is a diagram showing the relationships between the sliders: 'Green arm' and 'Red arm' both influence 'Speed', and 'Red arm' also influences 'Volume'. A 'Next' button is visible at the bottom right.

Figure 7. The knowledge application part. After sharing different plans, students have managed to arrive at a consensus and have already pressed Apply once. They have used the Adjust the slider and Message buttons for the discussion. (Currently, they are deciding about the second step, the Apply button being inactive.)

member clicks on it, he/she experiences his/her sliders moving to the particular spots his/her peer desires to use in the next step. In the case of all four steps, the Apply button only becomes active if the students share the same plan, i.e. if their sliders are in the same position.

The first plan shared at the start and in every further step is not shown immediately. The message which appears in the chat window only says that Student 1 or 2 has his/her plan (Figure 7). It can only be viewed after the other member has also shared his/her own plan. Thus, students cannot simply copy a plan that has been previously shared by their peers. They are forced to develop their own.

To come to a common plan, members can communicate via the Relations, Adjust the slider and Message buttons. The last one contains 23 similar messages in all the knowledge application phases (Figure 8). As there are no unavailable elements in this phase, the Share and Request to share buttons are eliminated.

Activities to be evaluated in the knowledge application phase

In Table 2 we review the activities which seem appropriate for indicator extraction in the knowledge application phase. In the cognitive dimension, it is possible again to observe the original variable of reaching the target value or not. However, the members of a dyad go through the process together throughout, which means this value cannot be differentiated, as the collaborators both succeed or fail. To be able to collect data on students' problem-solving skills on an individual level, we came up with the idea of having them think about their first plan in the four steps on their own. It is possible to have the system dynamically monitoring and selecting combinations of sliders whose application can lead toward the solution in the four steps. The evaluation of the social dimension is again based on pre-defined messages, use of other communication options and a combination of these.



Figure 8. The pre-defined message set under the Message button in the knowledge application phase.

Discussion and outlook

In this paper a research agenda was presented on possible solutions for dealing with the issues on realizing and evaluating interactions in CPS instruments. We stated that both the H–H and H–A assessment lines of measurement have their benefits for different assessment aims; furthermore, they also both have a great deal of potential for improvement. In an effort to boost the generalizability level of H–H instruments, a design was proposed in which the group composition changes in every task. The importance of retaining the pre-determined chat design of H–A assessment tools was pointed out to improve their ecological validity. The agenda recommended that each H–A instrument should have its own H–H pre-version to build on in order to create a realistic agent and to advance the evaluation system in H–A approaches.

Moreover, we discussed that the exchange of pre-defined messages is the method which provides an opportunity to understand the content of the interaction and ensures automated data coding at the same time. Nevertheless, it was strongly recommended that innovations be developed to compensate for the inflexibility of constrained communication. Creating the largest interaction space possible was another suggestion: CPS instruments should make their complete pre-defined message sets constantly available; furthermore, new options should be explored for constrained communication beyond pre-defined messages. The importance of taking into account the context of the communication was also stressed. Supporting this idea, a hybrid evaluation system was recommended in H–H measurement tools, in which both pre-defined messages and behavioural indicators play a great role.

To demonstrate an example of possible routes to realizing some of the proposals, we introduced a new CPS assessment tool. The instrument is not only the first example of transforming the so-called MicroDYN approach to make it suitable for H–H collaboration. It is also a pre-version of a future H–A assessment tool. The online measurement tool shows the potential for changing the group composition task by task. Furthermore, it offers several innovative solutions to replacing free chat with

Table 2. Examples of activities to be extracted as indicators assigned to the different CPS skills in the knowledge application phase.

	<i>Observed skill</i>	<i>Activity to be evaluated</i>
<i>Cognitive dimension</i>	Knowledge application	- Reaching the target value - Sharing one's own plan with the <i>Plan</i> button, which moves toward the solution
<i>Social dimension</i>	Adaptability	- Sending the pre-defined message "What do you think?"
	Coordination	- Sending the pre-defined message "Please share your plan."/"Our plans are not similar yet." - Pressing the <i>Plan</i> , <i>Apply</i> or <i>Next</i> button
	Decision making	- Sending the pre-defined message "This way we can reach the target values faster."/"That way we would overstep the target value."/"I think we should do that differently."/"Count it."/"Do you agree?"/"I agree."/"I disagree."/"Why?" - Sending any statement (whether correct or not) with the <i>Relations</i> button - Sending the pre-defined message "Yes" or "No" after the message "Do you agree?"
	Interpersonal	- Sending the pre-defined message "Cool!"/ "Thanks."/"That's true."/ "You're right."/"Keep going like this." - Pressing the <ul style="list-style-type: none"> • <i>Plan</i> button after receiving the pre-defined message "Please share your plan." • <i>Plan</i> button after receiving the pre-defined message "Our plans are not similar yet." - Moving the slider to the spot requested with the <i>Adjust the slider</i> button - Sending the pre-defined message "No." after receiving the pre-defined message "Please share your plan." - Sending the pre-defined message "No." after one's partner has pressed the <i>Adjust the slider</i> button
	Leadership	- Sending the pre-defined message "Hey, are you there?"/"Can we go on?" - Pressing the <i>Adjust the slider</i> button
	Communication	- Sending the pre-defined message "I did it."/"Yes."/"No."/"All right."/"Ok."

Note: The activities to be evaluated in reverse are indicated with a grey background.

besides the fixed, constantly available pre-defined message set. Some alternatives for future evaluation were also presented. The outlined evaluation model can again be considered as groundbreaking, as it unifies two methods for assessment: besides the pre-defined messages, a number of behavioural patterns have been identified which are relevant as an indicator of different CPS skills.

The next stage in developing the instrument is to test and improve it based on students' feedback. It will be key to collect their opinions to ascertain what pre-defined messages and other modifications are still required to make the environment user-friendly. Our aim is to reach the point where they report that they barely miss free communication within the test. Only if this goal is fulfilled can we consider scaling up the data collection and start concentrating on the coding scheme. With a stable and effective Human–Human version, we can make arrangements to embed the computer agent in the instrument. Certainly, we are at the beginning of a process of multiple years with a number of tasks ahead of us. Nevertheless, we believe the theoretical considerations and good practices in the instrument can serve as an inspiration even at this early stage of our research and contribute to overcoming the very complex challenge of developing effective CPS assessment tools.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study has been conducted with support from the National Research, Development and Innovation Fund of Hungary, financed under the OTKA K135727 funding scheme; Anita Pásztor-Kovács was supported by the “For the Young Talents of the Nation” (NTP-NFTÖ-19-B-0043) Scholarship from the Ministry of Human Capacities, Hungary.

Notes on contributors

Anita Pásztor-Kovács earned her master’s degree in Psychology in 2010 and her phd degree in Educational sciences in 2018 at the University of Szeged, Hungary. She is an assistant professor at the Institute of Education at the University of Szeged. In her researches she investigates the possibilities of assessing collaborative problem solving skills by technology-based methods. Her researches were supported by the New National Excellence Program of the Hungarian Ministry of Human Capacities through a Doctoral Candidate Research Scholarship in 2016 and a “For The Young Talents Of The Nation” Scholarship in 2017 and 2019.

Attila Pásztor received a master’s degree in Psychology in 2009 and a phd degree in Educational sciences in 2016 at the University of Szeged, Hungary. He is a research fellow at the MTA-SZTE Research Group on the Development of Competencies and an assistant professor at the Institute of Education at the University of Szeged. In his researches he focuses on the technology-based assessment and development of thinking skills such as inductive reasoning, combinatorial reasoning, creativity and problem solving. From 2017 he has been one of the senior professional developers in a large-scale project of the Educational Authority of Hungary referring to the establishment and improvement of technology-based educational assessment systems connected to public education.

Gyöngyvér Molnár is a full professor and the head of Institute of Education at the University of Szeged, Hungary. Her main areas of interest include: technology-based assessment, improving cognitive skills, studying the quality of school learning, and the potential for using ICT in education – all of which are aimed at improving the quality of learning. She heads eDia, an online diagnostic testing system used in numerous countries. In 2016, she won the Apáczai Csere János Prize – a Ministerial acknowledgement – for her outstanding scholarly work in support of educational practice.

ORCID

Anita Pásztor-Kovács  <http://orcid.org/0000-0002-5194-4437>

Attila Pásztor  <http://orcid.org/0000-0001-8441-446X>

Gyöngyvér Molnár  <http://orcid.org/0000-0003-4890-6904>

References

- Adams, R., Vista, A., Scoular, C., Awwal, N., Griffin, P., & Care, E. (2015). Automatic coding procedures for collaborative problem solving. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 115–132). Springer.
- 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>
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Martin, R., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Springer.
- Care, E., & Griffin, P. (2017). Assessment of collaborative problem solving processes. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 227–243). OECD Publishing.
- Care, E., Griffin, P., Scoular, C., Awwal, N., & Zoanetti, N. (2015). Collaborative problem solving tasks. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 85–104). Springer.
- 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. [https://doi.org/10.1016/S0747-5632\(99\)00032-1](https://doi.org/10.1016/S0747-5632(99)00032-1)
- Csapó, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). Technological issues of computer-based assessment of 21st century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). Springer.
- Csapó, B., & Funke, J. (Eds.). (2017). *The nature of problem solving: Using research to inspire 21st century learning*. OECD Publishing.
- Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology, 10*, 1522. <https://doi.org/10.3389/fpsyg.2019.01522>

- Dingler, C., von Davier, A. A., & Hao, J. (2017). Methodological challenges in measuring collaborative problem-solving skills over time. In *Team dynamics over time (Research on managing groups and teams)*. (Vol. 18, pp. 51–70). Emerald Publishing Limited. <https://doi.org/10.1108/S1534-085620160000018003>
- Dowell, N. M., Nixon, T. M., & Graesser, A. C. (2019). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior Research Methods*, 51(3), 1007–1041. <https://doi.org/10.3758/s13428-018-1102-z>
- Fiore, S. M., Graesser, A., & Greiff, S. (2018). Collaborative problem-solving education for the twenty-first-century workforce. *Nature Human Behaviour*, 2(6), 367–369. <https://doi.org/10.1038/s41562-018-0363-y>
- Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., Massey, C., O'Neil, H., Pellegrino, J., Rothman, R., Soulé, H., & von Davier, A. (2017). Collaborative problem solving: Considerations for the National Assessment of Educational Progress.
- Fiore, S. M., & Kapalo, K. A. (2017). Innovation in team interaction: New methods for assessing collaboration between brains and bodies using a multi-level framework. In A. A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 51–64). Springer.
- Fiore, S. M., & Wiltshire, T. J. (2016). Technology as teammate: Examining the role of external cognition in support of team cognitive processes. *Frontiers in Psychology*, 7, 1531. <https://doi.org/10.3389/fpsyg.2016.01531>
- Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *The Journal of Problem Solving*, 4(1), 19–42. <https://doi.org/10.7771/1932-6246.1118>
- Fischer, A., Greiff, S., & Funke, J. (2017). The history of complex problem solving. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 107–121). OECD Publishing.
- Fischer, A., Greiff, S., Wüstenberg, S., Fleischer, J., Buchwald, F., & Funke, J. (2015). Assessing analytic and interactive aspects of problem solving competency. *Learning and Individual Differences*, 39, 172–179. <https://doi.org/10.1016/j.lindif.2015.02.008>
- Fuks, H., Pimentel, M., & de Lucena, C. J. P. (2006). RU-Typing-2-Me? Evolving a chat tool to increase understanding in learning activities. *International Journal of Computer-Supported Collaborative Learning*, 1(1), 117–142. <https://doi.org/10.1007/s11412-006-6845-3>
- Graesser, A. C., Dowell, N., & Clewley, D. (2017). Assessing collaborative problem solving through conversational agents. In A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 65–80). Springer.
- Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018a). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, 19(2), 59–92. <https://doi.org/10.1177/1529100618808244>
- Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C., & Germany, M. L. (2018b). Challenges of assessing collaborative problem solving. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills* (pp. 55–73). Springer.
- Graesser, A. C., Forsyth, C. M., & Foltz, P. (2017). Assessing conversation quality, reasoning, and problem solving with computer agents. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 245–261). OECD Publishing.
- Greiff, S., & Funke, J. (2017). Interactive problem solving: Exploring the potential of minimal complex systems. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 93–105). OECD Publishing.
- Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement*, 36(3), 189–213. <https://doi.org/10.1177/0146621612439620>
- Griffin, P., & Care, E. (Eds.). (2015). *Assessment & teaching of 21st century skills. Methods and approach*. Springer.
- 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 A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 135–156). Springer.
- He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (2017). Collaborative problem solving measures in the programme for international student assessment (PISA). In A. A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 95–111). Springer.
- 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. <https://doi.org/10.1016/j.chb.2018.07.035>
- Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 37–56). Springer.
- Hsieh, I. L., & O'Neil, H. F. (2002). Types of feedback in a computer-based collaborative problem-solving group task. *Computers in Human Behavior*, 18(1), 699–715. [https://doi.org/10.1016/S0747-5632\(02\)00025-0](https://doi.org/10.1016/S0747-5632(02)00025-0)
- Krkovic, K., Pásztor-Kovács, A., Molnár, G., & Greiff, S. (2014). New technologies in psychological assessment: The example of computer-based collaborative problem solving assessment. *International Journal of e-Assessment*, 1(1).
- Krkovic, K., Wüstenberg, S., & Greiff, S. (2016). Assessing collaborative behavior in students. *European Journal of Psychological Assessment*, 32(1), 52–60. <https://doi.org/10.1027/1015-5759/a000329>

- Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013). *Handbook of latent semantic analysis*. Psychology Press.
- Liu, L., Hao, J., von Davier, A., Kyllonen, P., & Zapata-Rivera, D. (2016). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-life skill development* (pp. 344–359). IGI Global.
- Molnár, G., & Csapó, B. (2019). Making the psychological dimension of learning visible: Using technology-based assessment to monitor students' cognitive development. *Frontiers in Psychology, 10*, 1368. <https://doi.org/10.3389/fpsyg.2019.01368>
- OECD. (2017). *PISA 2015 results (volume V): Collaborative problem solving*.
- O'Neil, H. F., Chuang, S., & Chung, G. K. W. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *Assessment in Education, 10*(3), 361–373. <https://doi.org/10.1080/0969594032000148190>
- O'Neil, H. F., Chung, G. K. W. K., & Brown, R. S. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O'Neil (Ed.), *Workforce readiness: Competencies and assessment* (pp. 411–452). Lawrence Erlbaum.
- Pásztor-Kovács, A. (2018). *A kollaboratív problémamegoldó képesség mérése* [The assessment of collaborative problem solving skills]. [Doctoral dissertation]. Doctoral School of Education, University of Szeged.
- Pásztor-Kovács, A., Pásztor, A., & Molnár, G. (2018). Kollaboratív problémamegoldó képességet vizsgáló dinamikus teszt fejlesztése [Development of an online interactive instrument for assessing collaborative problem solving competence]. *Magyar Pedagógia, 118*(1), 73–102. <https://doi.org/10.17670/MPed.2018.1.73>
- Reilly, J. M., & Schneider, B. (2019). Predicting the quality of collaborative problem solving through linguistic analysis of discourse. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 149–157).
- Rosé, C. P., Howley, I., Wen, M., Yang, D., & Ferschke, O. (2017). Assessment of discussion in learning contexts. In A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 81–94). Springer.
- Rosen, Y. (2017). Assessing students in human-to-agent settings to inform collaborative problem-solving learning. *Journal of Educational Measurement, 54*(1), 36–53. <https://doi.org/10.1111/jedm.12131>
- Rosen, Y., & Foltz, P. (2014). Assessing collaborative problem solving through automated technologies. *Research and Practice in Technology Enhanced Learning, 9*(3), 389–410.
- Rosen, Y., & Mosharraf, M. (2016). Computer agent technologies in collaborative assessments. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-life skill development* (pp. 319–343). IGI Global.
- Rosen, Y., Wolf, I., & Stoeffler, K. (2020). Fostering collaborative problem solving skills in science: The Animalia project. *Computers in Human Behavior, 104*, 105922. <https://doi.org/10.1016/j.chb.2019.02.018>
- Scoular, C., & Care, E. (2020). Monitoring patterns of social and cognitive student behaviors in online collaborative problem solving assessments. *Computers in Human Behavior, 104*, 105874. <https://doi.org/10.1016/j.chb.2019.01.007>
- Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for operationalizing collaborative problem solving for automated assessment. *Journal of Educational Measurement, 54*(1), 12–35. <https://doi.org/10.1111/jedm.12130>
- 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. <https://doi.org/10.1016/j.compedu.2020.103964>
- 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. <https://doi.org/10.1016/j.chb.2019.05.033>
- 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. <https://doi.org/10.1016/j.compedu.2019.103672>
- Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving — More than reasoning? *Intelligence, 40*(1), 1–14. <https://doi.org/10.1016/j.intell.2011.11.003>
- Yuan, J., Liu, H., & Liu, Y. (2019). Assessment of collaborative problem solving based on process stream data: A new paradigm for extracting indicators and modeling dyad data. *Frontiers in Psychology, 10*, 369. <https://doi.org/10.3389/fpsyg.2019.00369>
- Zhang, J. (1998). A distributed representation approach to group problem solving. *Journal of the American Society for Information Science, 49*(9), 801–809. [https://doi.org/10.1002/\(SICI\)1097-4571\(199807\)49:9<801::AID-ASIS>3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1097-4571(199807)49:9<801::AID-ASIS>3.0.CO;2-Q)