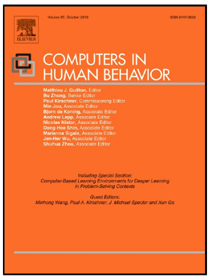


The Assessment of Collaborative Problem Solving in PISA 2015: Can Computer Agents Replace Humans?

Katharina Herborn, Matthias Stadler, Maida Mustafić, Samuel Greiff

PII:	S0747-5632(18)30357-1
DOI:	10.1016/j.chb.2018.07.035
Reference:	CHB 5624
To appear in:	*Computers in Human Behavior*
Received Date:	15 January 2018
Accepted Date:	23 July 2018

Please cite this article as: Katharina Herborn, Matthias Stadler, Maida Mustafić, Samuel Greiff, The Assessment of Collaborative Problem Solving in PISA 2015: Can Computer Agents Replace Humans?, *Computers in Human Behavior* (2018), doi: 10.1016/j.chb.2018.07.035

The Assessment of Collaborative Problem Solving in PISA 2015: Can Computer Agents Replace Humans?

Katharina Herborn[1], Matthias Stadler[1], Maida Mustafić[2], and Samuel Greiff[1]

University of Luxembourg[1]

University of Applied Sciences Northwestern Switzerland[2]

Author Note

This research was funded by the Organisation for Economic Co-operation and Development (OECD) and the Fonds National de la Recherche Luxembourg (ATTRACT „ASKI21") awarded to Samuel Greiff. We thank Nick Schweitzer and Tomas Kamarauskas for their contributions to the implementation of this study and for data collection.

Correspondence concerning this article should be addressed to Katharina Herborn, ECCS unit, University of Luxembourg, 11, Porte des Sciences, 4366 Esch, Luxembourg. Email: katharinamona.herborn@gmail.com

The Assessment of Collaborative Problem Solving in PISA 2015:

Can Computer Agents Replace Humans?

Abstract

Despite the relevance of collaborative problem solving (CPS), there are limited empirical results on the assessment of CPS. In 2015, the large-scale Programme for International Student Assessment (PISA) first assessed CPS with virtual tasks requiring participants to collaborate with computer-simulated agents (human-to-agent; H-A). The approach created dynamic CPS situations while standardizing assessment conditions across participating countries. However, H-A approaches are sometimes regarded as poor substitutes for natural collaboration, and only a few studies have identified if the collaborations with agents capture real dynamics of human interactions. To address this, we validated the original PISA 2015 CPS assessment by investigating the effects of replacing computer agents with real students in classroom tests (human-to-human; H-H). We obtained the original PISA 2015 CPS tasks from the OECD and replaced agents with real students to provide more real-life collaboration environments with less control over conversations; the H-H was less constrained than the H-A but still limited by predefined sets of possible answers from which the humans' would make selections. The interface remained nearly identical to the original PISA 2015 CPS assessment. Students were told the types of collaboration partners, namely humans versus agents. We applied structural equation modeling and multivariate analyses of variance to a sample of 386 students to identify the dimensionality of the CPS construct and compared the effects in CPS performance accuracy and number of behavioral actions. Results indicated no significant differences between type of collaboration partner. However, students performed a larger number of actions when collaborating with a human agent.

Keywords: collaborative problem solving, PISA 2015, assessment, validation, agent technologies

The Assessment of Collaborative Problem Solving in PISA 2015:

Can Computer Agents Replace Humans?

The majority of 21st century jobs require people to solve problems in collaboration with others as innovation is usually an outcome of interconnected individuals who share and integrate their expertise (OECD, 2017). In such collaborations, people are required to adapt to different groups and collaboration partners with varying proficiencies, skills and personalities in a range of problem-solving environments and group goals, using a variety of communication channels. It is not surprising that the greatest increase in the demand for employees' skills in the final decades of the 20th century occurred for those who had nonroutine analytical skills (i.e., the skills involved in problem solving; OECD, 2017) and social skills, including collaboration skills (Autor, Levy, & Murnane, 2003).

In order to ensure that current students and future career entrants will sufficiently meet the demands of the 21st century workplace, an increasing number of educational and governmental initiatives have added assessments of students' skills in solving problems in collaboration with others. For example, such initiatives include the Assessment and Teaching of 21st Century Learning (ATC21S; Griffin & Care, 2015) and the 2015 Programme for International Student Assessment (PISA). These programs have aimed to evaluate students' current CPS proficiencies in order to potentially make sensible changes to educational systems around the globe (e.g., Breakspear, 2012). In the current study, our aim was to validate one of the measures used to assess students' CPS skills in the PISA 2015 assessment.

1. Introduction

1.1. Theoretical Background

In psychology, the construct of Collaborative Problem Solving (CPS) is defined as solving problems in collaboration with others. CPS is a conjoint construct that is comprised of the two components of problem solving on the one hand and social collaboration on the other. We assume that problem solving accounts for the cognitive component, involving the ability to transform a current problem state into a desired goal state (Mayer & Wittrock, 2006), whereas social collaboration accounts for the skill that allows a person to interact in synchrony with other participants (Griffin & Care, 2015). Their combination defines the interdependent skill of problem solving to move toward a common goal with other people (Fiore et al., 2010; Griffin, 2014). More specifically, PISA 2015 defined CPS as follows, after carefully assembling existing CPS definitions (for more information, see OECD, 2013):

"Collaborative problem solving competency is 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, 2017).

The construct of CPS can be applied to various real-world settings, for example, the sharing of knowledge with colleagues in the workplace to overcome the individual boundaries of expertise. It can go beyond the workplace to the planning of tasks we complete with our families and friends in our private lives. Especially in our increasingly globalized and hyperconnected world, through digitalization we apply CPS in diverse aspects of 21st century life. Due the increasing significance of CPS, educational and political initiatives, including PISA 2015 and ATC21S, are assessing CPS to ensure that students demonstrate proficiency in CPS skills at the end of compulsory education. Other programs at the national levels are currently discussing the integration of CPS (e.g., in the US National Assessment of Educational Progress; NAEP; Fiore et al., 2017), so the relevance of CPS assessments in education is expected to remain high in the foreseeable future.

However, even though the construct of CPS is receiving increasing educational attention, there is a general debate on the ideal methodology for the assessment of CPS due to a lack of empirical evidence in academic research (von Davier & Halpin, 2013). Traditional assessments, such as situational judgment tests in which individuals react to hypothetical role-relevant scenarios (Mumford, Campion, & Morgeson, 2006; Patterson, Zibarras, & Ashworth, 2016) or paper-based questionnaires (e.g., Aguado, Rico, & Sánchez-Manzanares, 2014; Wang, MacCann, Zhuang, Liu, & Roberts, 2009), however, are increasingly being replaced by computer-based assessment approaches, especially in large-scale settings such as applied in ATC21S (Griffin & Care, 2015) and PISA 2015 (OECD, 2017). These can simulate complex and dynamic CPS situations in virtual tasks, similar to the situations people face in real life.

To assess the collaboration aspect of CPS, virtual CPS tasks require participants to collaborate with either computer-simulated agents (the human-to-agent technology: H-A) or real humans (human-to-human technology: H-H). Both approaches have advantages and disadvantages in the assessment of CPS (e.g., O'Neil, Chuang, & Chung, 2003). H-A approaches, as applied in PISA 2015, can offer standardized assessment conditions, which are especially crucial for student comparisons on the individual level. However, such conditions are often criticized for being limited in the extent to which they can allow natural collaboration to unfold because they limit conversational interactions between team partners (Graesser, Kuo, & Liao, 2017). In comparison, H-H approaches, such as applied in ATC21S, assess CPS during collaborations between humans and therefore provide better representations of natural collaboration. However, they lack controllability, which was crucial for the PISA 2015 CPS assessment, which aimed to compare students' CPS skills across countries. Also, H-H logfiles with natural speech information are very complex to analyze (Liu et al., 2015) and would take too long to be implemented in large-scale assessments (Care, Scoular, & Griffin, 2016).

At this point in the science, a very limited body of research has explored differences between H-A and H-H assessment approaches (e.g., Rosen, 2015; Rosen & Tager, 2013). In academic research, this concern has been addressed in theoretical reviews (Graesser, Kuo, & Liao, 2017) but not empirically. Despite the advantages of the H-A approach, which offers assessment standardization and controllability of effects, the extent to which the H-A methodology in the PISA 2015 CPS assessment was able to capture the real dynamics of H-H interactions given the a priori constraints of the H-A approach has yet to be determined. Therefore, the extent to which the CPS skills assessed in PISA 2015 represent the way students would interact with human partners needs to be identified. The validity of the PISA 2015 H-A approach is of critical interest, considering the large impact of the PISA 2015 CPS results across the globe on educational systems and policies as well as the research opportunities it provides for academic research.

We conducted the current study to validate the PISA 2015 CPS assessment by investigating the effects of replacing computer agents with real students in classroom tests (human-to-human; H-H). For this purpose, we obtained the otherwise confidential PISA 2015 CPS tasks so that we could generate additional data for this study and extend the main PISA 2015 trial. The interface in the H-H tasks remained nearly identical to the original PISA 2015 CPS assessment. We adopted the predefined chat design from the original PISA H-A tasks, so the H-H condition was constrained. More specifically, students selected from a predefined set of possible answers with one agent being replaced with a real student in the tasks. Therefore, there were always two humans interacting by selecting from a fixed set of chat options. This H-H condition was indeed constrained by these chat options, but less constrained than the H-A condition. Students were also informed about which types of partners they were collaborating with (computer-agents or computer-agents and a real classmate) in order to emphasize a likely effect of the collaboration partners' nature on the main test takers. We identified the dimensionality of the underlying CPS construct and compared the agent effects (H-H versus H-A) on CPS performance accuracy and behavioral actions. If no substantial differences in collaborative activities occur, and if no variation in CPS scores is identified between the H-A and H-H formats, such outcomes would support the validity of the PISA 2015 CPS assessment as an authentic representation of collaborative behavior by the computer agents.

1.2. The PISA 2015 CPS Assessment

PISA 2015 assessed CPS by employing different computer-based tasks that required active social collaboration with simulated agents during the solving of real-life problem scenarios in digital tasks (OECD, 2013, 2017). The students' social collaboration was based on selecting predefined messages from lists of possible messages and sending them through a chat window in which the computer agent and students exchanged information in order to solve the required problem in the task space. Actions were performed in the action space of the task in order to solve problems that required non-chat actions, such clicking, dragging and dropping, or moving the elements on the screen in the task. Each correctly selected message or action that was chosen reflected a specific CPS skill for which students received credit (1 point; in a few cases, 2 points) or no credit (0) otherwise.

Overall, 12 distinguishable CPS skills were assessed in the PISA 2015 tasks (the PISA 12-cell matrix; OECD, 2013), and each CPS skill was conceptualized on the basis of four individual problem solving processes, i.e., (A) exploring and understanding, (B) representing and formulating, (C) planning and executing, and (D) monitoring and reflecting, which was crossed with three newly conceptualized social collaboration dimensions, i.e., (1) establishing and maintaining a shared understanding, (2) taking appropriate action to solve the problem, and (3) establishing and maintaining team organization. Table 1 displays the PISA 12-cell matrix that represents the CPS framework (OECD, 2013).

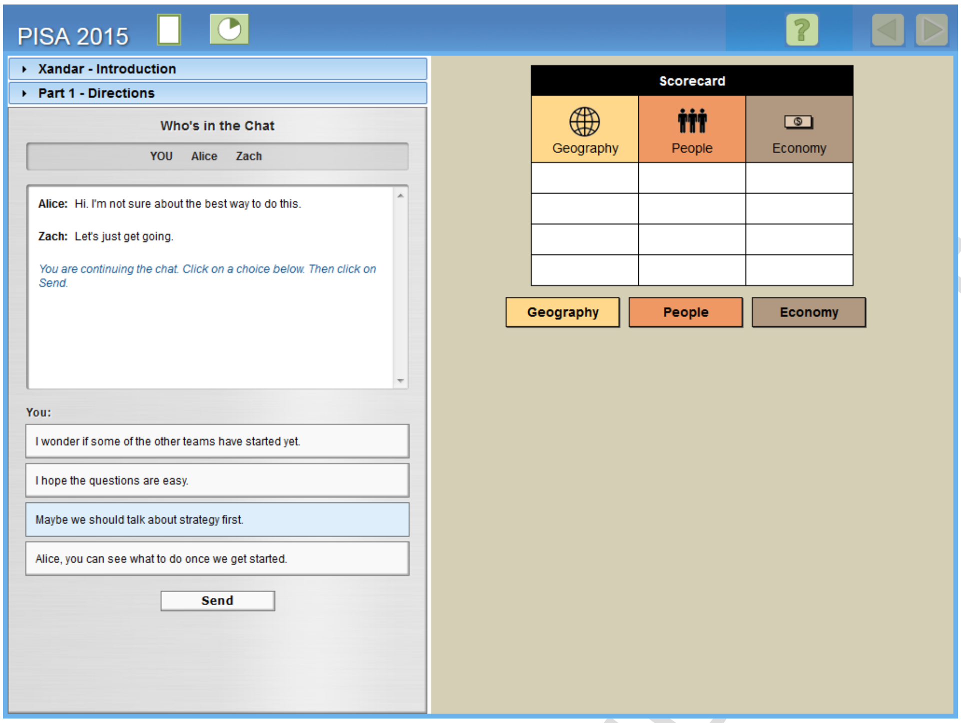
Table 1

The 12-Cell Matrix Illustrating the 12 CPS Skills in the PISA 2015 Assessment.

	(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organisation
(A) Exploring and Understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve problem
(B) Representing and Formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organisation (communication protocol/rules of engagement)
(C) Planning and Executing	(C1) Communicating with team members about the actions to be/ being performed	(C2) Enacting plans	(C3) Following rules of engagement, (e.g., prompting other team members to perform their tasks.)
(D) Monitoring and Reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organisation and roles

Note. Drawn from the OECD CPS Draft Report in PISA 2015 (2013).

To provide an example, Figure 1 illustrates the original PISA 2015 CPS task called “Xandar,” which was assessed in the main PISA 2015 assessment (OECD, 2017). In Xandar, students were required to compete in a contest along with their collaboration partners Alice and Zach in which they had to answer questions about the geography, people, and economy of the fictional country called Xandar (OECD, 2017). In general, Xandar assessed students’ decision-making, coordination, and consensus-building collaboration skills through correctly selected predefined messages and actions (OECD, 2017). More specifically, in the first part of Xandar as illustrated in Figure 1, students were required to communicate with team members about the actions to be/being performed in this specific problem scenario (CPS skill “C1” in the PISA 12-cell matrix; OECD, 2013). Therefore, the third message “Maybe we should talk about strategy first” was scored as the correct message.



Alice: Hi. I'm not sure about the best way to do this.

Zach: Let's just get going.

You are continuing the chat. Click on a choice below. Then click on Send.

I wonder if some of the other teams have started yet.

I hope the questions are easy.

Maybe we should talk about strategy first.

Alice, you can see what to do once we get started.

Figure 1. Illustration of the PISA 2015 CPS task Xandar as retrieved from the official OECD report (OECD, 2017). Screenshot 1 illustrates a typical predefined message selection scenario for communicating with the computer agents. Message 3 is scored as the correct message representing the CPS Skill C1: communicate with team members about the actions to be/being performed in this specific problem scenario.

1.3. H-A versus H-H Assessments

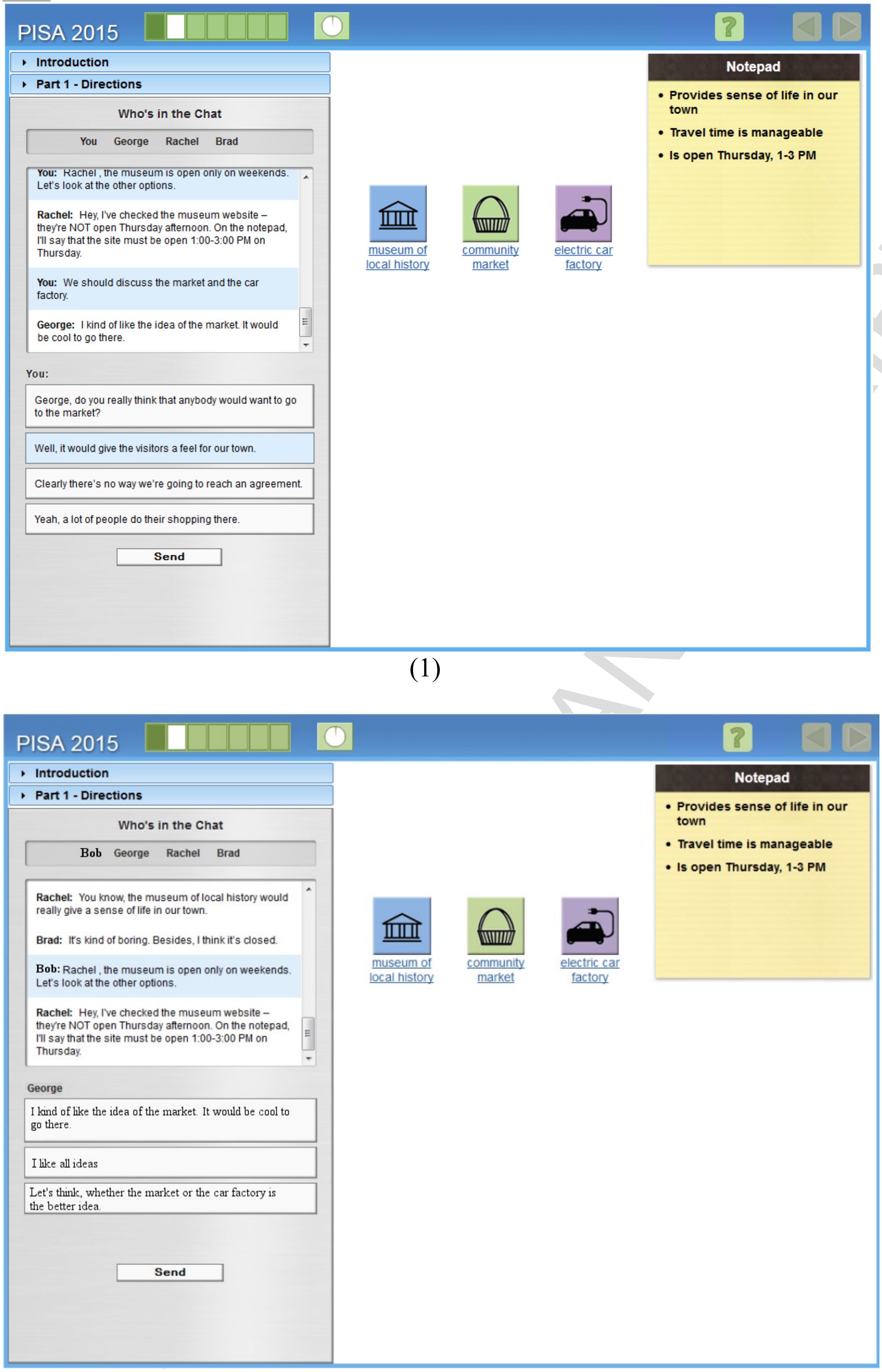
PISA 2015 aimed to compare differences in educational systems, evaluate their impact on students' CPS proficiencies (OECD, 2017), and eventually draw implications for current educational policies. For this, PISA 2015 applied the H-A method to create dynamic CPS situations while standardizing assessment conditions across participating countries in a controlled manner. Computer agents can be validly used as conversation partners in students' collaborative learning (e.g., Biswas et al., 2010) and CPS assessment (e.g., Rosen, 2015) and possibly offer the advantage of assessing a wider spectrum of CPS skills (Rosen, 2015). In addition, the H-A technology allows CPS assessments to control for external effects, such as group composition effects (e.g., Wildman et al., 2012), personality effects (Herborn, Mustafic, & Greiff, 2018), or the partner's CPS activity and proficiency. In addition to controlling for these external effects, PISA 2015 could assess range of CPS skills and abilities with agents in CPS scenarios that were both varied and standardized across the students who participated in the PISA 2015 CPS assessment. In turn, this enabled the comparison of PISA 2015 CPS results between cultures and languages to a great extent.

The H-A method enabled researchers to standardize the agents' responses regardless of the messages selected by the students or the CPS level. Each scenario had a fixed sequence of assessment episodes that all students received; each assessment episode had the same starting point and converged on the same end point after interactions between the student and agent. Students with strong CPS skills who selected a response that offered zero points early in the test could still score well overall due to later standardized conditions in the task; and likewise, students with low CPS skills who chose a response that offered points early in the test would still be identified as low performers on the basis of their subsequent (poor) responses.

However, assessing CPS with standardized H-A technologies also has some key drawbacks. As applied in the ATC21S project (Griffin & Care, 2015), H-H technologies assess CPS during collaboration between humans. Therefore, H-H assessment approaches provide more natural human collaboration situations that are closer to the kinds of CPS situations students encounter in real life. Unexpected responses or actions, which might not be assessed by standardized algorithms, would therefore not be captured in H-A approaches (Rosen, 2015; Graesser, Kuo, & Liao, 2017). For example, Graesser, Kuo, and Liao (2017) stated that conversational interactions can be limited when comparing them with free chat sessions between collaboration partners. Students were allowed to type individual messages during their collaborations in ATC21S. Therefore, communication and actions that were not coded in a standardized algorithm could be captured, and logstream data enabled researchers to conduct both qualitative and quantitative analyses (Care, Scoular, & Griffin, 2016). This raises the question of the extent to which the H-A methodology used in the PISA 2015 CPS assessment was able to capture the real dynamics of H-H interactions given the a priori constraints of the H-A approach.

1.4. The Present Study

We conducted this study to investigate whether the original PISA 2015 CPS tasks were able to reflect the extent to which students' collaborations with computer agents represented the way students would interact with human partners. In other words, our long-term goal was to determine whether agents can replace humans as collaboration partners in CPS assessments. This study does not fully achieve this long-term goal but does take an initial step in addressing the issue. In particular, some of the original PISA 2015 CPS tasks were reformatted and redesigned into a constrained H-H format by replacing one of the agents with a classmate in each task to allow real human interaction to take place. One of the computer agents was replaced by a classmate, a peer of equal status to the student. It is important to note that the computer-agents replaced by classmates were not in the role of the experts, but rather, the role within the group was defined by the students' CPS skills preforming the computer-agent. The predefined chat communication was adopted and extended in the new H-H tasks. More specifically, the original PISA 2015 H-A approach, as illustrated in Figure 1, was fully adopted, and only the type of collaboration partners was changed (computer-agents or computer-agents and a real classmate). Students in the role of the collaboration partners also received predefined messages to choose from. Figure 2 illustrates a reformatted PISA task in H-H format and provides an example chat turn of an assessment episode.



You: Rachel , the museum is open only on weekends. Let's look at the other options.

Rachel: Hey, I've checked the museum website – they're NOT open Thursday afternoon. On the notepad, I'll say that the site must be open 1:00-3:00 PM on Thursday.

You: We should discuss the market and the car factory.

George: I kind of like the idea of the market. It would be cool to go there.

You:

George, do you really think that anybody would want to go to the market?

Well, it would give the visitors a feel for our town.

Clearly there's no way we're going to reach an agreement.

Yeah, a lot of people do their shopping there.

- Provides sense of life in our town
- Travel time is manageable
- Is open Thursday, 1-3 PM

(1)

Rachel: You know, the museum of local history would really give a sense of life in our town.

Brad: It's kind of boring. Besides, I think it's closed.

Bob: Rachel , the museum is open only on weekends. Let's look at the other options.

Rachel: Hey, I've checked the museum website – they're NOT open Thursday afternoon. On the notepad, I'll say that the site must be open 1:00-3:00 PM on Thursday.

George

I kind of like the idea of the market. It would be cool to go there.

I like all ideas

Let's think, whether the market or the car factory is the better idea.

- Provides sense of life in our town
- Travel time is manageable
- Is open Thursday, 1-3 PM

(2)

Figure 2. Illustration of the PISA 2015 CPS task the Visit as retrieved from the official OECD report on released field trial cognitive items (OECD, 2017). Screenshot 1 illustrates a typical predefined message selection scenario for communicating with the computer agents. The computer-agent George replies with the standardized message "I kind of like the idea of

the market. It would be cool to go there". Screenshot 2 illustrates the chat turn in H-H format. The student replacing George also received predefined messages to choose from.

Using the example of Figure 2, students of equal status to the main test taker were in the role of the collaboration partners and replaced George in the task "the Visit". These students acted as George within the group, and also received predefined messages to select from and to reply in the group chat. Among the predefined messages is George's original message "I kind of like the idea of the market. It would be cool to go there" that the agent George sent to the chat in the H-A format (Figure 1). Based on the CPS proficiency levels as published in the PISA 2015 CPS report (OECD, 2013), George's original message was rated as medium collaboration proficiency. In addition, the two further messages "I like all ideas" (low collaboration proficiency) and "Let's think, whether the market or the car factory is the better idea" (high collaboration proficiency) were also offered to the students replacing George, so that they also had three messages to select from.

In a first step, this study investigated the factorial validity of both approaches in assessing CPS using several consecutive confirmatory factor analyses. The reformatting allowed stipulating the following research questions for this study.

Research Question 1: Are there differences in factorial validity when assessing students' CPS performance using computer agents versus classmates?

In a second step, this study further compared the validity by investigating into the effects in CPS performance accuracy and behavioral actions between type of format (H-A versus H-H) by looking at the accuracy as well as the number of students' interactions with the problem in each task individually. We examined the differences in the correctness scores and number of actions made by students assessed using only computer agents with that of students assessed using a classmate.

Research Question 2: Are there differences in CPS performance accuracy and behavioral actions when assessing students' CPS performance using computer agents versus classmates?

2.0 Method

2.1 Sample

The Luxembourgish National Commission for Data Protection (CNPD) and the Educational Ministries of Rhineland and Hesse approved the quantitative data collection, which was conducted between January 2016 and April 2016. Schools were recruited over email and received a donation of 160 Euro for each class. A total of *N* = 748 students in 35 classes in Grades 9 and 10 (PISA population) from eight secondary grammar schools in Germany voluntarily participated in the quantitative studies (for more information about the different school tracks in Germany, see Paulick, Watermann & Nückles, 2012). Two trained test administrators assessed students during regular class time on a single day per class (approximately 4.5 hr). After data management procedures (e.g., exclusion of students without informed consent forms, the exclusion of *N* = 71 students in pilot studies, data that were missing by design, and the exclusion of two students due to missing values), the final sample included 386 students (*M*age = 15.69, *SD* = 0.64, 59.3% identified as female[1]). We ensured that these students had not participated in the official PISA 2015 CPS assessment as such students would therefore have already been familiar with the PISA 2015 CPS tasks. Of note, we have confidence in the quality of the current sample with regards to CPS performance per gender. One of the key implications in the report is that girls significantly outperform boys in every participating country and economy, such as in Germany (OECD,

[1] The overrepresentation of girls occurred for 10 girls-only classes.

2017). Also in our data, which represents German students, girls achieved higher CPS performance scores, and significantly outperformed boys in three out of the four CPS tasks.

2.2 Procedure

Each student separately picked a number when entering the classroom to be randomly assigned to one of the Macbook laptops that had been set up. Each laptop belonged to one particular assessment group out of four (Groups 1 to 4). Each group completed the PISA 2015 CPS tasks in an H-A and/or an H-H format in a particular sequence in the classroom test sessions. For example, Group 1 completed the CPS Tasks 1 and 2 in the H-A format and subsequently completed the CPS Items 3 and 4 in the H-H format. Table 2 illustrates the sequence of CPS tasks in the H-A design or the H-H design per group. Students were allocated randomly in an effort to minimize systematic differences between and within groups.

Table 2
The Allocation of Students into Groups and the Specific Order of H-A and H-H PISA 2015 CPS Task Completion

		PISA 2015 CPS tasks			
		Task 1	Task 2	Task 3	Task 4
Tasks in specific format per group	Group 1 (*N* = 97)	H-A	H-A	H-H	H-H
	Group 2 (*N* = 97)	H-H	H-H	H-A	H-A
	Group 3 (*N* = 98)	H-A	H-A	H-A	H-A
	Group 4 (*N* = 94)	H-H	H-H	H-H	H-H

2.3 Measures

PISA CPS. Each student individually completed four original PISA 2015 CPS tasks out of the seven PISA 2015 CPS tasks that we obtained from the OECD. Unfortunately, not all PISA tasks were assessed due to the given time constraints in the classroom test sessions. Each task consisted of several parts, and each task required the students to solve problems in collaboration with a minimum of one and a maximum of three simulated computer agents (as illustrated in Figures 1 and 2). It is important to mention that students were informed when they were collaborating with computer agents by the sentence “You are now collaborating with agents” (H-A design) in order to emphasize a likely effect of the collaboration partners’ nature on the main test takers. The collaboration with agents was based on predefined chat messaging in a chat window on the left side of the screen in which students were able to re-view the chat history at any time. Problem solving was performed, for example, by using drag and drop, cut and paste, and clicking on an action space on the right side of the screen. When students chose the message that was identified as the correct one or they performed the specific actions as required, they received full credit (1 point; in a few cases, 2 points) or no credit (0 points). While the students solved the CPS tasks, the number of actions they performed were automatically logged. The number of actions included students’ clicks and double clicks, keystrokes, as well as “drag and drop” actions such as moving elements of the task when performing the task. The sum of all actions reflected each student’s overall number of actions score. For further information on the PISA 2015 CPS tasks, see the official OECD report on the PISA 2015 CPS approach and results (OECD, 2013, 2017).

PISA CPS in constrained H-H. The PISA 2015 CPS tasks were reformatted into an H-H design that enabled the creation of a more human real-life collaboration environment. For this, one classmate, who was a peer of equal status to the main test-taker, replaced one of the agents per CPS task and was in the role of the collaboration partner throughout the tasks. It is important to mention that the sentence “You are now collaborating with agents and a real classmate” was displayed on the screen to inform participants about the types of collaboration partners. In the PISA CPS H-H tasks, classmates who were in the role of collaboration partners experienced the same task design as the main test taker and were also presented with predefined messages to choose from when responding in the group chat. Among the predefined messages were the original messages that the agent sends into the group chat in the original H-A format. In addition to that, two further messages were offered, so that students replacing the agent had three messages to select from and reply with one of them in the group chat. Hereby, the three messages represented low, medium and high collaboration, in order to offer messages to the students that reflect different CPS proficiency levels. Apart from the mode of interaction (collaboration with agents or with agents and a classmate), the interface remained identical between the H-H and the H-A designs as illustrated in Figure 2. It is important to note that the existing software code underlying the PISA H-A CPS tasks was entirely reused in the H-H format (the code was provided by the Educational Testing Service; ETS).

2.4 Statistical Analysis

We implemented a structural equation modeling approach (SEM; Bollen, 1989) in MPlus Version 7.0 (Muthén & Muthén, 2012). We chose the maximum likelihood estimator (ML) for our models. The model fit was evaluated according to standard fit indices consisting of the Comparative Fit Index (CFI; cut-off for good fit: CFI > .95), Tucker Lewis Index (TLI; cut-off for good fit: TLI > .95), Root Mean Square Error of Approximation (RMSEA; cut-off for good fit: RMSEA < .05), and Standardized Root Mean Square Residual (SRMR; cut-off for good fit: SRMR < .05; Hu and Bentler, 1999). Descriptive statistics were calculated in SPSS version 22.

2.5 Missing Data

Due to the study design of four groups completing the tasks in a specific order of H-A and/or H-H formats as outlined in Table 2, each student was missing some data by design. To account for these missing data, we performed multiple imputations to replace each missing value in the data set with m pseudo-random values, thereby creating m "complete" data sets (Longford, 2005). As our data were missing by design, we could consider them to be missing completely at random (von Davier, 2013), but given the relatively large amount of missing data, we decided to run 50 imputations (Bodner, 2008). Only the main test takers were included in the analyses for RQ1 and RQ2, and students in the role of the collaboration partners were excluded.

3.0 Results

3.1 Validity differences when assessing students' CPS performance using computer agents or classmates (RQ1).

Research Question 1 was aimed at investigating differences in the factorial validity of assessing CPS performance using computer agents or classmates. In order to empirically identify this potential interface effect, we thus defined three consecutive latent factor models.

Model A defined CPS as a one-dimensional model that represented CPS as a general factor comprised of H-A and H-H items. In other words, the one-dimensional model assumed no difference in type of format. Given that the H-A and H-H items shared the same content, correlated errors between the respective items were allowed. This one-dimensional CPS model showed a very good fit (see Table 3 for the specific fit indices of all models). Notably, the H-A items (*Mdn* $\lambda = .63$) had slightly weaker loadings on the common factor than the H-H items (*Mdn* $\lambda = .65$), $\chi^2(1) = 6.81$, $p = .009$. For a graphical illustration of Model A, see Figure 3.

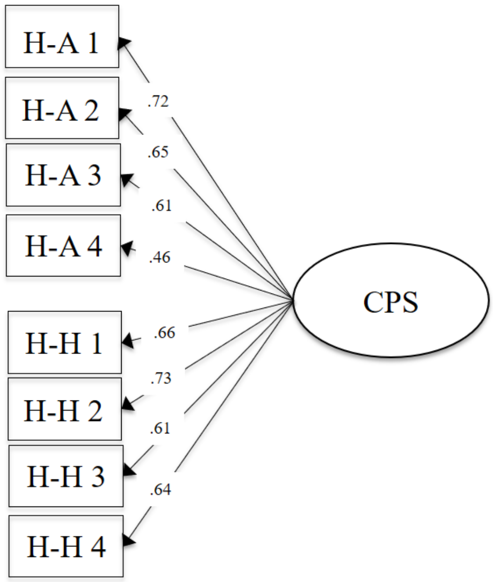


Figure 3. One-dimensional model (Model A) representing CPS as a common factor comprised of the original PISA 2015 CPS items in the H-A and H-H formats. The factor loadings of the eight H-A and H-H items ranged from .46 to .73.

Table 3
Goodness of Fit Indices for the One-Dimensional Model (no Difference between H-A and H-H) and the Two-Dimensional Model (Treating the H-A and H-H Collaborations as Separate Dimensions)

PISA 2015 internal structure	χ^2	*df*	*p*	CFI	TLI	RMSEA
Model A (one-dimensional)	21.55	21	.436	.98	.98	.008
Model B (two-dimensional)	21.82	19	.293	.924	.888	.020
Model C1 (bifactor H-A)	1321.42	16	<.001	.53	.15	.450
Model C2 (bifactor H-H)	1281.29	16	<.001	.60	.19	.432

Note. χ^2 and *df* were estimated with ML; *df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error Approximation.

To contrast against Model A, we modeled CPS as a two-dimensional model (Model B) that represented the H-A and H-H performances as two separate factors. In this model, the components that were specific to the H-A (computer-agent) and H-H (human-agent) collaborations could not be captured by a general CPS factor. This two-dimensional CPS model also showed an acceptable model fit (Table 3) with all items loading substantially on their respective factors. However, the correlation between the two factors did not deviate significantly from 1, $\chi^2(1) = 1.24$, $p = .49$, thereby indicating that the model in which H-A and H-H were separated into two factors should be rejected. Model B is illustrated in Figure 4.

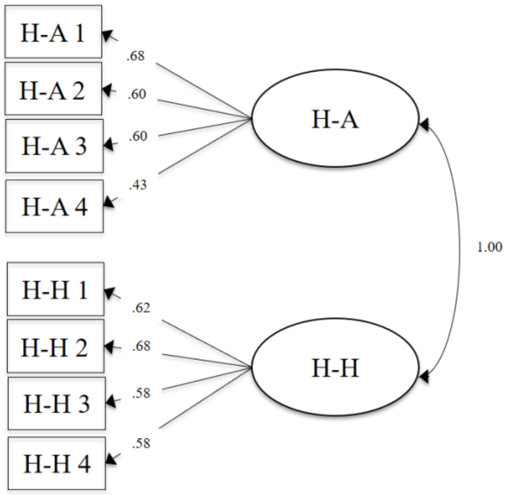


Figure 4. Two-dimensional model (Model B) in which CPS was represented as two separate factors on which the original PISA 2015 CPS items in the H-A and H-H formats loaded, respectively. The factor loadings of the four H-A items ranged from .43 to .68, and for the H-H items, from .58 to .68.

Finally, we defined two bifactor models in which the variance that was common to the H-A and H-H items was separated from potential interface-specific components. The advantage of this model over the correlated factor model (Model B) is that it allowed us to partition the variance, and thus we could separate what was common to the two factors and what was specific to a particular factor. In other words, the bifactor model enabled us to assess the extent of how much the H-A and H-H items had in common and whether there would be any specific information contained in the H-A or H-H items that could not be explained by a general CPS factor. The model included a common CPS reference factor for all H-A and H-H items, and either a specific H-A factor (see Figure 5 for Model C) or H-H factor. Both models had poor fit, and thus, we were able to reject the idea that there were separate interface-specific variance components (see Table 3).

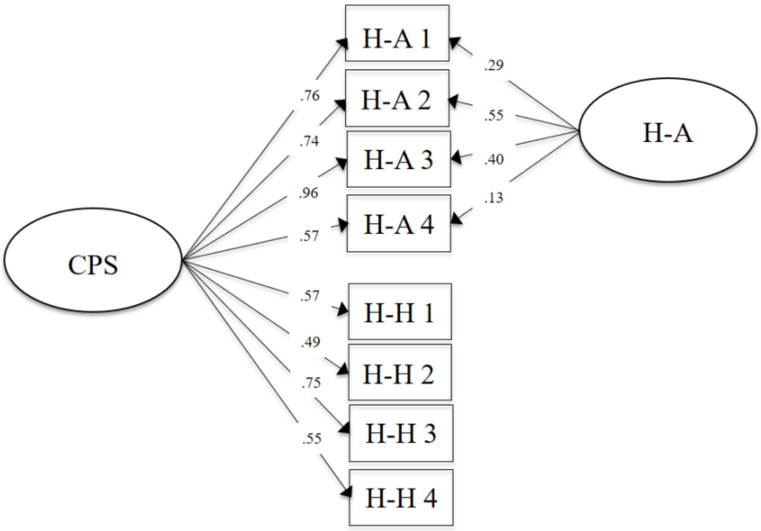


Figure 5. Bifactor model representing CPS as a general factor and a specific H-A factor (Model C).

3.2 Are there differences in CPS performance accuracy and behavioral actions when students' CPS performance is assessed using computer agents versus classmates (RQ2)?

Research Question 2 was aimed at investigating differences in students' CPS performance accuracy and behavioral actions when CPS was assessed using computer agents or classmates. Table 4 provides descriptive statistics for students' CPS performance accuracy and behavioral actions using computer-agents versus classmates as collaboration partners. First, we compared the CPS performance of students assessed using only computer agents with that of students assessed using a classmate by conducting multivariate analyses of variance (MANOVAs) with the four CPS tasks in H-A and H-H format as dependent variables. Using Pillai's trace, there was no significant effect of interface on students' CPS performance, V = 0.02, $F(4, 186) = 0.77$, $p = .548$. These results did not suggest that any performance differences were caused by the different interfaces for any of the four tasks and thus rendered any follow-up analyses redundant.

Table 4
Descriptive Statistics for Students' CPS Performance Accuracy and Behavioral Actions *in the H-A versus H-H assessment.*

		Performance Accuracy		*Behavioral Actions*	
	Format	*M*	*SD*	*M*	*SD*
Task 1	H-A	16.44	3.78	79.02	14.70
	H-H	15.84	3.91	85.74	24.54
Task 2	H-A	24.84	6.23	83.82	14.10
	H-H	24.16	5.46	92.24	26.06
Task 3	H-A	24.27	3.66	82.25	17.76
	H-H	23.45	4.58	84.85	29.74
Task 4	H-A	8.21	1.46	39.51	7.68
	H-H	7.91	1.85	42.00	12.37

Regarding behavioral actions during the assessment, we compared the number of actions performed by students assessed using only computer agents with that of students assessed using a classmate. Again, MANOVAs were computed with the four tasks as dependent variables. Using Pillai's trace, there was a significant effect of interface on the number of actions students performed, V = 0.11, *F*(4, 186) = 5.55, *p* < .001. Follow-up analyses on the individual tasks revealed that students collaborating with a classmate interacted slightly more frequently than students collaborating with only computer agents on Task 1, *F*(1, 189) = 7.99, *p* < .001, $\eta_p^2 = .04$, and Task 2, *F*(1, 189) = 10.36, *p* < .001, $\eta_p^2 = .05$, but not on Task 3, *F*(1, 189) = 0.03, *p* = .869, $\eta_p^2 = .00$, or Task 4, *F*(1, 189) = 2.02, *p* = .157, $\eta_p^2 = .01$.

4.0 Discussion

4.1 Implications

The aim of the present study was to identify the extent to which the original PISA 2015 CPS H-A methodology and standardized assessment is valid for capturing the real dynamics of H-H interactions given the a priori constraints of the H-A approach (RQ1). In other words, the long-term goal of this study was to assess whether students' CPS results in the original PISA 2015 CPS assessment would be found to resemble students' real CPS skills exhibited in interactions with humans. Further, this study further compared the validity by investigating into the effects in CPS performance accuracy and number of behavioral actions between type of format (H-A versus H-H) by looking at the success as well as the students' interactions with the problem in each task and format individually. Especially, we explored the difference in the number of behavioral actions made by students assessed using only computer agents with that of students assessed using a classmate (H-A versus H-H; RQ2). Considering the general lack of empirical information on H-A and H-H comparisons for validation purposes, especially from the original PISA 2015 CPS assessment, coupled with the expected impact of the PISA 2015 CPS results, this study aimed to generate empirical results that could validate the approach.

To do so, we obtained the original PISA 2015 CPS tasks from the OECD and reformatted the original PISA 2015 CPS tasks into a H-H format, in which one computer-simulated agent was replaced by a real classmate, and created more real-life collaboration environments between humans with less control over the conversation. The H-H condition was a constrained human-to-human collaboration as the predefined chat communication from which the humans' would make selections was adopted and extended, and free chat response not enabled. Therefore, this study does not fully achieve this long-term goal but does take an initial step in addressing the issue. Likewise to the H-A format, each H-H scenario had a fixed sequence of assessment episodes that all students received; each assessment episode had the same starting point and converged on the same end point after interactions between the student and agent. After assessing both formats in students, this study investigated the factorial validity of both approaches for assessing CPS using several consecutive confirmatory factor analyses. For this, we identified a potential interface effect in the collaboration with computer-agents or classmates, and we thus defined three consecutive latent factor models.

For RQ1, the one-dimensional model identified CPS as a general factor in both types of formats (H-A versus H-H). Second, the two-dimensional model (Model B: Figure 4) identified CPS as two separate H-A and H-H formats. Finally, two different bifactor models allowed for a general CPS factor plus a specific method factor for the H-A (Model C1) and H-H tasks (Model C2). Overall, the models supported the general CPS factor in both types of formats and did not support the separation into two factors or the necessity of an additional method factor. Therefore, this study offers support for the use of computer agents as collaboration partners as implemented in the standardized H-A approach and discussed in the body of literature on the use of computer agents in CPS assessments (Rosen, 2015; Graesser & McDaniel, 2008; Millis et al., 2011). However, it still needs to be considered that the H-H condition in this study was constraint and did not allow free response collaboration when drawing this implication.

For RQ2, we investigated the differences in students' correctness scores and number of actions made by students assessed using only computer agents with that of students assessed using a classmate by applying multivariate analyses of variance (MANOVAs). First, we compared CPS performance accuracy and correctness scores of students assessed using only computer agents with that of students assessed using one real classmate in addition to the agents. The results did not suggest any performance accuracy differences. These findings in which we identified no significant difference in CPS performance between type of format have been found before in other academic studies (e.g., Rosen & Tager, 2013). Regarding the number of behavioral actions during the assessment, we compared the number of behavioral actions (i.e., clicking, dragging and dropping, or moving elements of the tasks) implemented by students assessed using only computer agents with those of students assessed using a classmate in addition to the agents. The results showed that students collaborating with classmates interacted slightly more frequently during the tasks than students collaborating with only the computer agents did. Differences in behavioral actions during the collaboration with computer agents or human agents have been identified in previous studies. For example, Rosen and Tager (2015) found no significant differences in time-on-task between type of format (H-H versus H-A) using predefined messaging; however, students still spent more time on CPS tasks in the H-H format. Because these small differences in behavioral actions did not affect actual performance, however, they do not seem to limit the comparability of H-A tasks and H-H tasks.

4.2 Limitations

We conducted this study to investigate whether the original PISA 2015 CPS tasks were able to reflect the extent to which students' collaborations with computer agents represented the way students would interact with human partners. However, the main limitations of the results should be stated in order to show the constraints of the generalizability of the results. The main study limitation in the design of this study is the constraint of the H-H condition. First, it is important to remember that only one agent was replaced by a human in the PISA 2015 CPS tasks in the constraint H-H format. The creation of more real-life collaboration environments between humans in the H-H tasks and the allowance of external effects that occur in real H-H interactions were therefore very limited. If more agents had been replaced, the collaborative behavioral action effects most likely would have been larger. Second, the H-H condition was not fully free human-to-human collaboration as the predefined chat communication was adopted from the original PISA 2015 CPS assessment and extended, and free chat response prohibited. The adoption of the predefined chat communication allowed the comparison of students' CPS performance in the

H-A and H-H condition. In the PISA CPS H-H tasks, classmates who were in the role of collaboration partners experienced the same task design, however their predefined messages were extended in order to act as collaboration partners. This did not allow for a direct comparison between the performance in the H-A tasks and H-H tasks of students who were in the role of the collaboration partners. Therefore, the generalizability of the results of the nature of the H-A approach in resembling students' real CPS skills exhibited in interactions with humans is limited.

However, replacing one agent with a classmate already allowed for more natural communication and external effects (e.g., group composition or the collaboration partner’s CPS proficiency) on the main test-taker’s performance. Therefore, this study has first identified the extent to which the H-A methodology used in the PISA 2015 CPS assessment was able to capture the real dynamics of H-H interactions given the a priori constraints of the H-A approach. In order to overcome these limitations of the study, future research design should enable to allow the comparing of the H-A approach with results obtained in H-H settings that involve only human agents and allow open-response chat communication, such as in ATC21S. As such a design would require considerable changes to be made to the tasks’ interface, it however might seriously limit the extent to which such tasks would be comparable to the original tasks. In addition to that, this study included the assessment of factorial validity in the PISA 2015 CPS assessment. Future research should integrate the assessment of validity from different perspectives in order to evaluate the validity and generalizability of the PISA 2015 CPS assessment and general use of computer-agents as collaboration partners. Lastly, data collection was conducted only in Germany and Luxembourg in the current study, thus accounting for why conclusions cannot be drawn about cultural implications and explaining why information about what these results imply for other cultures is therefore limited. However, ETS designed the PISA 2015 CPS tasks to be culturally independent to the greatest possible extent, such as tested in their Field Trials. Therefore, we believe that the results and implications are theoretically generalizable to other cultures.

4.3 Conclusion

Clearly, computer agents will not replace actual humans in collaborations anytime soon, however they are increasingly integrated in educational settings and of critical interest, such as shown by this study, and workplace environments. For the sake of assessing students' CPS skills in a standardized way in PISA 2015, the H-A approach seems to be comparable to the H-H approach. Given the many advantages of the H-A approach regarding standardization, application, and interpretation, these results justify the further development and use of H-A assessment instruments in educational large-scale studies.

References

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

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics, 118*(4), 1279-1333.

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.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling, 15*(4), 651-675.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance*. OECD Education Working Papers, No. 71, OECD Publishing, Paris.

Care, E., Scoular, C., & Griffin, P. (2016) Assessment of Collaborative Problem Solving in education environments. *Applied Measurement in Education, 29*(4), 250-264, DOI: 10.1080/08957347.2016.120920

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors, 52*(2), 203-224.

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. Retrieved from: https://nces.ed.gov/nationsreportcard/pdf/researchcenter/collaborative_problem_solving.pdf

Graesser, A.C., & McDaniel, B. (2008). *Conversational agents can provide formative assessment, constructive learning, and adaptive instruction*. In C. A. Dwyer (Ed.), The future of assessment: Shaping teaching and learning (pp. 85-112). New York, NY: Routledge.

Graesser, A., Kuo B.-C., & Liao, C-H. (2017). Complex Problem Solving in assessments of Collaborative Problem Solving. *Journal of Intelligence, 5*(10), 1-14. doi:10.3390/jintelligence5020010

Griffin, P., McGaw, B., & Care, E. (Eds.) (2012). *Assessment and teaching of 21st century skills*. Dordrecht: Springer.

Griffin, P. (2014). Assessment and teaching of C21 Skills (ATC21S). *Measuring collaborative skills: Challenges and opportunities.* Melbourne, Australia: University of Melbourne.

Griffin, P., & Care, E. (2015). *The ATC21S method.* In P. Griffin & E. Care (Eds.), Assessment and teaching of 21st century skills (pp.3-33). Dordrecht, The Netherlands: Springer.

Herborn, K., Mustafic, M., & Greiff, S. (2018). *Computer-based Collaborative Problem Solving in PISA 2015 and the role of the Big Five*. Manuscript submitted for publication.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Liu, L.; Von Davier, A.; Hao, J.; Kyllonen, P.; Zapata-Rivera, D. *A tough nut to crack:*

Measuring collaborative problem solving. In Handbook of Research on Technology Tools for Real-World Skill Development; Yigal, R., Ferrara, S., Mosharraf, M., Eds.; IGI Global: Hershey, PA, USA, 2015; pp. 344–359.

Longford, N. (2005). *Missing data and small-area estimation: Modern analytical equipment for the survey statistician.* New York: Springer.

Mayer, R. E., & Wittrock, M. C. (2006). *Problem solving*. In P. A. Alexander & P. H. Winne (Eds.), Handbook of educational psychology (pp. 287-303). Mahwah, NJ: Erlbaum.

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-Verlag.

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2006). *Situational judgment in work teams: A team role typology*. In J. A. Weekley & R. E. Ployhart (Eds.), Situational judgment tests: Theory, measurement, and application (pp. 319-343). Mahwah, NJ: Lawrence Erlbaum.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide (7th edition).* Los Angeles, CA: Muthén & Muthén.

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: Principles, Policy & Practice, 10*, 361–374.

Organisation for Economic Co-operation and Development (OECD) (2013). *PISA 2015 draft collaborative problem solving framework*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD). (2017). *PISA 2015 Results (Volume V): Collaborative problem solving*. Paris: OECD Publishing.

Patterson, F., Zibarras, L. and Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100, *Medical Teacher, 38*(1), pp. 3-17.

Paulick, I., Watermann, R., & Nückles, M. (2013). Achievement goals and school achievement: The transition to different school tracks in secondary schools. *Contemporary Educational Psychology*, *38*(1), 75-86.

Rosen, Y., & Tager, M. (2013). *Computer-based assessment of collaborative problem-solving skills: Human-to-agent versus human-to-human approach*. Boston, MA: Pearson Education.

Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of H-A approach. *International Journal of Artificial Intelligence in Education, 25*(3), 380-406. doi:10.1007/s40593-015-0042-3

von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 175–201). Boca Raton: Chapman Hall/CRC.

von Davier, A. A., & Halpin, P. F. (2013). Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations. *ETS Research Report Series, 2013*(2), i-36.

Wang, L., MacCann, C., Zhuang, X., Liu, O. A., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multimethod approach. *Canadian Journal of School Psychology, 24*(2), 108-124. DOI: 10.1177/0829573509335470.

Wildman, J. L., Shuffler, M., Lazzara, E. H., Fiore, S., Burke, C. S., Salas, E., & Garven, S. (2012). Trust development in swift starting action teams: A multilevel framework. *Group & Organization Management, 37*, 138-170.

Highlights

- We validated the original PISA 2015 Collaborative Problem Solving tasks.
- We found no significant differences per type of collaboration partner (agents or classmates).
- Students performed a larger number of actions when collaborating with classmates.