Ghost in the Machine: A Symposium on Collaboration Between Human and Computerized Agents in Educational Contexts

Matthias Stadler (chair), Ludwig-Maximilians-Universität München, matthias.stadler@lmu.de
Frank Fischer (chair), Ludwig-Maximilians-Universität München, frank.fischer@lmu.de
Pantelis M. Papadopoulos, Aarhus University, pmpapad@tdm.au.dk
Anika Radkowitsch, Ludwig-Maximilians-Universität München, anika.radkowitsch@psy.lmu.de
Ralf Schmidmaier, Ludwig-Maximilians-Universität, Ralf.Schmidmaier@med.uni-muenchen.de
Martin Fischer, Ludwig-Maximilians-Universität, martin.fischer@med.uni-muenchen.de
Haiying Li, Rutgers University, haiying.li@gse.rutgers.edu
Jiangang Hao, Educational Testing Service, jhao@ets.org
Samuel Greiff, University of Luxembourg, Samuel.Greiff@uni.lu
Arthur C. Graesser (discussant), The University of Memphis, art.graesser@gmail.com

Abstract: The importance of collaboration is growing in an era where knowledge-based tasks are increasingly accomplished by teams of people with complementary roles and expertise, as opposed to individuals doing isolated work. Moreover, the nature of collaboration is shifting to a more sophisticated skillset that includes accomplishing tasks through mediated interactions with peers halfway across the world or even computer-generated agents. The 4 papers presented in this symposium focus on this new opportunity of introducing computer-generated agents in collaboration as well as the challenges entailed. The presentations will cover a broad range of topics ranging from the assessment of collaborative problem-solving skills to the use of computer-generated agents in intelligent tutoring systems. In addition, they will illustrate the validity and practicability of various state-of-the-art approaches to introduce non-human collaborators into human collaboration.

Overall focus of the symposium

Collaboration between humans and computerized agents, a scenario seemingly taken right out of science fiction just a few decades ago, has become part of our everyday life. Whether learning to play chess using online tutoring systems, playing PC games with non-player characters, or using intelligent personal assistants for work, most people interact and collaborate with computerized agents more frequently than they may be aware of. Research on 21st century collaboration therefore needs to investigate the possibilities and challenges that come with human-agent collaboration. The aim of this symposium will therefore be to present research projects that focus on the collaboration between humans and computerized agents as well as discuss the implementation, benefits, challenges and potential pitfalls entailed.

Historically, there have been two approaches to human-agent interaction (Terveen, 1994). The first approach assumes that the way to get computers to collaborate with humans is to endow them with human-like abilities, to enable them to act like humans. The second approach assumes that the way to get computers to collaborate with humans is to exploit their unique abilities, to complement humans, beginning from the premise that computers and humans have fundamentally asymmetric abilities. These two approaches are not completely distinct of course as will be obvious from our symposium. Both of them require adequate models of human ability and the computational modeling underlying artificial intelligence. The study of human-agent interaction is therefore inherently interdisciplinary and necessitates to be approached from various perspectives at once. The proposed symposium recognizes this requirement in combining research in psychology, educational sciences, and computer science.

Specific contributions of the presentations

In total, the symposium will include 4 presentations covering several broad areas within the field of human-agent collaboration: Automated tutoring, language processing, computer supported learning, and the use of computerized-agents in the assessment of collaborative problem-solving skills. Obviously, these areas are neither representative in and by themselves, nor are they comprehensively and fully covered within this limited selection of papers. They do however represent a good overview of the current state-of the-art on a long journey to our understanding of collaboration in the 21st century. We will begin by discussing whether human-agent collaboration can actually be compared to human-human collaboration in assessments of collaborative problem-solving in educational contexts such as PISA 2015 (OECD, 2017). This is followed by two presentations that employ conversational agents to assess and improve collaborative skills and learning. One study discusses how automated text analysis is used to predict participants' success in collaboratively solving problems together with computerized-agents. The other study demonstrates how conversational pedagogical agents can potentially be used in large collaborative learning activities to improve the quality of peer dialog. Finally, we present a learning environment using human-agent collaboration to teach collaborative diagnostic competencies to medical students. All four presentations will be discussed by Art Graesser, an internationally renowned expert in computerized assessment, intelligent tutoring and human-computer interaction.

The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans?

Matthias Stadler and Samuel Greiff

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. 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). 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. 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). 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 complex to analyze and would take long to be implemented in large-scale assessments.

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. 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 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. Students of equal status to the main test taker were in the role of the collaboration partners and replaced one of the simulated agents in each task. 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 (Fig. 1). Based on the CPS proficiency levels as published in the PISA 2015 CPS report, 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?

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

Results and discussion

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 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 and H-H tasks. 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 (e.g., Rosen, 2015). 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.

Language and group performance on science inquiry in simulated synchronous collaboration

Haiying Li, Jiangang Hao, and Art Graesser

Collaborative problem solving (CPS) involves a high level of social and cognitive skills, which is critically important for career and life success (OECD, 2017). Even though CPS is in high demand in workplace and life, collaboration is not explicitly taught or assessed in schools. Instead, it is acquired through group work in core academic subjects, such as science or extracurricular activities. Recently, researchers have developed computer-assisted, simulated environments to provide platforms for students to augment CPS skills for supporting science learning (Hao et al., 2017; Lin et al., 2013). Computer simulations with online text chats serve as group cognitive tools to facilitate mutual understanding of the problem and quest for solution through group discussion (Gijlers et al., 2009). Researchers found that simulated environments effectively enhanced learning performance during the CPS (Hao et al., 2015), boosted high-level cognitive skills (Lin et al., 2014), and augmented active engagement (Chang et al., 2017). Previous studies on group discourse analyses concentrated on the utterances in sequential order based on thread analyses that were related to specific sub-tasks during problem solving, such as simulation run (Chang et al., 2017). To date, no studies have examined whether language used by groups is correlated with group performance, which is the focus of the present study. The study on language use in CPS is fundamentally important and significant because language is an essential means of communication among members of a group in CPS to share and negotiate ideas, regulate and coordinate behaviors, and sustain the interpersonal exchanges to solve a shared problem (Liu et al., 2016). In this study, we aimed to answer two research questions: (1) Does language used by groups predict group performance in a simulated CPS task? and (2) What language used by groups can augment group performance in the CPS task?

Method

956 Participants were recruited through Amazon Mechanical Turk, a crowdsourcing data collection platform to evaluate online learning environments (Li & Graesser, 2017). These participants were randomly paired into 478 dyadic groups. Two participants in each group synchronously collaborated to interact through text chats with two virtual agents to complete a set of science inquiry practices on volcanoes which includes data collection and data interpretation. In this study, we only analyzed the team interactions between dyads. Interactions were used to measure CPS competency, including sharing ideas, negotiating ideas, regulating problem-solving activities, and maintaining communication (Liu et al., 2016). Language used during group interaction was measured by 18 language features at the multiple textual levels (see Table 1), including descriptive (e.g., the number of turns and words in group interactions, syllables in a word, words in a sentence), word information (e.g., pronouns, word frequency, age of acquisition), syntactic complexity (e.g., the number of modifiers per noun phrase, sentence

syntactic similarity), referential cohesion (e.g., noun overlap), and situation model that represent deep cohesion and clear causality to fill the information gap during communication (e.g., causal verb, intentional verb, expanded temporal connectives). We used these features to measure language because these features enabled us to predict the quality of students' scientific explanations in the form of constructed responses during science inquiry within an intelligent tutoring system (Li et al., 2018). All these features were exacted through Coh-Metrix, an automated text analysis tool (Li et al., 2018). Performance of each group on inquiry competency was evaluated by the total scores of responses to seven multiple choice items and four constructed responses. Multiple choice responses were randomly chosen from one team member, whereas four constructed responses were submitted by one randomly chosen team member.

Results, discussion, and conclusions

Results indicated significant correlations between group performance and all language features except the number of words during interaction, with correlations ranging from small to medium (see Table 1). Results of a multiple linear regression with stepwise 10-fold cross-validation showed a significant regression model, with 9 language features explaining 21.91% of the total variance in group performance on science inquiry practices. Table 1 displays significant predictors as well as the constant and coefficients. Specifically, more turns between group members and longer sentences in each turn positively predicted group inquiry performance, whereas the number of words in group interaction negatively predicted group performance. These findings imply that more high quality of communication elicits explicit and detailed information, which ultimately facilities group performance. The minimal use of pronouns but more use of first and third plural pronouns in high quality of group inquiry practices implies that the use of more formal language to deliver information but the use of "*we*" or "*us*" to emphasize their embeddedness into social relationships did enhance group performance. The more use of rare words that people never or rarely encounter and of spoken words that are acquired in later than earlier ages denotes that the use of more formal or academic words is related to high quality group performance. Scores of group performance increased with the less use of intentional verbs, which signal actions that are enacted by team members, motivated by plans in pursuit of solutions for the problems. It is surprising that language features such as referential cohesion and situation model were not significant predictors. The possible explanation is that team members shared contexts, so no much information gaps exist during their interaction.

These findings confirmed the use of more formal, academic language in inquiry correlated with higher performance on inquiry practices, such as longer utterances, more turns, more use of third person plural nouns, more less frequently-used words, and words acquired at later ages (Li et al., 2018). However, findings also demonstrated the unique characteristics of group interaction in CPS, such as the more use of first personal plural pronouns to signify team identity during collaboration as well as the less use of action verbs or plan words to decrease the demand or request from group members so as to demonstrate collaboration. Based on findings such as these, teachers or virtual agents will be able to identify and address the specific language used inappropriately in CPS that may cause social or cognitive issues during collaboration. In the future, real-time, automated scaffolds could be designed and provided to students to enhance their CPS competence.

Table 1: The correlations between group performance and language features

Textual Levels	Language Features	Mean	SD	Pearson *r*	Coefficients
Descriptive	Number of turns	79.68	34.93	-0.09*	0.011
	Number of words	391.72	24.46	-0.02	-0.003
	Number of words per sentence	4.55	1.18	0.35**	0.638
	Number of syllabus per word	1.25	0.05	0.25**	-----------
Word Information	Pronoun	129.12	25.79	-0.13**	-0.010
	1st person singular pronoun	64.13	19.11	-0.11*	-----------
	1st person plural pronoun	16.39	9.47	0.10*	0.025
	2nd person pronoun	15.64	9.50	-0.11*	-----------
	3rd person singular pronoun	4.02	4.78	0.01	-----------
	3rd person plural pronoun	5.55	4.55	0.05	0.031
	CELEX word frequency for content words	2.52	0.11	-0.13**	-1.827
	Age of acquisition for content words	330.88	23.66	0.15**	0.006
Syntactic Complexity	Number of modifiers per noun phrase, mean	0.47	0.11	0.21**	-----------
	Sentence syntax similarity, adjacent sentences	0.16	0.05	-0.12**	-----------
Referential Cohesion	Noun overlap, all sentences	0.02	0.01	0.14**	-----------
Situation Model	Causal verb	63.99	14.24	-0.12*	-----------
	Intentional verb	40.52	13.08	-0.17**	-0.019

Textual Levels	Language Features	Mean	SD	Pearson *r*	Coefficients
	Expanded temporal connectives	13.25	6.74	0.10*	-----------

Note. * $p < .05$. ** $p < .01$. The constant was 12.339 in regression model. --- refers to insignificant predictors.

Conversational agents for collaborative learning activities in the lab and "in the wild"

Pantelis Papadopoulos

Peer dialogue, written or verbal, as the externalization and sharing of one's thoughts and understandings with peers, is a key element in collaborative learning in order to reach a common understanding. Despite the unequivocally crucial role of peer dialogue, empirical evidence suggested that not all dialogue-based activities ensure effective collaborative behavior among students (Vogel et al., 2016). Therefore, additional questions emerge regarding the nature of an effective peer dialogue. Towards this direction, the Academically Productive Talk (also known as APT or Accountable Talk) framework is based on the analysis of what could be deemed as effective classroom discussion for the purpose of promoting academic learning and reasoned student participation (Michaels et al., 2008; Resnick et al., 2010). According to APT (Resnick et al., 2010), students' discussion should be accountable to the learning community (students should build upon their peers' understandings), accurate knowledge (use explicit evidence and refer to shared knowledge to validate contributions), and rigorous thinking (connect logically their claims). APT suggests a set of different types of strategic interventions (talk moves, e.g., "Do you agree or disagree with what your partner said about …? Why?") a teacher could use in the classroom to trigger valuable discourse (adding-on, elaborate on agreement/disagreement, re-voicing, pressing for accuracy, building on prior knowledge, pressing for reasoning, expanding on reasoning). One crucial characteristic of APT (and one that perhaps invites the use of conversational agents into the learning design) is that it focuses on students' reasoning over correctness, allowing also the teacher to hand over the control of the discussion to the students. This presentation will discuss the use, potential, and practicability of conversational pedagogical agents (i.e., software tools that can use text/voice to interact with the user through natural language) in collaborative learning activities in the context of the MentorChat tool and the colMOOC project.

Studies with MentorChat

By utilizing the APT framework, conversational agents can model effective teacher-student interactions and scaffold both one-to-one and group discussions (e.g., Dyke et al., 2013; Stahl, 2015). This is also the case with the MentorChat tool (e.g., Tegos et al., 2016, 2017) that can employ both directed (available to one student) or undirected (available to both students) interventions in pair, chat-based, discussions. Technically, MentorChat is based on three elements: the domain model of the subject matter, the intervention model, and the peer interaction. The teacher (domain expert) is responsible for configuring the first two, first by creating a concept map with concepts and relationships, and second by deciding on the discussion patterns that will trigger agent interventions. During peer discussion, the system monitor students' utterances in the chat room and builds respective domain models that represent the knowledge of students and pairs. These models are continuously compared to the domain model created by the teacher and agent interventions occur when a defined pattern is recognized (e.g., the agent recognizes that the discussion revolves around a concept and asks the students to extend their discussion to a linked concept). Additional factors, such as the time lapsed after agent's last intervention, the pace of the peer discussion, etc. are also taken into consideration by the system before the agent intervenes. Despite the usual limitations of similar agents in dealing with natural language in a group discussion, the simple prompts used can still trigger student thinking and explicit reasoning. Even though MentorChat focused on interventions regarding prior knowledge, the same tool could have been used for other types as well. Corroborating previous studies, the empirical evidence recorded in the MentorChat study series showed significant domain knowledge gains for the students (both as individual and as pairs) that received agent's interventions. In addition, interventions significantly affected explicit reasoning, which in turn served as a mediator, allowing students under the directed condition to outperform students in the undirected one.

Agents in MOOCs and the colMOOC Project

Based on the outputs of MentorChat, colMOOC is a recently started Erasmus+ project that aims at integrating conversational agents and learning analytics in the context of MOOCs (Demetriadis et al., 2018). While the MentorChat activities were conducted as controlled experiments, starting in Fall 2019 colMOOC will explore the potential of conversational agents in real-life settings focusing both on the learning gains MOOC participants can reap and on how their engagement is affected (aiming at reducing the dropout rate). Similar efforts have already demonstrated a positive impact on behalf of the conversational agents (Rosé et al., 2015). To validate the effectiveness and practicability of the colMOOC approach, the agent is going to be used in three subjects: Programming for Non-Programmers, Computational Thinking, and Educational Technologies in the Classroom. The MOOCs are going to be offered in four languages in total (English, Spanish, German, and Greek), while SPOC versions of the MOOCs are going to be tested in a formal education context, allowing the comparison of the agent's potential is different situations.

Practicability of agents

Regarding the validity and practicability of using conversational agents in the classroom and "in the wild" (i.e., MOOCs), empirical evidence has already revealed both a great potential for additional learning gains and a series of factors that can hinder any positive impact mentioned earlier. Despite the simple interface and the straightforward task of configuring a domain model, the fact remains that the effectiveness of an agent depends largely on how capable the teacher is in defining the domain model and the types of interventions needed (acting both as a domain expert and as a learning designer). Higher reusability and interoperability between agents could allow teachers to use pre-configured agents in the same or similar topics. Another issue for the teacher, especially in the case of MOOCs, is to ensure student availability. This is a practical task that can nevertheless affect the outcome. In most cases, grouping students in synchronous activities in MOOCs occurs in an ad hoc basis, with little room for effective matching of learners' characteristics. The expected differences in MOOC participants' prior knowledge, skills, and motivation may not be easily addressed by an agent.

Learning to diagnose collaboratively: Validating a simulation for medical students

Anika Radkowitsch, Ralf Schmidmaier, Martin Fischer, and Frank Fischer

Physicians with different professional backgrounds often collaboratively solve a patients' problem. In those situations, physicians are expected to be able to diagnose individually by gathering and integrating case-specific information with the goal to reduce uncertainty to make a medical decision (Wildgans et al., 2018). But physicians additionally need collaborative problem-solving competences for sharing the relevant information, negotiation, as well as regulation skills for the interaction (Liu et al., 2016). Combining both, we define collaborative diagnostic competence as the competence to accurately and efficiently diagnose a patient's problem by sharing relevant information, negotiating evidence and regulating the interaction based on clinical knowledge about symptoms and diseases and meta-knowledge about the collaboration partner's discipline (e.g., Hesse, Care, Buder, Sassenberg, Griffin, 2015). Our objective is, to investigate and facilitate collaborative diagnostic competences of advanced medical students by simulating a physician with whom learners can interact to solve a patients' problem. By simulating a physician, we expect that medical students can repeatedly engage in beneficial activities that help to reconfigure internal collaboration scripts (Fischer, Kollar, Stegmann, Wecker; 2013). In order to ensure the validity of the simulation, we develop a validity argument based on Kane (2006). We see the following aspects as evidence for a satisfactory validity: if practitioners from the field rated the simulated collaboration as authentic (Shavelson, 2012); if medical students and medical practitioners with high prior knowledge showed better test performance (i.e., better and more efficient collaboration) and lower intrinsic cognitive load compared to medical students with low prior knowledge (Sweller, 1994).

Research questions of the validation study

1. To what extent do medical practitioners perceive the simulated collaborative process as authentic?
2. To what extent do medical students and practitioners with different levels of prior knowledge differ with respect to (a) their diagnostic performance (i.e., diagnostic efficiency, diagnostic accuracy, and information sharing skills) within the simulation and (b) the reported intrinsic cognitive load?

Method

In a quasi-experimental validation study, 45 medical students (5th-7th semester, low prior knowledge), 27 medical students (10th semester and higher, intermediate prior knowledge), and 26 internal specialists (more than 3 years of experience, high prior knowledge) participated. All participants worked on five case scenarios in which they first individually inspected patient information from a health record, then collaborated with a simulated radiologist by requesting a radiologic examination that was to be justified by sharing patient information and differential diagnoses, and finally solved the patient case individually by suggesting a diagnosis. All participants completed an interim-test and a post-test to assess intrinsic cognitive load (1 item, 5-point Likert scale, Opferman, 2008), as well as the perceived authenticity with respect to the collaborative process (3 items, 5-point Likert-Scale, Schubert, Friedmann, Regenbrecht, 2001). To assess the performance, we used the diagnostic accuracy (solution of the patient case and its backing with symptoms and findings), the diagnostic efficiency (diagnostic accuracy weighted by the time needed to solve a single patient case), and the information sharing skills (the inverted proportion of requests rejected by the simulated radiologist due to insufficient justification). We then calculated the mean of authenticity for both measurement times and contrasted it to a threshold (3.0) using a one-sample t-test. A mean authenticity above 3.0 indicates that practitioners on average perceive the simulation as rather authentic or authentic. Additionally, we examined the skewness of the authenticity ratings. Highly negatively skewed distributions indicate higher authenticity ratings. Further, we conducted ANOVAs with the independent variable prior knowledge and the dependent variables diagnostic accuracy, diagnostic efficiency, information sharing skill, as well as intrinsic cognitive load.

Results

Concerning research question 1, participants with high prior knowledge rated the perceived authenticity of the simulated collaborative process as $M = 3.57$ ($SD = 0.91$) which is significantly above the threshold ($t(24) = 3.14$, $p < .01$). Further, the distribution of the authenticity ratings is highly negatively skewed (skewness = -1.45, $SE = 0.64$). These results indicate that practitioners with high prior knowledge perceive the simulation as authentic. Additionally, results show that the groups differ significantly with respect to the efficiency of the diagnostic process ($F(2,93) = 15.33$, $p < .001$, $\eta^2 = 0.25$), to the information sharing skills ($F(2,92) = 6.22$, $p < .003$, $\eta^2 = 0.12$), and to intrinsic cognitive load ($F(2,95) = 26.12$, $p < .001$, $\eta^2 = 0.36$) with advanced students and practitioners being more efficient, showing better information sharing skills and lower intrinsic cognitive load compared to students with low prior knowledge. No differences between groups were found with respect to diagnostic accuracy ($F(2,94) = 3.05$, $p = .052$, $\eta^2 = 0.06$). However, we found solution rates (i.e., the final diagnosis) up to .94 for three of the five patient cases indicating ceiling effects.

Discussion

The objective of the validation study was to assess whether the developed simulation is a valid instrument to assess and facilitate collaborative diagnostic competences. We collected validity evidence (Kane, 2006), which, to a large extent, supports the validity argument: Participants with higher levels of prior knowledge are better able to justify their requests and thus convince the simulated radiologist to conduct a test by sharing more relevant information; they collaborate more efficiently, and experience less intrinsic cognitive load. However, we found ceiling effects for the case solutions making it difficult to interpret the results with respect to diagnostic accuracy. Our findings thus allow concluding that the interaction with the simulated physician is sufficiently authentic. The simulation entails valid cases and content that potentially help less knowledgeable students to advance their individual diagnostic competences and their collaborative competences by interacting with a simulated physician (Fischer et al., 2013). In further studies, we want to focus more on how to facilitate and scaffold the interdisciplinary collaboration and further take into account negotiation skills and meta-knowledge as important aspects of the interdisciplinary collaboration (Hesse et al., 2015).

References

Chang, C.J., Chang, M.H., Liu, C.C., Chiu, B.C., Fan Chiang, S.H., Wen, C.T., ... & Chai, C.S. (2017). An analysis of collaborative problem-solving activities mediated by individual-based and collaborative computer simulations. Journal of Computer Assisted Learning, 33(6), 649–662.

Dyke, G., Adamson, D., Howley, I., & Rosé, C. P. (2013). Enhancing scientific reasoning and discussion with conversational agents. Learning Technologies, IEEE Transactions on, 6(3), 240-247.

Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. Educational Psychologist, 48(1), 56–66.

Gijlers, H., Saab, N., van Joolingen, W.R., De Jong, T., & Van Hout-Wolters, B.H.A.M. (2009). Interaction between tool and talk: How instruction and tools support consensus building in collaborative inquiry-learning environments. Journal of Computer Assisted Learning, 25, 252–267.

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.

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2015). Assessing collaborative problem solving with simulation based tasks. In O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine, & S. Ludvingsen (Eds.), Exploring the material conditions of learning: The computer supported collaborative learning (CSCL) Conference 2015 (Vol. 2, pp. 544–547). Gothenburg, Sweden: International Society of the Learning Sciences, Inc.

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. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), Innovative Assessment of Collaboration (pp. 135–156). New York: Springer.

Hesse, F., Care, E., Buder, J., Sassenberg, K., Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), Assessment and teaching of 21st century skills (pp. 37–56). Dordrecht: Springer.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational Measurement (pp. 17 – 64). Westport: Praeger.

Li, H., & Graesser, A.C. (2017). Impact of pedagogical agents' conversational formality on learning and engagement. In E. André, R. Baker, X. Hu, M. Rodrigo, & B. du Boulay (Eds.), Artificial Intelligence in Education. AIED 2017. Lecture Notes in Computer Science (Vol. 10331, pp. 188–200). China: Springer.

Li, H., Gobert, J., Dickler, R., & Morad, N. (2018). Students' academic language use when constructing scientific explanations in an intelligent tutoring system. In C. P. Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), Artificial Intelligence in Education (AIED): 19th International Conference, AIED 2018 (Vol. 1, pp. 267–281). Cham, Switzerland: Springer.

Lin, T.J., Duh, H.B.L., Li, N., Wang, H.Y., & Tsai, C.C. (2013). An investigation of learners' collaborative knowledge construction performances and behavior patterns in an augmented reality simulation system. Computers & Education, 68, 314–321.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. Studies in Philosophy and Education, 27, 283-297.

OECD (2017). PISA 2015 results (Volumn V): Collaborative problem solving. Paris: PISA, OECD Publishing.

Opfermann, M. (2008). There's more to it than instructional design: The role of individual learner characteristics for hypermedia learning. Berlin: Logos.

Resnick, L. B., Michaels, S., & O'Connor, C. (2010). How (well structured) talk builds the mind. In R. Sternberg & D. Preiss (Eds.) From Genes to Context: New Discoveries about Learning from Educational Research and Their Applications (pp. 163-194). New York: Springer.

Rosé, C. P., Ferschke, O., Tomar, G., Yang, D., Howley, I., Aleven, V., … Baker, R. (2015). Challenges and Opportunities of Dual-Layer MOOCs: Reflections from an edX Deployment Study. In Proceedings of the 11th International Conference on Computer Supported Collaborative Learning, 848–851.

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.

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.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The Experience of Presence: Factor Analytic Insights. Presence: Teleoperators & Virtual Environments, 10, 266-281.

Shavelson, R. J. (2012). Assessing business-planning competence using the Collegiate Learning Assessment as a prototype. Empirical Research in Vocational Education and Training, 4(1), 77-90.

Stahl, G. (2015). Computer-supported academically productive discourse. Socializing intelligence through academic talk and dialogue, 213-224. Retrieved on May 1, 2016 from http://gerrystahl.net/pub/lrdc2015.pdf

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction, 4, 295-312.

Tegos, S., & Demetriadis, S. (2017). Conversational Agents Improve Peer Learning through Building on Prior Knowledge. Educational Technology & Society, 20 (1), 99–111.

Tegos, S., Demetriadis, S., Papadopoulos, P. M., Weinberger, A. (2016). Conversational Agents for Academically Productive Talk: A Comparison of Directed and Undirected Agent Interventions. International Journal of Computer-Supported Collaborative Learning, 11(4), 417-440.

Terveen, L. G. (1995). Overview of human-computer collaboration. Knowledge Based Systems, 8(2), 67-81.

Vogel, F., Wecker, C., Kollar, I., & Fischer, F. (2016). Socio-Cognitive Scaffolding with Computer-Supported Collaboration Scripts: a Meta-Analysis. Educational Psychology Review, 1-35.

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.

Wildgans, A., Fink, M. C., Pickal, A. J., Weber, C. …, DFG research group COSIMA. (2018). Erfassung des Diagnoseprozesses und der Diagnosequalität im Rahmen von simulationsbasierten Lernumgebungen [Assessing the diagnostic process and the diagnostic quality in the context of simulation-based learning environments]. Presentation held at GEBF 2018, Switzerland: Basel.