Does giving students feedback on their concept maps through an on-screen avatar or a humanoid robot make a difference?

Sisman, Burak; Steinrücke, Johannes; de Jong, Ton

DOI
10.1007/s12369-024-01144-y

Publication date
2024

Document Version
Final published version

Published in
International Journal of Social Robotics

International Journal of Social Robotics
https://doi.org/10.1007/s12369-024-01144-y

Does giving students feedback on their concept maps through an on-screen avatar or a humanoid robot make a difference?

Burak Sisman[1] · Johannes Steinrücke[2] · Ton de Jong[3]

Accepted: 6 May 2024

Abstract

Active or engaged learning is often seen as a way to improve students' performance concerning STEM topics. When following such a form of self-directed learning, students often need to receive feedback on their progress. Giving real-time feedback on an individual basis is usually beyond the teacher's capacity; in digital learning environments, this opens the door for exploring automated feedback. In the current study, a posttest only design was used to investigate the effect of providing students with different forms of automated feedback while they were creating a concept map about photosynthesis in an online inquiry learning environment. Participants were high school students ($N = 138$), divided over two experimental groups. In one group, feedback was given by a humanoid robot and in the other group via an avatar. The effects of the different feedback forms were compared for the two groups in terms of the frequency with which students consulted the feedback, concept map quality, and students' attitudes. Results showed that the robot group consulted feedback more often than the avatar group. Moreover, the robot group had higher scores on a scale measuring enjoyment than the avatar group. Both of these differences were statistically significant. However, the average quality of the concept maps created by both groups was similar.

Keywords Humanoid robots · Avatar · Feedback · Inquiry-based learning · Concept mapping

Learning science topics often poses a challenge to students. Students frequently have difficulty grasping the essence of the science topic to be learned and do not reach a level of deep knowledge, as noted in the PISA 2015 report, "It is also worrying to see how many young people fail to reach even the most essential learning outcomes" [30, p. 3]. Recent PISA numbers [31] show that student performance in science did not improve over the past years and that for mathematics scores even dropped considerably This lack of mastery of science topics is often attributed to students' low engagement due to directive and non-interactive teaching methods which has spurred a search for more engaging science learning experiences (e.g., [13]). These engaging or active learning approaches include, for example, inquiry learning with online labs and/or the use of interactive concept maps for knowledge expression. Overall, it is important for this active way of learning to be supported, so that students actually profit from being in charge of their own learning process [14]. For inquiry learning with online labs, this support may consist of online scaffolds [43], while students may be provided with feedback when creating a concept map [25]. In previous work we developed an online automated feedback mechanism for digital concept mapping in the context of an inquiry learning process, in which students received feedback through an on-screen avatar [1]. In the current study we introduce a humanoid robot to deliver the same feedback as the avatar does, to explore if some of the disadvantages of the avatar (e.g., students ignoring the avatar) are overcome.

In inquiry learning, students follow an inquiry cycle that resembles a scientific inquiry process. In this cycle, processes such as setting up hypotheses, designing an experiment, and drawing conclusions are central [34]. Through these processes students are active processers of information and

✉ Burak Sisman
b.sisman@tudelft.nl

Johannes Steinrücke
j.steinrucke@utwente.nl

Ton de Jong
a.j.m.dejong@utwente.nl

1 Department of Cognitive Robotics, Delft University of Technology, Delft, Netherlands

2 Department of Cognition, Data and Education, University of Twente, Enschede, Netherlands

3 Department of Instructional Technology, University of Twente, Enschede, Netherlands

Published online: 06 June 2024

engage in extending and adapting their knowledge base, which is assumed to lead to deeper knowledge [12, 17]. Inquiry processes, however, are rather complex, and require good structuring and guidance in the learning environment in order to be effective [27]. Additionally, for inquiry learning to be effective, students need adequate initial knowledge to build upon [19, 20]. For example, to be able to create informative hypotheses, learners need to have sufficient knowledge of the variables in the domain involved [24]. One way to create such an initial knowledge base is to have students produce a concept map in the starting phases of the inquiry cycle.

Concept maps are used to display relationships between concepts. In other words, concept maps are 2-dimensional diagrams that enable information to be organized by visualizing concepts and their organization [21]. Concept maps are used as a tool in teaching and learning, as well as in evaluating conceptual understanding and knowledge. In the context of inquiry learning, creating concept maps can be considered part of the orientation and conceptualization phases of the inquiry cycle, which occur at the beginning of an inquiry cycle [34]. Students re-activate prior knowledge by creating a concept map, and after receiving feedback, they can also construct a solid foundation for the subsequent inquiry process. However, giving feedback to each individual student takes too much time in a face-to-face educational environment and thus may be not doable by the teacher. Furthermore, the timing of feedback is also important. If a student's expression in a concept map needs corrective feedback, the feedback should be provided as soon as possible so that the learner can react accordingly. If feedback is not provided until the end of the task, learners will not be able to correct their concept map during the inquiry process [35]. Automated feedback may be a solution for this, and some tools have been designed for that purpose (e.g., [22]). More specifically for our context, Anonymous [1] developed an automated feedback tool that is part of the Go-Lab ecosystem, an online sharing and learning platform for inquiry-based learning [15]. Anonymous [1] demonstrated that their tool could effectively assess the quality of concept maps and provide accurate and helpful feedback on a number of specific shortcomings that are frequently visible in students' concept maps. However, their results showed that students with feedback available frequently did not consult it or did not fully utilize the feedback provided to them. In the Anonymous [1]'s study, feedback was given via a virtual agent (avatar). In the current study we investigate what happens if the feedback is given in a way that is more attractive and engaging than through an onscreen avatar, that is, via a humanoid robot.

The use of robots is becoming an increasingly common practice influencing different aspects of daily life [7]. More specifically, the use humanoid robots has become popular in the educational field [37]. Research into the use of robots in education has highlighted the positive influence of robots on the cognitive and affective dimensions of learning, attributing this impact mostly to the robots' ability to display social behavior that encourages learners to participate in the learning process [5]. Research has also indicated an increase in learners' achievement of cognitive learning objectives following the robots' presentation of content [28]. Furthermore, it has also been noted that robots can display socially supportive behavior and provide personalized aid by naming the learners and referring to previous interactions [5]. Their distinguishing characteristics of repeatability (i.e., the ability to consistently perform specific tasks or behaviors), humanoid appearance, intelligence, sensing capability, flexibility, interaction, body motion capability and adaptability allow robots to interact with learners in varying roles, as teaching assistant, peer, teacher and/or teaching resource/material in the class [4, 8].

In the literature, there are studies comparing robots and avatars in terms of various variables in different domains. Pan and Steed [33] conducted a study to compare users' trust in expertise in avatar-, video-, and robot-mediated interaction. They analyzed participants' advice-seeking behavior in limited advice and risk situations as an indicator of trust. They found that participants were less likely to choose advice from the avatar, regardless of whether the avatar was an expert or not. In the study, the avatar scored the lowest on the trust assessment, while the robot and video were rated similarly. van den Berghe et al. [41] conducted a study to compare children receiving a programming training with a non-humanoid robot or an avatar. In the study, they compared learning to program a robot or an avatar. According to results of the study,although no differences in self-reported motivation or cooperation during the training were found, children showed higher learning outcomes when learning to program a robot rather than an avatar. Moreover, in the study of [11] on emotional storytelling using virtual and robotic agents, it was reported that the physically embodied robot garnered greater narrative attention from listeners compared to a virtual embodiment. Additionally, the study found that human voice narration was favored over the current text-to-speech technology. Furthermore, the results revealed a multifaceted relationship between the emotional content of the story, the facial expressions of the narrating agent, and the emotional responses of the listener. Notably, the empathetic engagement of the listener was demonstrated through observable facial expressions.

In the research, humanoid robots are used most frequently in foreign language teaching and in the field of special education [32]. However, robots can also be suitable supports for science instruction. Robots can provide motivation for students to learn science, relieve students' anxiety and create a fun learning environment. For instance, humanoid robots can be interactive feedback providers that may improve interest and motivation to learn scientific subjects. Only a few articles have explored this type of activity in science education (e.g., [3, 9]). In the current study, students were asked to create a concept map in an early stage of an inquiry process and received feedback on their concept map either from an avatar or a humanoid robot. Both avatar and robot followed the same rules for generating and delivering feedback. Our study examines the following research questions:

What effect, if any, does provision of feedback by a humanoid robot as compared to an avatar?

- *RQ1* What effect, if any, does provision of feedback by a humanoid robot as compared to an avatar have on the frequency of students' accessing available feedback?
- *RQ2* What effect, if any, does provision of feedback by a humanoid robot as compared to an avatar have on the quality of the students' concept maps?
- *RQ3* What effect, if any, does provision of feedback by a humanoid robot as compared to an avatar have on students' attitudes?

1 Method

In this study, students were asked to create a concept map on the topic of photosynthesis. Students were randomly assigned to one of two groups. While a humanoid robot provided feedback for the concept maps of the students in one condition (HRC), the feedback was provided by an avatar in the other condition (AC).

1.1 Participants

In total, 138 students (58% male) from two Dutch secondary schools participated in the experiment. The students were all in their second year and aged around 13 years. Participants from each class of each school were randomly assigned to one of the two conditions: the humanoid robot condition ($n = 73$, 64.4% male) or the avatar condition ($n = 65$, 50.8% male).

1.2 Learning Environment

An inquiry learning environment (ILS, Inquiry learning Space) for the topic of photosynthesis in biology that was designed for our previous study Anonymous [1] formed the basis for the ILS in the current study. The ILS in the current study focused on the starting phases of the inquiry learning cycle. The ILS we used had a basic text-based introduction on photosynthesis as the starting phase, it then proceeded by presenting an example concept map about fruits and vegetables to demonstrate how to use the concept mapping tool, and in its third phase gave a concept mapping tool to create the concept map about photosynthesis. Students were asked to create a concept map on photosynthesis from scratch without any predefined concepts available. After that the ILS presented students with a brief questionnaire about participants' experiences with the avatar or humanoid robot while they were creating their concept map. All materials were presented in the students' native language, what is presented in this article are translations.

The concept map was at the center of the current experiment. Figure 1 displays an 'expert' concept map for the topic of photosynthesis, which was used as the reference concept map in this study.

1.3 Feedback

Feedback was based on the algorithms from our previous study [1]. The concept mapping tool used those algorithms to evaluate every action by a student (adding new concepts, editing existing concepts, creating a relation between concepts, etc.) while creating their concept map. Both humanoid robot and avatar used the same algorithms.

After each change to the concept map, a list of all possible feedback was created and sorted, and the student was presented with the most relevant feedback prompt. The types of feedback students could receive were: (a) a suggestion to add a specific concept to the concept map, (b) a suggestion to add a proposition (link between two concepts) to the concept map, (c) a question asking the student if an irrelevant concept in their concept map is really necessary, (d) a question asking if an irrelevant proposition is necessary, (e) a suggestion to change the direction of a proposed relation, (f) a suggestion to change a label, or (g) a suggestion to add an intermediate concept in a proposition. These suggestions and questions were based on a comparison between the reference map (Fig. 1) and the student concept map. In mapping the student concept map onto the reference map, synonyms (e.g., 'sugar' replacing 'glucose' and 'O_2' replacing 'oxygen') and potential typos (based on Levenshtein distance, [26, 29]) were taken into account.

The timing of the feedback was based on the student's actions in the concept map. This means that feedback was presented within a certain timeframe after a meaningful change (excluding, e.g., changing the position of a concept) or as a result of the student's responses to the preceding feedback. The relevance of feedback prompts was based on a combination of the type of feedback and the state of the student concept map. When starting their concept map, students were guided towards adding the main concepts and propositions. When the concept map had reached a certain threshold, feedback was aimed at making specific improvements to the concept map. As the feedback algorithm relied on being able to identify content in the student concept map, fixing potential

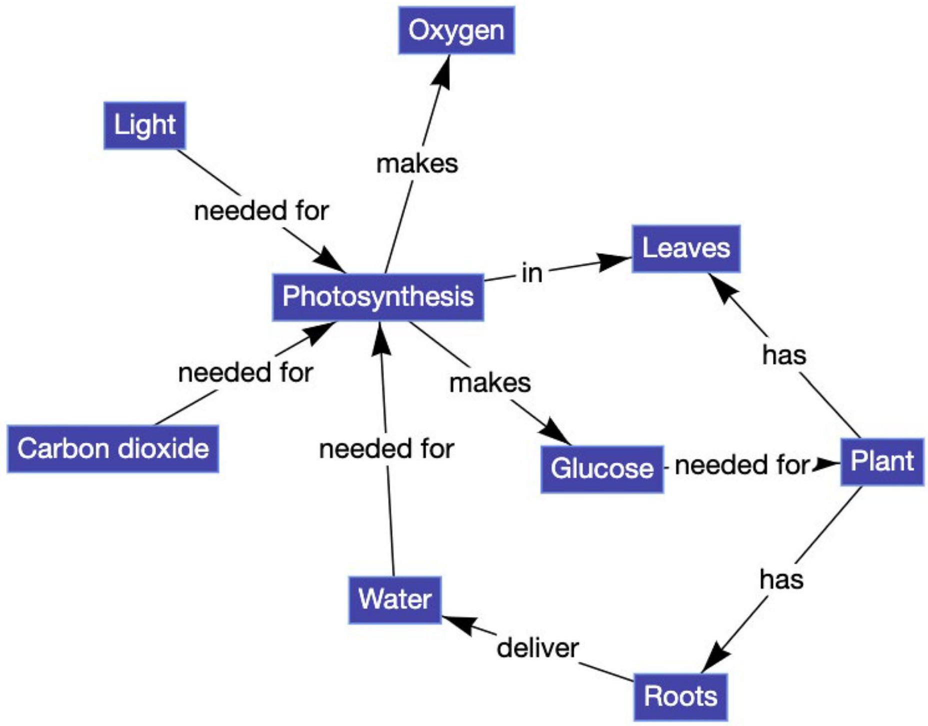

Fig. 1 The reference concept map for photosynthesis

typos and removing irrelevant information was prioritized. In all cases, students had the option of immediately implementing the prompt (e.g., adding, removing, or relabelling concepts and propositions) or ignoring and suppressing it. Each specific prompt was only suggested to a student once. Feedback prompts (including all the possible prompts that were not presented), student concept maps, and changes to those concept maps were saved in the learning analytics logs for later use. The intention of the feedback was to assist students in developing effective concept maps. If a student successfully creates an effective concept map independently, then they may not require any feedback. Therefore, it is understandable that students who produce inadequate concept maps will receive more feedback—as this aligns with the purpose of providing constructive feedback.

1.3.1 Humanoid Robot Feedback

A NAO model humanoid robot was used as a feedback provider. The robot was placed on the table in a standing position next to the computer that the student used. A picture of the humanoid robot's setting is given in Fig. 2. The robot was in alive mode, which means that it tracked the student's face with its head to make eye contact. To alert students to available feedback (determined based on the feedback rules outlined in the previous section), the humanoid robot raised its right hand and waited for 10 s. If a student touched one of its hands within 10 s, the robot gave the feedback by speaking to the students in their native language. If the student did not touch one of the hands, this was taken as a sign that no feedback was wanted. In that case, the robot lowered its hand. If the feedback was a question requiring an answer from the student, the robot listened to the student's response. When the robot got a response, it transmitted the response to the concept mapping tool, which then acted according to the response. For instance, suppose that the student had the concept 'orange' in their concept map, and assume that the feedback from the robot was "Do you need the concept 'orange' in your concept map?". If a student responded orally, "Yes", the tool did not perform any action; if the response was "No", the tool deleted the concept 'orange' from the concept map.

1.3.2 Avatar Feedback

The feedback process for the avatar was essentially identical to that for the humanoid robot, but the avatar had a different physical presence. To alert students to available feedback, the avatar popped up in the bottom-right corner of their computer screen. If the student clicked on the avatar, indicating that they would like to see the feedback, the feedback was presented next to the concept map as a speech balloon coming from the avatar. Possible responses were identical in function to those used in the HRC, though the specific wording might

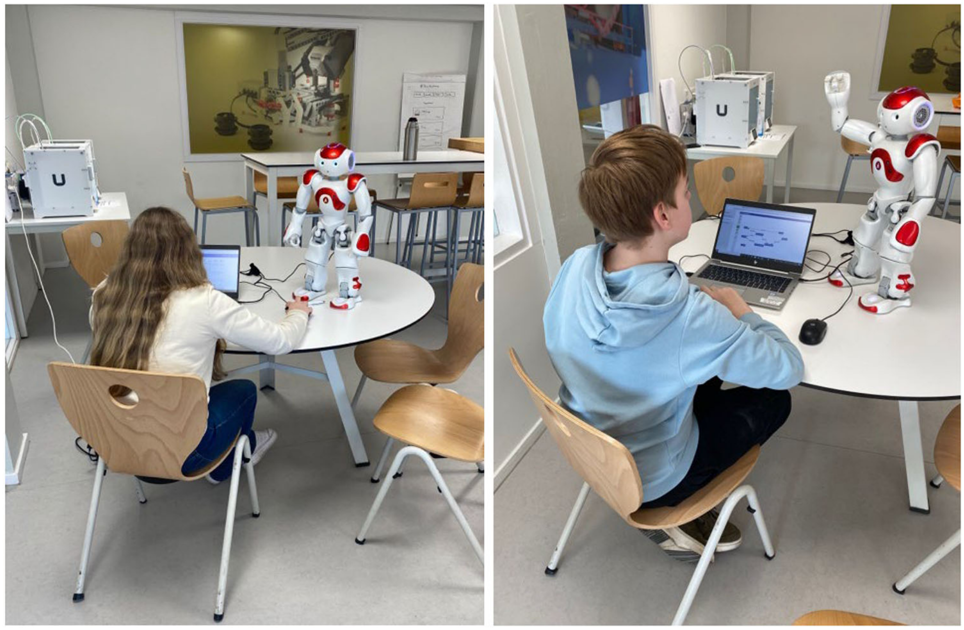

Fig. 2 Pictures from the humanoid robot setting

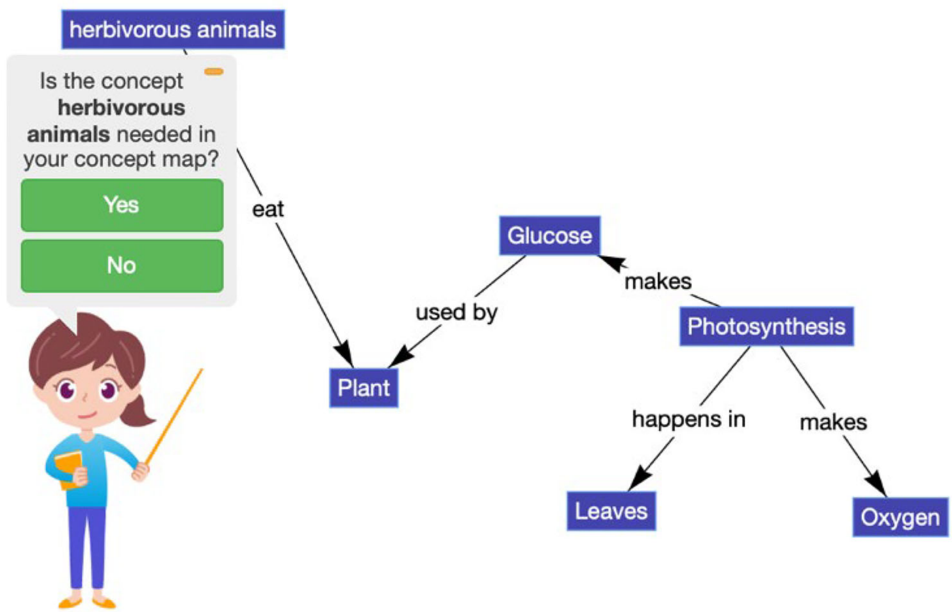

Fig. 3 Screenshot of feedback by avatar for a superfluous concept

be changed to improve clarity. A screenshot of the avatar is shown in Fig. 3.

1.4 Measurement Instruments

1.4.1 Students' Use of Feedback

To answer the first research question, interaction logs such as occurrences of feedback being offered and consulted were saved. We calculated how often students used the feedback that was offered, by dividing the number of times feedback was consulted by the number of times it was offered.

1.4.2 Concept Map Quality

In order to evaluate the quality of concept maps, the first author and research assistants used four criteria based on the relevant literature [38] in a hand-coding process. The criteria used were: (1) topic-relevant concepts, (2) relevant propositions, (3) correct concepts, and (4) correct propositions. The team performed the hand-coding process jointly and aimed to reach consensus on the quality of the concept maps. This strategy facilitated an iterative and collective assessment procedure, which enhanced both dependability and validity of scoring. Furthermore, the first author and research assistants compared the concept map with a reference concept map for correctness.

Topic-relevant concepts and propositions were simply the number of concepts and propositions—nodes and edges—present in the student concept map and relevant for the topic in general (in this case, photosynthesis). The inclusion of a greater number of relevant elements in a student concept map was assumed to be associated with greater understanding of the topic, with students being able to name more concepts relevant to the topic and identify more of the connections between them.

Evaluation of the remaining two criteria, correct concepts and propositions, was done by comparing the students' concept maps to a reference concept map (see Fig. 1). A concept was marked as correct when it was present in the reference map; a similar approach was used for propositions. It should be noted that when comparing propositions, the causal direction and label for the proposition were ignored. For example, the proposition "photosynthesis requires water" was considered equivalent to the proposition "water is used in photosynthesis", as both propositions connect photosynthesis and water.

For each criterion, we counted the number of concepts or propositions in each student concept map. Next, we classified the quality of the concept map based on a four-point scale ranging from excellent through good and fair to poor, with the use of the rubrics shown in the Appendix. To have a numerical indication of the quality of the concept map, each of these four criteria was transformed into a numerical score (see the Appendix). The total concept map quality score was determined as the average of the scores obtained for these four criteria.

For example, Fig. 4 shows a concept map created by a student (Translated from the student's native language). This concept map contains 11 concepts and 10 propositions. Regarding the first criterion, we gave a rating of Excellent (10 points) because that rating applies when there are 9 or more relevant concepts, and the student's concept map contains 11 relevant concepts. Similarly, for the second criterion, we gave a rating of Excellent (10 points) because the map contains 10 relevant propositions. For the third criterion, we gave a rating of Fair (6 points), because although the majority (64%) of the ideas are correct, some concepts are incorrect, such as substances of interest, ground, sun, men, and animals, and some concepts are missing, such as roots and leaves. Finally, for the fourth criterion, we gave a rating of Good (8 points), because three of the 10 propositions are false (70% correct). For example, while the proposition linking glucose and plant is in the reference map, in the student map there are separate propositions linking photosynthesis and plant, as well as photosynthesis and glucose. Averaging the scores for the four criteria gives a total score of 8.5 for the student concept map in Fig. 4.

1.4.3 Attitude Test

Sisman et al. [37] developed and validated a scale for measuring attitudes towards robots, which consists of the subscales Engagement, Intention, Enjoyment, and Anxiety. The scale items were created as five-point Likert-type questions (from 1 = strongly disagree to 5 = strongly agree),according to a reliability analysis, the internal consistency coefficient for the whole scale was 0.90 [37]. For this study, the scale was slightly adapted. The item "I enjoy lessons that are handled using a robot" means that the lesson was taught by the robot. However, the robot just gave feedback in this study, it did not teach. Therefore, that item was removed. A new item, "I like getting feedback from the humanoid robot" was added to the Enjoyment subscale, in line with similar feedback studies in the literature [10]. This item directly refers to receiving feedback from the robot. A parallel version of the questionnaire for the use of an avatar was created based on the opinion of two domain experts. The new version of the scale is given in Table 1. The scale was translated into Dutch and checked and corrected by two English and Dutch language experts.

Cronbach's alpha for the total Attitude scale in this study was 0.92. For the Engagement subscale $\alpha = 0.79$, for the Anxiety subscale $\alpha = 0.73$, for the Intention subscale $\alpha = 0.81$, and for the Enjoyment subscale $\alpha = 0.86$. Moderate skewness and kurtosis were observed for the Enjoyment, Anxiety and Attitude (sub) scales (see Table 2).

1.5 Procedure

The experiment was conducted at two schools in the east of the Netherlands. In each school, two NAO humanoid robots were used. All students were randomly assigned to the HRC or AC groups before the experiment. The logistical arrangement differed slightly between the two schools. In the first school, two rooms were set up for the AC and one larger room for the HRC. The rooms for the AC each had one notebook computer, and the other larger room for the HRC had two humanoid robots and two notebooks that were next to the robots. The larger room was used by two participants at the same time. The space was large enough to isolate the participants from each other and the space was isolated in such a way that the participants could not see or hear each other. In the second school, students in the AC worked in a room with two notebooks, so that two participants used the space at the same time. The other two rooms each had a notebook and a humanoid robot that was standing next to the notebook. In both schools, on the morning of the actual experiment, one of the researchers informed students in each participating class as a group about the goals of the experiment, extent and purpose of data gathered, and students' right to withdraw at any time. The participants were also

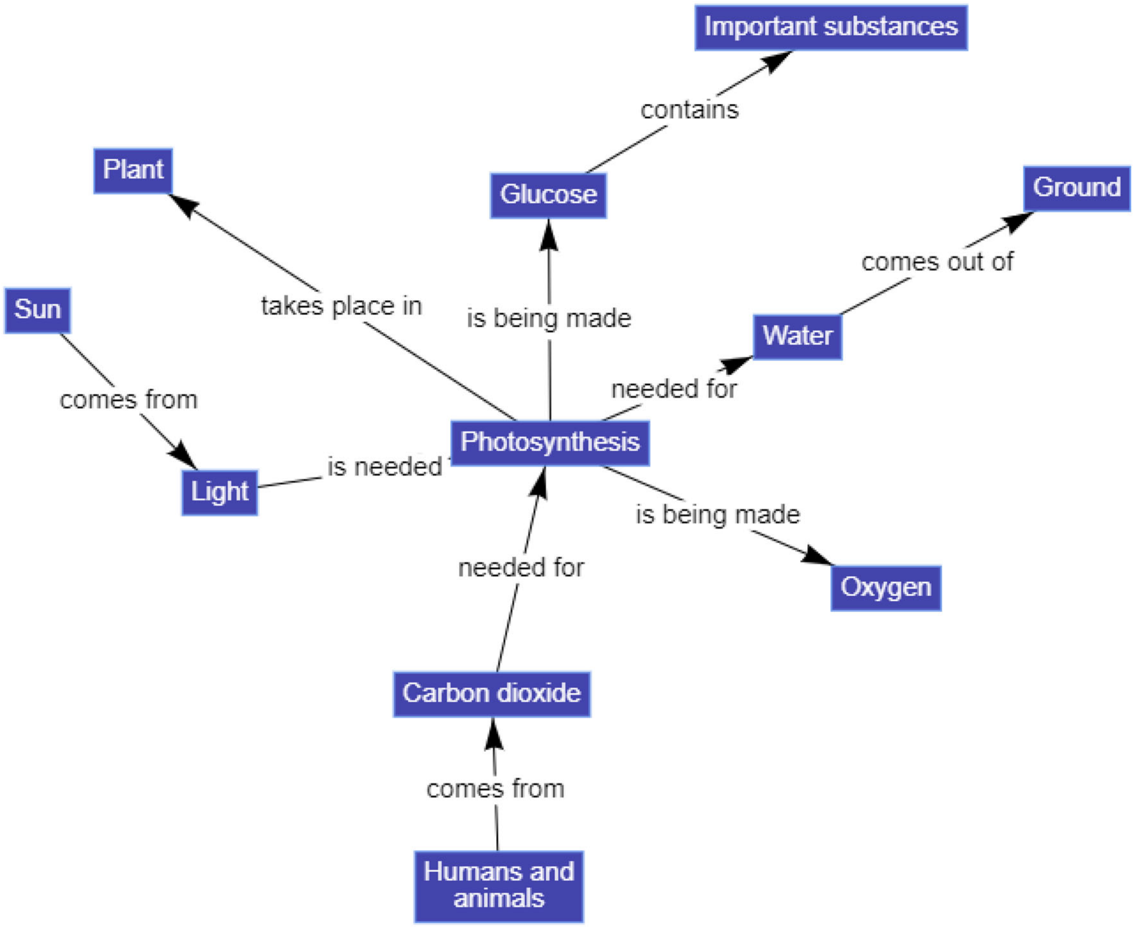

Fig. 4 Example of a student concept map

given an introduction to concept maps and the concept mapping tool in general. They were also instructed on how they would be called to the experiment individually throughout the day. The experiment was organized in such a way that four students (two from each condition) could participate in the study at the same time. The participating students were asked to direct the next students, or one of the researchers directed the next students to the relevant rooms. Students used their student number to log in to the ILS. One of the researchers or a research assistant gave a quick introduction about the experiment. Students were provided with a brief refresher on photosynthesis in the first phase of the ILS. In this section, there was a short paragraph about what photosynthesis is and the process of photosynthesis accompanied by two visuals, one showing a very general impression of the role of photosynthesis in food production and the other one showing a plant in the soil with a sun in the sky and indicating that in a leaf of the plant carbon dioxide goes in and oxygen comes out. Following that, one of the researchers or a research assistant introduced the use of the concept mapping tool for the example topic of fruits and vegetables. After this, students had 15 min to complete the concept map on photosynthesis. After 10 min, the student was informed by the experiment leader that there were 5 min left. After 15 min, students were instructed to stop working on their concept map and to fill out the questionnaire.

1.6 Ethical Consent

The experimental procedure was approved by the ethical committee of the University of Twente. Both secondary schools have agreements with all parents covering research for the purpose of improving education. The topic of the learning environment was aligned with students' regular curriculum to minimize the impact on students. Given these circumstances, students' passive consent was deemed appropriate.

Students were informed of the purpose of the experiment, the extent and purpose of data gathered during the experiment, and their right to withdraw from the experiment at any time. Contact information for the researchers was provided directly to the students, and it was made clear that they could also direct any questions to their teacher – who would contact the researchers if needed. Data collection took place in

Table 1 Items from the attitudes questionnaires

Sub-category		Questionnaire items for the robot	Questionnaire items for the avatar
Engagement	1	I try to make sense of the lesson better when a robot is used in the lesson	I try to make sense of the lesson better when an avatar is used in the lesson
	2	I can pay more attention to the lesson when there is a robot in the class	I can pay more attention to the lesson when there is an avatar in the learning environment
	3	I can consider the robot as a teacher as it talks and moves like a human	I can consider the avatar as a teacher as it talks and moves like a human
		I feel more interested in the topic when there is a robot in the class	I feel more interested in the topic when there is an avatar in the learning environment
Intention	5	I would like robots to take part in the lessons in the future	I would like avatars to take part in the lessons in the future
	6	I would like to have a robot that would keep me company while studying at home	I would like to have an avatar that would accompany me while studying at home
	7	I would like robots to be utilized in other lessons as well	I would like avatars to be utilized in other lessons as well
	8	I think robots should be used in the lessons because the age of technology necessitates this	I think avatars should be used in the lessons because the age of technology necessitates this
Enjoyment	9	It is interesting to see a robot acting and talking like a human	It is interesting to see an avatar acting and talking like a human
	10	Communicating with the robot is fun	Communicating with the avatar is fun
	11	I like communicating with the robot on a one-on-one basis	I like communicating with the avatar on a one-on-one basis
	12	I enjoy it when the robot understands what I am saying and responds to me	I enjoy it when the avatar understands what I am doing and responds to me
	13	I like getting feedback from the humanoid robot	I like getting feedback from the avatar
Anxiety*	14	I don't like the robot to be in the class	I don't like the avatar to be in the learning environment
	15	The existence of a robot in the class causes disruption	The existence of an avatar in the learning environment causes disrupt
	16	Communicating with the robot is difficult	Communicating with the avatar is difficult
	17	I feel nervous while talking with the robot	I feel nervous while talking with the avatar

*Negative items, reverse-scored

Table 2 Skewness and kurtosis per (sub) scale

(Sub)scale	N	Skewness	Kurtosis
Engagement	138	− 0.34	2.66
Intention	138	− 0.33	2.74
Enjoyment	138	− 0.72	3.50
Anxiety	138	− 0.56	3.44
Attitude	138	− 0.54	3.20

the school during school hours. Students were guaranteed confidentiality of data handling.

1.7 Data Analysis

Multiple analyses were conducted. First, to analyze students' consultation of feedback, logfiles were analyzed with regard to feedback interaction frequency. As data were not normally distributed, a Mann–Whitney U-test was used to investigate between-condition differences.

Second, the quality scores for the concept maps were determined on the basis of the criteria presented above. To test for significant differences in total quality scores for the concept maps between the HRC and the AC, an independent samples t-test was carried out.

Table 3 Feedback consultation rate per condition

Condition	n	Mean	SD
HRC	39	62.31	26.44
AC	37	19.51	20.85

Third, students' attitudes were compared between the HRC and AC conditions using analysis of variance (ANOVA; Attitude scale) or multivariate analysis of variance (MANOVA; on subscale level), as the four attitude subscales were moderately correlated. Subsequently, estimated marginal means were computed to investigate the direction of potential differences. Prior to all analyses, descriptive statistics were computed for all variables analyzed.

2 Results

The study was conducted at two schools, but for technical reasons data on avatar and robot feedback access were not available for the first school. For that reason, results for the attitude questionnaire and concept map quality are based on data from two schools, whereas data on the use of feedback are only available for one school.

2.1 Students' Accessing of Feedback

Students' access of feedback was calculated from the logfiles, where we determined how often students clicked the avatar to see the feedback or tapped the hand of the robot to hear the feedback. The consultation rates pertaining to the feedback messages were computed for both conditions. Descriptive statistics are given in Table 3. To determine whether these feedback consultation rates data were normally distributed, a Shapiro–Wilk test was performed. The result ($W = 0.914$, $p < 0.01$) showed that the data were not normally distributed. Therefore, we performed a non-parametric Mann–Whitney U-test to determine whether a significant difference existed between the HRC and AC conditions. A total of 62.31% feedback messages were consulted by the students in the HRC, whereas only 19.51% of feedback messages were consulted by the students enrolled in the AC. There was a significant difference between the conditions in terms of consultation rate in favor of the HRC ($W = 172$, $p < 0.001$, $r = 0.35$). Statistical power was determined to be 0.51, sensitivity analysis yielded a required effect size of 0.44 to achieve 0.8 power.

2.2 Concept Map Quality

With skewness and kurtosis values between − 1.5 and 1.5, these data followed a normal distribution [40]. On average, students scored 4.80 points (SD = 2.42). Skewness and kurtosis were − 0.08 and − 1.01 respectively. The means and standard deviations of the total scores and the scores for each criterion are presented in Table 4 for both conditions. Results of independent samples t-tests are also given in Table 4. The results showed no significant difference between the total scores for the two conditions ($p = 0.838$), and no significant differences on the separate criteria. Therefore, we can conclude that there was no difference in terms of concept map quality between the two conditions. Sensitivity analyses yielded that, given 0.8 power, an effect size of 4.8 would have been required.

2.3 Attitudes Towards Robot and Avatar

The Attitude questionnaire was administered at both participating schools. Table 5 presents descriptive statistics summed over both schools.

To test for significant differences in students' attitudes towards the robot and avatar, an ANOVA was carried out on the total score (Attitude), yielding no significant difference: $F(1, 136) = 0.39$, $p = 0.532$. Assuming statistical power of 0.8, sensitivity analyses yielded a required effect size of 0.24. As described in Table 6, the subscales correlate significantly. Consequently, MANOVA analyses with the subscales as dependent variables were performed, yielding a significant between-condition difference: $F(4, 133) = 6.38$, $p < 0.001$, $V = 0.06$. This significant between-condition difference stemmed from the Enjoyment subscale, where participants in the HRC had significantly higher scores than their peers in the AC: $t(136) = –2.47$, $p = 0.015$. No significant differences were found for the subscales Engagement, Intention and Anxiety. For the MANOVA analysis we achieved a statistical power of 0.62. Sensitivity analysis yielded a required Pillai's trace V of 0.08, given a statistical power of 0.8.

3 Conclusion and Discussion

In the current study, students had to construct a concept map and received feedback during the construction process. This is a common approach to help improve the quality of the concept map and traditionally this feedback is given by humans, either the teacher (e.g., [36]) or fellow students (e.g., [16]). Feedback given by humans is hard for the receiver to ignore, but also time-consuming for the feedback giver. Therefore, alternatives in the form of automated feedback have been developed. However, this feedback, often presented by prompts or avatars, is easier for students to ignore than human feedback Anonymous [1]. The current study sought to find out if offering students feedback from a humanoid robot meant that students consulted the feedback more frequently compared to feedback coming from an avatar and looked into the effects of the humanoid robot versus an avatar on students' experiences and the quality of their concept maps.

Table 4 Concept map quality as total score and per criterion

	Robot			Avatar			*t*	d*f*	*p*
	n	M	SE	n	M	SE			
Total score	73	4.85	.28	65	4.76	.30	.20	136	.838
Criterion 1	73	5.53	.29	65	5.89	.32	− .83	136	.406
Criterion 2	73	5.00	.33	65	4.74	.39	.52	136	.606
Criterion 3	73	5.25	.31	65	5.02	.32	.52	136	.607
Criterion 4	73	3.60	.33	65	3.40	.31	.45	136	.657

Table 5 Descriptive statistics for attitude scale and its subscales per condition

	Robot			Avatar		
	n	M	SD	*n*	M	SD
Engagement	73	3.32	.83	65	3.18	.91
Intention	73	3.32	.92	65	3.30	.91
Enjoyment	73	3.91	.74	65	3.58	.783
Anxiety*	73	3.62	.89	65	3.87	.72
Attitude	73	3.56	.72	65	3.49	.71

*A score of 1 is the lowest level of the measured construct and 5 is the highest. The Anxiety score was reversed to line up with the other subscales, with a higher score meaning a more favorable attitude

Table 6 Correlations between the subscales

	Engagement	Intention	Enjoyment	Anxiety
Engagement				
Intention	0.66*			
Enjoyment	0.70*	0.63*		
Anxiety	0.56*	0.53*	0.59*	

* $p < .001$

Within the scope of RQ1, our results indicate that students tended to consult the robot feedback more frequently than the avatar feedback. This may be because the robot's embodied presence is more noticeable and trustable than a non-physical object such as an avatar. Bainbridge et al. [2] investigated the impact of a robot's physical presence on human evaluations of the robot as a social collaborator. Their results indicated that individuals were more inclined to complete trust-related activities when they interacted with the robot in vivo rather than through live video transmission, the latter being similar to a screen avatar. In addition, the fact that the humanoid robot raised its hand when there was feedback available may also have attracted the students' attention. This feature of the robot may have served as a visual and audible cue to alert students to engage with the feedback.

For RQ2, we had expected that students would produce higher quality concept maps if they were willing to follow the advice of an agent. Since students more often consulted the advice given by the robot, we might have expected higher concept map quality in the HRC condition. However, we found no significant difference between the quality of the concept maps created by students in the HRC versus the AC. Students generated fairly low-quality products ($M = 4.85$ for HRC, $M = 4.76$ for AC) regardless of the type of agent they interacted with. The most obvious explanation may be that students dismissed the recommendations offered by the robot, although in a few cases the robot advice may have been suboptimal. For example, this was the case when students spoke in a low voice and the robot did not understand the student's response. In that case, the robot prompted for a repetition. However, if no suitable reply was received even after two attempts, no automatic modification of the concept map was applied, leaving the concept map as it was.

Within the scope of RQ3, in terms of attitudes, no significant overall difference was found between students' attitudes towards the robot and the avatar. On the subscale level, students who interacted with the robot reported a higher level of enjoyment than students who received feedback through the avatar. This may have been because students perceived the robot as more entertaining than an avatar, due to its three-dimensional human-like appearance and interactive behavior. In relation to this, research has shown that human-like social robots tend to evoke more positive emotional responses from users than systems lacking the humanoid form [6]. The presence of a robot may also play a significant role in students' motivation [23, 42]. Therefore, based on the findings of previous studies and our current research results, it could be concluded that using humanoid robots to provide feedback to students may have advantages in terms of attracting student attention, increasing engagement with and utilization of the feedback, and enhancing students' enjoyment and motivation [39].

No significant differences in engagement and intention were found between those using the robot or the avatar but the descriptive data here also showed an advantage for the HRC. Previous research has shown that the impact of the physical presence and embodiment of social robots on engagement and intention is complex and may depend on various factors such as robot behavior, task demands, and individual user characteristics [18]. This, of course, makes the proper design of robot instruction a challenging task. What may complicate matters could be that a robot may evoke more anxiety for some students than an avatar. We did not find a significant difference in anxiety between the two conditions, but the descriptive data suggested that the lack of physical presence in the avatar condition could create a less threatening atmosphere for some students, resulting in a lower level of anxiety. However, due to the lack of evidence for this, we can only speculate about this.

The limitations of our study include the relatively small sample size, the use of only one type of humanoid robot, and the short duration of our intervention. Future research could investigate the impact of different designs for robot behavior and various types of robot tutors on students' engagement, motivation, intention to use feedback, and anxiety levels in order to gain a more comprehensive understanding of how social robots can be most effectively implemented in educational settings. In this context, it may have been the case that the robot by its very nature was more attractive to students than the avatar. A study set-up in which also a condition with a robot-like avatar is included could have shed more light on this aspect. Also, the intervention in our study only lasted for 15 min, which is shorter than a typical instructional intervention. It should be tested if students' results and attitudes are similar when a longer intervention is used. Furthermore, the study did not account for individual student characteristics such as prior experience with or attitudes towards technology, which may have affected their engagement and motivation levels. Additional research could explore how the social robot tutor can adapt to individual differences to provide more personalized education and support. Another limitation may be not having a control group to determine the effectiveness of the robot and avatar conditions compared to a traditional human-led tutoring approach or a control group without feedback. Additionally, another limitation of the study is that the intervention duration is limited, which may not adequately reflect the long-term impacts. The fact that no study has been conducted to understand the anxiety of students due to communication with robots can be stated among the limitations of this study. In addition, the application of a pre-test to determine the level of familiarity of students with robots could have been useful in terms of evaluating parameters such as the novelty effect. For future studies, we can suggest that researchers should consider issues such as the robot's inability to understand the speech of students who speak in a low voice, rarely the touch sensors do not work, and rarely communication errors.

Our study contributes to the growing body of literature on the use of robots and avatars in educational settings. Our results suggest that robots can be more effective than avatars in promoting enjoyment of the learning process and lead to higher accessing of feedback than an avatar does. Implementing robots in the actual school practice schools/teachers would, of course, still have to deal with aspects as costs, set-up, maintenance etc. Future research may consider exploring the mechanisms behind the effects we found and ways to optimize the use of humanoid robots and avatars to enhance student learning experiences.

Acknowledgement The authors express their sincere gratitude to the participants in this study for actively participating and being receptive towards a new technological innovation. Additionally, recognition is given to Firdevs Sirma, Busra Bozkurt, Tugba Sevval Gumus, and Senanur Hiz - the research assistants responsible for coding students' concept maps. A special acknowledgement is also extended to Karel Kroeze who developed the feedback mechanism and avatar system utilized in this investigation and who also aided in the data collection. Furthermore, Joris de Vries (Het Noordik Lyceum) and Gerbrand Wilbrink Med (Canisius) are acknowledged for allowing their students' participation within this study along with providing valuable support throughout its duration.

Data Availability The dataset generated during and /or analyzed during this current study are not publicly available due to the protection of participants' privacy but are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Numerical Indication of the Quality of the Concept Map

This appendix consists of the quantitative representation of the concept map's quality, evaluated against four criteria derived from the literature.

	Excellent (10 points)	Good (8 points)	Fair (6 points)	Poor (4 points)
Number of concepts	The concept map includes a large number of concepts (9 or more) that are relevant to the topic and cover a wide range of related ideas. The concepts are well-organized and demonstrate an in-depth understanding of the subject matter	The concept map includes a sufficient number of concepts (between 6 and 8) that cover the main ideas of the topic. The concepts are well-organized and indicate a solid understanding of the subject matter	The concept map includes a limited number of concepts (between 3 and 5) that only cover some of the main ideas of the topic. The concepts may be less well-organized and indicate a somewhat limited understanding of the subject matter	The concept map includes a very small number of concepts (2 or fewer) that do not adequately represent the topic. The concepts may be poorly organized and fail to demonstrate a clear understanding of the subject matter
Number of propositions	The concept map includes a large number of relations (9 or more) that show how the concepts are connected to each other in a comprehensive and meaningful way. The connections between concepts are clear, well-organized, and effectively demonstrate the relationships between the concepts	The concept map includes a sufficient number of relations (between 6 and 8) that show the main connections between the concepts in a clear and organized way. The connections between concepts are well-defined and demonstrate a solid understanding of the relationships between the concepts	The concept map includes a limited number of relations (between 3 and 5) that only show some of the connections between the concepts. The connections may be less well-defined, and may not fully capture the relationships between the concepts	The concept map includes very few (2 or fewer) relations, making it difficult to understand the connections between the concepts. The connections may be absent or unclear, and may fail to demonstrate a clear understanding of the relationships between the concepts
Correctness/incorrectness of concepts	The concepts in the concept map are all correct and accurately represent the topic (i.e., 100% accuracy)	The majority of the concepts in the concept map are correct and accurately represent the topic (i.e., 70–99% accuracy)	Some of the concepts in the concept map are correct, but others are inaccurate or not fully representative of the topic (i.e., 50–69% accuracy)	Most of the concepts in the concept map are inaccurate or do not adequately represent the topic (i.e., less than 50% accuracy)
Correctness/incorrectness of propositions	The relations between the concepts in the concept map are all correct and accurately represent the connections between the concepts (i.e., 100% accuracy)	The majority of the relations (i.e., 70–99% accuracy) in the concept map are correct, but there may be a few minor inaccuracies	Some of the relations in the concept map are correct (i.e., 50–69% accuracy), but others are inaccurate or do not fully represent the connections between the concepts	Most of the relations in the concept map are inaccurate (i.e., more than 50% inaccuracy) or do not adequately represent the connections between the concepts, or there is no relation

References

1. Kroeze KA, Van Den Berg SM, Veldkamp BP, De Jong T (2021) Automated assessment of and feedback on concept maps during inquiry learning. IEEE Trans Learn Technol 14(4):460–473
2. Bainbridge WA, Hart JW, Kim ES, Scassellati B (2011) The benefits of interactions with physically present robots over video-displayed agents. Int J Soc Robot 3:41–52. https://doi.org/10.1007/s12369-010-0082-7
3. Barak M, Zadok Y (2009) Robotics projects and learning concepts in science, technology and problem solving. Int J Technol Des Educ 19(3):289–307. https://doi.org/10.1007/s10798-007-9043-3
4. Baxter P, Ashurst E, Read R, Kennedy J, Belpaeme T (2017) Robot education peers in a situated primary school study: personalisation promotes child learning. Plos One. https://doi.org/10.1371/journal.pone.0178126

5. Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F (2018) Social robots for education: a review. Sci Robot. https://doi.org/10.1126/scirobotics.aat5954
6. Breazeal C, Kidd CD, Thomaz AL, Hoffman G, Berlin M (2005) Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In 2005 IEEE/RSJ international conference on intelligent robots and systems. IEEE. pp 708–713
7. Carradore M (2022) People's attitudes towards the use of robots in the social services: a multilevel analysis using eurobarometer data. Int J Soc Robot 14(3):845–858. https://doi.org/10.1007/s12369-021-00831-4
8. Causo A, Vo GT, Chen IM, Yeo SH (2016) Design of robots used as education companion and tutor. Robot Mech 37:75–84. https://doi.org/10.1007/978-3-319-22368-1_8
9. Cejka E, Rogers C, Portsmore M (2006) Kindergarten robotics: Using robotics to motivate math, science, and engineering literacy in elementary school. Int J Eng Educ 22(4):711–722
10. Conde MA, Sanchez-Gonzalez L, Matellan-Olivera V, Rodriguez-Lera FJ (2017) Application of peer review techniques in engineering education. Int J Eng Educ 33(2):918–926
11. Costa S, Brunete A, Bae B-C, Mavridis N (2018) Emotional storytelling using virtual and robotic agents. Int J Humanoid Rob 15(03):1850006. https://doi.org/10.1142/s0219843618500068
12. de Jong T (2006) Scaffolds for scientific discovery learning. In: Elen J, Clark RE (eds) Dealing with complexity in learning environments. Elsevier Science Publishers, Netherlands, pp 107–128
13. de Jong T (2019) Moving towards engaged learning in STEM domains; there is no simple answer, but clearly a road ahead. J Comput Assist Learn 35(2):153–167. https://doi.org/10.1111/jcal.12337
14. de Jong T (2022) The guided inquiry learning principle in multimedia learning. In: Mayer RE, Fiorella L (eds) The Cambridge handbook of multimedia learning, 3rd edn. Cambridge University Press, Cambridge, pp 394–402
15. de Jong T, Gillet D, Rodríguez-Triana MJ, Hovardas T, Dikke D, Doran R, Dziabenko O, Koslowsky J, Korventausta M, Law E, Pedaste M, Tasiopoulou E, Vidal G, Zacharia Z (2021) Understanding teacher design practices for digital inquiry-based science learning: the case of Go-Lab. Educ Tech Res Dev 69:417–444. https://doi.org/10.1007/s11423-020-09904-z
16. Dmoshinskaia N, Gijlers H, de Jong T (2021) Giving feedback on peers' concept maps in an inquiry learning context: the effect of providing assessment criteria. J Sci Educ Technol 30(3):420–430. https://doi.org/10.1007/s10956-020-09884-y
17. Eysink THS, de Jong T (2012) Does instructional approach matter? How elaboration plays a crucial role in multimedia learning. J Learn Sci 21(4):583–625. https://doi.org/10.1080/10508406.2011.611776
18. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3–4):143–166. https://doi.org/10.1016/S0921-8890(02)00372-X
19. Frey N, Fisher D, Hattie J (2017) Surface, deep, and transfer? Considering the role of content literacy instructional strategies. J Adolesc Health 60(5):567–575. https://doi.org/10.1002/jaal.576
20. Hattie JAC, Donoghue GM (2016) Learning strategies: a synthesis and conceptual model. npj Sci Learn 1:16013. https://doi.org/10.1038/npjscilearn.2016.13
21. Horzum T (2018) The investigation of preservice mathematics teachers' knowledge about quadrilaterals through concept maps. Turk J Comput Math Educ 9(1):1–30. https://doi.org/10.17762/turcomat.v9i1.169
22. Ifenthaler D (2010) Bridging the gap between expert-novice differences. J Res Technol Educ 43(2):103–117. https://doi.org/10.1080/15391523.2010.10782564
23. Kanda T, and Ishiguro H (2005) Communication robots for elementary schools. In: Proceedings of the symposium on robot companions: hard problems and open challenges in robot-human interaction. AISB. pp 54–63
24. Kuang X, Eysink THS, de Jong T (2022) Effects of providing domain information on facilitating hypothesis generation in inquiry learning. J Educ Res 115:285–297. https://doi.org/10.1016/j.learninstruc.2018.01.011
25. Lachner A, Burkhart C, Nückles M (2017) Mind the gap! Automated concept map feedback supports students in writing cohesive explanations. J Exp Psychol Appl 23(1):29–46. https://doi.org/10.1037/xap0000111
26. Lalwani M, Bagmar N, Parikh S (2013) Efficient algorithm for auto correction using n-gram indexing. Intl J Comput Commun Technol (IJCCT) 4(2):108–112. https://doi.org/10.47893/IJCCT.2013.1177
27. Lazonder AW, Harmsen R (2016) Meta-analysis of inquiry-based learning: effects of guidance. Rev Educ Res 86(3):681–718. https://doi.org/10.3102/0034654315627366
28. Leyzberg D, Spaulding S, Toneva M and Scassellati B (2012) The physical presence of a robot tutor increases cognitive learning gains. In: Proceedings of the annual meeting of the cognitive science society
29. Mehta A, Salgond V, Satra D, Sharma N (2021) Spell correction and suggestion using Levenshtein distance. Int Res J Eng Technol 8(8):1977–1981
30. OECD (2016) PISA 2015 results (volume II): policies and practices for successful schools. OECD Publishing, Francia. https://doi.org/10.1787/9789264267510-en
31. OECD (2023) PISA 2022 results (Volume I): the state of learning and equity in education. OECD Publishing, Francia. https://doi.org/10.1787/53f23881-en
32. Özkan SB, Toz M (2022) A review study on the investigation of the effects of using robots in education on metacognitive behaviors. Comput Appl Eng Educ 30(4):1277–1288
33. Pan Y, Steed A (2016) A comparison of avatar-, video-, and robot-mediated interaction on users' trust in expertise. Front Robot AI. https://doi.org/10.3389/frobt.2016.00012
34. Pedaste M, Mäeots M, Siiman LA, de Jong T, van Riesen SAN, Kamp ET, Manoli CC, Zacharia ZC, Tsourlidaki E (2015) Phases of inquiry-based learning: definitions and inquiry cycle. Educ Res Rev 14:47–61. https://doi.org/10.1016/j.edurev.2015.02.003
35. Ramani S, Krackov SK (2012) Twelve tips for giving feedback effectively in the clinical environment. Med Teach 34(10):787–791. https://doi.org/10.3109/0142159x.2012.684916
36. Roessger KM, Daley BJ, Hafez DA (2018) Effects of teaching concept mapping using practice, feedback, and relational framing. Learn Instr 54:11–21. https://doi.org/10.1016/j.learninstruc.2018.01.011
37. Sisman B, Gunay D, Kucuk S (2019) Development and validation of an Educational Robot Attitude Scale (ERAS) for secondary school students. Interact Learn Environ 27(3):377–388. https://doi.org/10.1080/10494820.2018.1474234
38. Strautmane M (2012) Concept map-based knowledge assessment tasks and their scoring criteria: an overview. Concept maps: theory, methodology, technology. In: Proceedings of the fifth international conference on concept mapping
39. Szafir D and Mutlu B (2012) Pay attention! Designing adaptive agents that monitor and improve user engagement. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp 11–20. https://doi.org/10.1145/2207676.2207679
40. Tabachnick BG, Fidell LS (2013) Using multivariate statistics, 6th edn. Pearson, London
41. van den Berghe R, Petersen H, Hellendoorn A, and van Keulen H (2020) Programming a robot or an avatar: a study on learning outcomes, motivation, and cooperation. In: 2020 ACM/IEEE international conference on human-robot interaction, Cambridge, United Kingdom. https://doi.org/10.1145/3371382.3378239

42. Velentza A-M, Fachantidis N, Lefkos I (2021) Learn with surprize from a robot professor. Comput Educ 173:104272
43. Zacharia ZC, Manoli C, Xenofontos N, de Jong T, Pedaste M, van Riesen SAN, Kamp E, Mäeots M, Siiman LA, Tsourlidaki E (2015) Identifying potential types of guidance for supporting student inquiry in using virtual and remote labs: a literature review. Educ Tech Res Dev 63:257–302. https://doi.org/10.1007/s11423-015-9370-0

Burak Sisman is a researcher in the Department of Cognitive Robotics at Delft University of Technology. His research focuses on educational robotics, STEM education, human-robot interaction, knowledge representation, and artificial intelligence. He has over 18 years of experience in teaching and research. He has conducted various research projects in these fields and has received several patents for his innovative work.

Johannes Steinrücke is Assistant Professor Data Science at the Cognition, Data and Education chair at the University of Twente. With his background in psychology and methodology and statistics, his research interests lies in psychometrics, educational measurement and AI in education. Currently he is working on various projects and initiatives surrounding item-response theory, psychometrics and machine learning, and AI in education. His previous research included the development of unobtrusive assessment algorithms to individualize adaptations in digital serious games.

Ton de Jong is emeritus professor of Instructional Technology at the University of Twente. He specializes in inquiry learning (mainly in science domains) supported by technology. He was coordinator of eight EU projects including the 7th framework Go-Lab project on learning with online laboratories in science and its H2020 follow-up project Next-Lab. He has published three papers in Science. He is AERA and ISLS fellow and is elected member of the Academia Europaea. In 2023 he received the EARLI oeuvre award.