When Positive Perception of the Robot Has No Effect on Learning

Jauwairia Nasir*[1], Utku Norman*[1], Barbara Bruno[1,2], and Pierre Dillenbourg[1]

***Abstract*— Humanoid robots, with a focus on personalised social behaviours, are increasingly being deployed in educational settings to support learning. However, crafting pedagogical HRI designs and robot interventions that have a real, positive impact on participants' learning, as well as effectively measuring such impact, is still an open challenge. As a first effort in tackling the issue, in this paper we propose a novel robot-mediated, collaborative problem solving activity for school-children, called JUSThink, aiming at improving their computational thinking skills. JUSThink will serve as a baseline and reference for investigating how the robot's behaviour can influence the engagement of the children with the activity, as well as their collaboration and mutual understanding while working on it. To this end, this first iteration aims at investigating (i) participants' engagement with the activity (Intrinsic Motivation Inventory—IMI), their mutual understanding (IMI-like) and perception of the robot (Godspeed Questionnaire); (ii) participants' performance during the activity, using several performance and learning metrics. We carried out an extensive user-study in two international schools in Switzerland, in which around 100 children participated in pairs in one-hour long interactions with the activity. Surprisingly, we observe that while a teams' performance significantly affects how team members evaluate their competence, mutual understanding and task engagement, it does not affect their perception of the robot and its helpfulness, a fact which highlights the need for baseline studies and multi-dimensional evaluation metrics when assessing the impact of robots in educational activities.**

***Keywords*—educational robotics; collaborative problem solving; computational thinking; engagement; mutual modelling; robot perception, human-robot interaction.**

I. Introduction

"Computational thinking (CT) is going to be needed everywhere. And doing it well is going to be a key to success in almost all future careers." The words of Stephen Wolfram[3] capture the urgency seen in the efforts to introduce CT in educational curricula before high school [1]. At the same time, the potential of robots is increasingly being explored in educational settings across the globe, under the intuition that robots could be an effective tool for advancing CT skills [2], as well as for increasing participants' engagement with the educational activity [3] and collaboration [4], [5]. However, crafting pedagogical designs and robot interventions that truly succeed in achieving such objectives is a challenging and to-date open question.

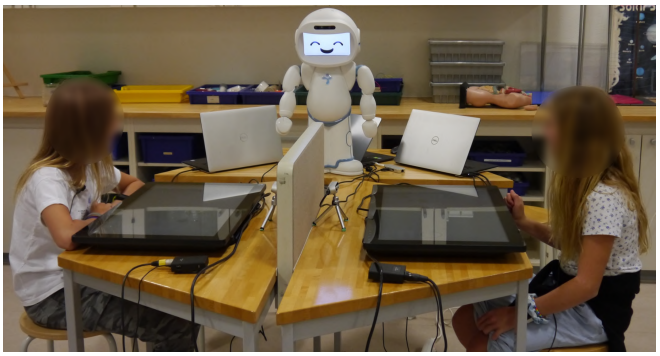


Fig. 1: QTrobot welcomes children to the JUSThink activity.

Inspired by this challenge, the JUSThink project[4] (see Fig. 1) aims to (i) improve the computational thinking skills of children by exercising their abstract reasoning with and through graphs (posed as a way to represent, reason about and solve a problem), (ii) promote collaboration between participants, by providing team members with different, complementary information at all times during the activity, (iii) serve as a platform for the design and evaluation of robot behaviours aiming to ultimately improve learning, by improving participants' engagement with the task as well as collaboration and mutual understanding between them [6].

From a research perspective, and in line with the objectives outlined above, the designed robot-mediated activity is also aiming to surface cues relevant to (i) participants' engagement with the task at hand, their partner and the robot, (ii) mutual understanding and misunderstandings between the participants.

The contribution of the paper is twofold:

1) design for a first version of a robot-mediated human-human collaborative learning activity in which the robot is intended to intervene only when required by the activity's pedagogical goals, without causing unnecessary distractions;
2) an analysis of participants' self-assessment of engagement, mutual understanding, and perception of the robot, both separately and in connection with performance and learning in the collaborative activity.

The second contribution serves a double purpose. On the one hand, it is meant as a baseline reference for future studies on the impact that robot behaviours have on participants'

*The first two authors contributed equally to this work.

[1]Computer-Human Interaction in Learning and Instruction (CHILI) Lab, Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland

[2]MOBOTS group within the Biorobotics Laboratory (BioRob), Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955.

[3]https://blog.stephenwolfram.com/2016/09/how-to-teach-computational-thinking/

[4]https://www.epfl.ch/labs/chili/index-html/research/animatas/justhink/

learning, performance, engagement, collaboration and mutual understanding. This is the reason why the robot's behaviour in this version is purposefully designed to be minimal and detached from the participants' situation. On the other hand, participants' assessment of a "useless robot", especially if they are struggling with the task at hand, is an interesting insight into the appropriateness of commonly adopted tools for robot evaluation in educational settings. For this reason, in our analysis we complement standard HRI questionnaires with learning and performance metrics.

Concretely, in this work we address the following research questions:

RQ1: How do participants assess their engagement, mutual understanding and perception of the robot, for the proposed JUSThink activity?

RQ2: Is the JUSThink activity effective in its pedagogical objective (i.e., does a high performance in the task also lead to a high learning gain)?

RQ3: Is there a correlation between the performance in the task, or the learning gain, and participants' self-assessment of engagement, mutual understanding, competence, stress and, above all, the robot's behaviour and its helpfulness?

From the above research questions, we derive the corresponding, following hypotheses:

H1: H1(a): Participants' self-assessment of engagement and mutual understanding lies more on the positive side of the spectrum than on the negative one.
H1(b): Their self-assessment of the robot lies more on the negative side of the spectrum than on the positive one, because of its few and limited interventions.

H2: Performance in the learning task correlates with learning gain.

H3: H3(a): Teams with high performance will rate their engagement, mutual understanding, self-competence higher than teams with low performance, and will have a more positive perception of the robot.
H3(b): Teams with low performance will rate their stress higher than teams with high performance, and will have a more negative perception of the robot and its helpfulness.

The article is organised as follows. Section II briefly examines the state of the art in robot-mediated educational activities, with a specific focus on collaborative activities, while Section III describes in detail the proposed JUSThink activity. Section IV describes the user study conducted to evaluate the activity and the robot's role in it, while the results are presented in Section V. Conclusions follow.

II. Related Work

A. Robot Interventions in Educational Settings

The principle is simple: when embedded in an educational activity, the robot must choose, based on its perception of the situation, an action that is in line with enhancing the activity's educational goals.

The implementation of this principle in practice, however, is not so straightforward. In a recent review, Belpaeme et al. point out that even experienced human instructors struggle to make the best choice and debate whether the same educational theories that apply in human-human settings hold for robot interventions [3].

As reported in the review, robot behaviours that have been found to have a positive effect on participants' learning gain and recall include "choosing an appropriate emotional support strategy based on the affective state of the child [7], assisting with a meta-cognitive learning strategy [8], deciding when to take a break [9], appropriate gestures [10], appropriate and congruent gaze behaviour [11], expressive behaviours and attention-guiding behaviours [12], timely nonverbal behaviours [13]".

However, Belpaeme et al. also warn that "merely increasing the amount of social behaviour for a robot does not lead to increased learning gains: certain studies have found that social behaviour may be distracting [14], [15]".

In line with the above considerations, the overarching goal of the JUSThink project is to design and evaluate robot behaviours which have a positive effect on participants' learning while specifically addressing their engagement and mutual understanding. The goal of the baseline version presented in the article is to endow the robot with basic behaviour and evaluate its perception by the participants, specifically in relation with their performance.

B. Activity Design for Collaborative Learning

Collaborative learning describes a situation in which particular forms of interaction among people are expected to occur, which would trigger learning mechanisms, but there is no guarantee that the expected interactions will actually occur [16]. Robots have been incorporated in collaborative learning activities to support the interaction in various ways. For instance, a robot equipped with emphatic competencies was used to support the interactions of a collaborative learning activity about sustainable development through constructing a sustainable city in a group setting by considering the affective states of the participants [17]. Within a learning-by-teaching paradigm [18], robots were used to promote children's responsibility in a collaborative learning activity in which children write on a tablet and robot gives corrective feedback on it [19], to aid the reading of children where a child and a robot collaboratively read stories [15], and to be tutored by children collaboratively in order to improve handwriting [5].

Constructivism entails actively building knowledge rather than passively receiving it [20], which means that the participants explore openly without directly receiving guidance. In collaborative designs where no direct feedback is given, collaborative learning can then be seen as a special case of constructivism where the participants have to achieve a shared goal through exploration, reflection, mutual regulation [21], conflict resolution [22], [23], and explanation of their decisions.

TABLE I: Pipeline of the JUSThink activity.

Stage	What are the participants supposed to do?	What does the robot do?	Level	Duration
Welcome	Enter their name, age and gender on the screen	Welcome the participants, ask them for personal details	individual	2 min
Introduction	Listen to the robot	Introduce the task goal: connecting the gold mines by spending as little money as possible	team	2 min
Pre-test	Answer a list of multiple-choice questions on the screen	Ask the participants to answer the pre-test questions	individual	$\leq$ 10 min
Demo	Listen to the robot and follow the illustrations on the screen	Explain the two game views and their functionalities	team	3 min
Learning Task	Find a cheapest railway network (a minimum spanning tree) connecting all gold mines by: i) drawing or erasing tracks that connect pairs of gold mines ii) submitting any agreed-upon solution to the robot for evaluation and feedback	At the submission of a solution: If the submitted solution is optimal, congratulate the participants and move to the post-test stage. Otherwise, reveal the cost difference between the submitted solution and an optimal one and motivate the participants to try harder. Point out the availability of the history of submitted solutions if the participants are not successful after several attempts.	team	$\leq$ 25 min
Post-test	Answer a list of multiple-choice questions on the screen	Ask the participants to answer the post-test questions	individual	$\leq$ 10 min
Questionnaire	Rate on a 5-point Likert scale a set of items about engagement, mutual understanding and the robot	Ask the participants to answer the questionnaire questions	individual	$\leq$ 5 min
Goodbye	See the robot wave goodbye	Thank the participants for their help, say goodbye	team	$\leq$ 1 min

A careful activity design is needed to maximise the chances for the learning mechanisms to occur. Our design enforces, through specific design choices, collaboration between the team members while also leaving space for exploration: thus, the participants are expected to have productive interactions [24] while contributing to a solution together. We expand on the design choices in the upcoming section.

III. Activity Design

The JUSThink activity is organised in a sequence of stages as described in Table I, the core of which is the learning task.

A. Learning Task Design

1) Swiss Gold Mines Scenario: The objective of the JUSThink activity is to give participants an intuitive knowledge about minimum-spanning-tree problems[5] and how to solve them. To introduce the minimum-spanning-tree problem to the participants as a game and with minimal terminology, we created a scenario based on a map of Switzerland. On the map, gold mines are depicted with mountains, animated with glittering gold on them, and labelled with names of Swiss cities (e.g. "Mount Zermatt" and "Mount Zurich"): these make up the nodes V of the graph $G = (V, E)$.

At the start of the Learning Task stage, the robot, acting as the CEO of a gold mining company, reiterates the problem by asking the participants to help it collect the gold by connecting the gold mines with railway tracks, while spending as little money as possible. Then, during the Learning Task stage, the participants collaboratively construct a solution by drawing and erasing tracks that connect pairs of gold mines, and submit it to the robot for evaluation (one of the two optimal solutions is shown in Fig. 2). The cost function and the graph layout draw inspiration from the *muddy city* problem[6]. Note that the cost function is strictly positive.

2) Scaffolding for Collaboration and Abstract Reasoning: A number of design choices have been made to promote collaboration between the participants.

Firstly, the screens display two different views that present only *partially observable information* to the participants, with a barrier preventing each participant from seeing the other's screen (see Fig. 1). At every point in time within the task, one of the participants is shown the *figurative view* (see Fig. 2a) and the other is shown the *abstract view* (see Fig. 2b). In the figurative view, nodes are shown as mountains and edges as railway tracks connecting two gold mines. Edges' costs are not visible. In the abstract view, nodes are shown as labelled circles, drawn edges as solid lines, while edges drawn and then deleted appear as dashed lines, superimposed over the figurative drawing as a semitransparent overlay. The costs of edges (solid or dashed) are indicated as a number near their center point.

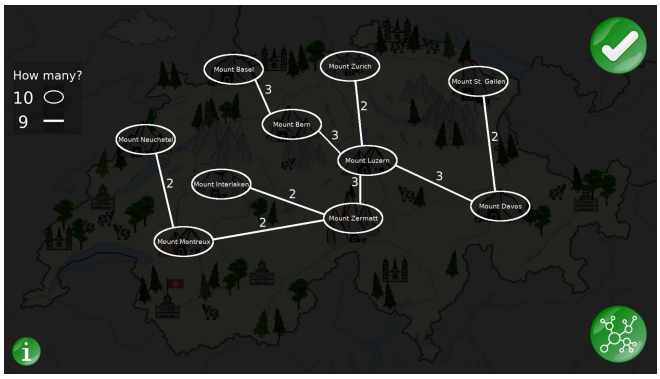
Secondly, the views offer *complementary functionality*, allowing different actions for constructing a solution. In the figurative view, one can edit the graph by drawing a track or erasing an existing track. In the abstract view, one can see the cost of the tracks, access the previous solutions and their costs, and bring back a previous solution after discarding the current selection. A track is explored after drawing it for the first time, and its cost is displayed in the abstract view until the constructed solution is submitted. Hence, in order to make an informed decision on which action to take (add/delete edges, submit), the participants need to communicate their understanding of what the best move would be based on the information available to them.

Thirdly, every two edits, the views of the participants are swapped, i.e. the participant in the figurative view then is in the abstract view and vice versa. This is so that there are no permanent roles, and that the participants could participate equally in the thought process associated with each view.

[5] Let $G = (V, E)$ denote a connected, undirected, edge-weighted graph. V is the set of nodes, $E \subseteq V \times V$ is the set of edges that connects node pairs, and $c : E \to \mathbb{R}$ is the edge cost function for G. A subgraph of G is said to "span" the graph G if it connects all nodes of G, i.e. each node is reachable from every other. The problem is to find a subgraph T of G that spans G and minimises $o_T(T) = \sum_{e \in E_T} c(e)$. An optimal solution T is called a minimum spanning tree for G.

[6] https://csunplugged.org/minimal-spanning-trees/

(a) figurative view



(b) abstract view

Fig. 2: The contents of the screens of the participants, where one participant is in the figurative view and the other participant is in the abstract view. The shown set of tracks forms a minimum spanning tree for the network of gold mines to be constructed together by the participants.

Fig. 3: Various robot behaviours during the activity.

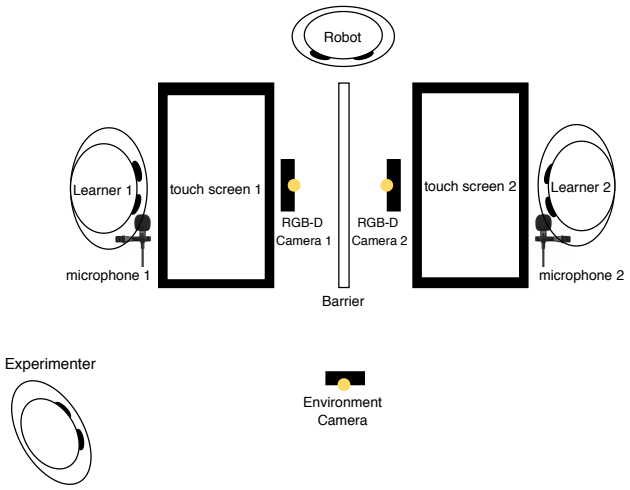


Fig. 4: The layout of the hardware setup for JUSThink.

Fourthly, the participants can submit only if their solution spans the whole graph. The participants can submit as many times as they want, until they find an optimal solution or the allocated time is over. This allows the participants to experiment with different solutions. The participants are informed of the remaining time only a few minutes before the allocated time is over.

Fifthly, the cost of each track is initially hidden and revealed only after it is drawn. This could promote reasoning about an edge in terms of a connection between two entities with an associated cost.

Lastly, in order to submit a solution, both participants have to select submit (for the same solution) by clicking the submit button on their respective screens. A selection for submission is revoked by an edit on the solution. Thus, the participants need to agree on a solution.

B. The Robot's Role

The robot's role in JUSThink is twofold: (i) mediate and automate the entire interaction (see Table I), pausing the participants' applications, giving instructions, and moving from a stage to the next upon its completion; as well as (ii) intervene at sparse moments to give feedback on the progress, provide basic hints (as mentioned in Table I) and support through minimal expressive behaviours. The expressive behaviours include verbal support, using participants' names, and the display of emotions and supporting gestures. Some of these behaviours can be seen in Fig. 3.

C. Setup Design

1) Hardware Setup: The hardware layout required for the JUSThink activity is shown in Fig. 4. Two children are sitting across each other, separated by a barrier. In front of each child, a touch screen is placed horizontally. The humanoid robot (QTrobot) is placed on the side, visible by both children. The children can see each other but not their partner's screen. The experimenter is at all times in the room, ready to intervene. The interaction is recorded by three cameras: one environment camera filming the whole scene and two RGB-D cameras each focused on a child's face. Audio is recorded with two lavalier microphones, clipped on children's shirts. Two computers, connected to the two touchscreens and to the robot's local network, manage the activity and the synchronous recording of the cameras and microphones. The face cameras are connected to a third computer to alleviate the burden of bandwidth in the local network.

TABLE II: Categorisation of the questions in the questionnaire.

No	Question	Group	Category
1	I was trying very hard to find the best solution.	Cognitive at Task Level (IMI)	Task Engagement
2	It was important to me to do well at this task.		
3	I thought this activity was quite enjoyable.	Affective at Task Level (IMI)	
4	I enjoyed trying to find the best solution.		
5	I was trying very hard while discussing with my friend about the activity.	Cognitive at Social Level (IMI)	Social Engagement
6	It was important for me to discuss with my friend while finding the best solution.		
7	Discussions with my friend were quite interesting.	Affective at Social Level (IMI)	
8	I enjoyed discussing with my friend about the activity.		
9	I think I did pretty well at this activity.	Perceived Competence (IMI)	Own Competence
10	I am satisfied with my performance at this task.		
11	I felt tense while doing this activity.	Pressure/Tension (IMI)	Stress
12	I think my friend understood my instructions very well.	Cognitive (IMI-like)	Mutual Understanding
13	I think my friend understood my emotions very well.	Affective (IMI-like)	
14	I think the robot is competent (capable).	Robot (Godspeed)	Robot
15	I think the robot is intelligent.		
16	I think the robot is friendly.		
17	I think the robot is likeable.		
18	I think the robot is distracting.	Robot (Godspeed-like)	Robot Behaviour
19	I think the robot should give more useful feedback.		
20	I liked the robot.		
21	I would like to play the same game with the same friend.	Game and Friend	
22	I would like to play the same game with another friend.		
23	I knew my friend well.	Known Friend	
24	How many minutes do you think you spent on the part where you played with your friend to find the best solution?	Perception of Time	

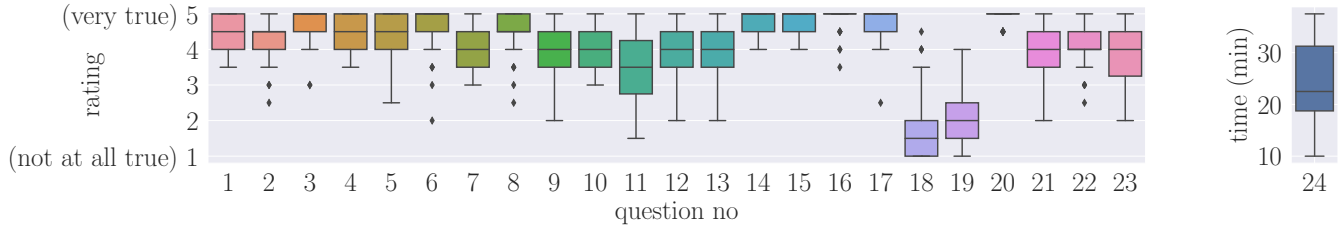


Fig. 5: Box plots showing the distribution of the ratings given in the questionnaire ($N = 39$ teams) for each question. The questions are listed in Table II.

2) Software Setup: Each participant interacts with an instance of the JUSThink participant application that is written in Python and uses pyglet as the windowing and multimedia library. Hence, a separate instance of the application is run for each participant in a team. The JUSThink robot behaviour application is also developed in Python and governs what the robot does and when. The applications communicate via the Robot Operating System (ROS).

IV. User Study

A. Evaluation Metrics

1) Learning Metrics: We generate our learning metrics from the scores of the pre-test and post-test, which are defined in a context other than Swiss gold mines and based on variants of the graphics in the *muddy city*[6] problem.

Specifically, pre-test and post-test are composed of 10 multiple-choice questions, assessing the following concepts:

C1: (exists-or-not, 3 questions). If a spanning tree exists, i.e. if the graph is connected. Example question: "In which map can a postman visit *all the houses* using only the roads?"

C2: (spans-or-not, 3 questions). If the given subgraph spans the graph. Example question: "In which map can a postman visit *all the houses* using *only the black roads*?"

C3: (minimum-or-not, 4 questions). If the given subgraph that spans the graph has a minimum cost. Example question: "In which map can the city build another path with *fewer* stones than the *black path shown* to visit all the houses?".

In C2 and C3, the black path illustrates the given subgraph. The questions are given here in verbatim, where the emphases (here in *italics*) are presented to the participants in uppercase. The post-test is obtained by randomly shuffling the questions and the response choices within and across the questions for the same concept, as well as mirroring the images given in the response choices vertically.

From pre- and post-test, we define two learning metrics:

(i) **absolute learning gain**, i.e. the difference between a participant's post-test and pre-test score, divided by the maximum score that can be achieved (10), which grasps how much the participant learned of all the knowledge available,

(ii) **relative learning gain**, i.e. the difference between a

participant's post-test and pre-test score, divided by the difference between the maximum score that can be achieved and the pre-test score, which grasps how much the participant learned of the knowledge that he/she didn't possess before the activity.

In the analysis, the absolute learning gains of two team members are averaged, to provide a measure of the team's absolute learning gain. The same procedure is used to obtain a team's relative learning gain.

2) Performance Metrics: Let error be the difference between the cost of a submitted solution and the cost of an optimal/correct solution (optimal cost), normalised by the optimal cost. Then, we define two metrics to measure the task performance as follows:

(i) **last error**, i.e. error of the last submitted solution. Note that if a team has found an optimal solution (error = 0) the game stops, therefore making last error = 0.

(ii) **minimum error**, i.e. the minimum of the error values, considering all submitted solutions. This metric is interesting since the last submission does not necessarily reflect the best solution of a team, in case they have not found an optimal solution.

3) Questionnaire: The questionnaire consists of 24 questions as reported in Table II. Among them, 11 belong to the Intrinsic Motivation Inventory (IMI) [25], which "is a multidimensional measurement device intended to assess participants' subjective experience related to a target activity in laboratory experiments" and relate to engagement, own competence and stress, 2 refer to mutual understanding and are completed by 3 other questions on the relationship with the team partner (21-23), 4 belong to the Godspeed questionnaire [26], one of the most widely used in HRI, and refer to the perception of the robot, which we complement with 3 additional questions on the robot's behaviour and its helpfulness. Question 24 is on the perception of time elapsed.

Items concerning the perception of the robot refer to competence, intelligence, friendliness and likeability and are complemented by behavioural items on being distracting and giving useful feedback. Engagement here entails the effort put in for solving the task (cognitive engagement at task level) as well as for discussions with the partner to solve the given problem (cognitive engagement at social level). It also includes the enjoyment that the participants had with regards to the task (affective engagement at task level) as well as when discussing with their partner (affective engagement at social level). Similarly, mutual understanding was also measured both in terms of understanding of their instructions to each other for solving the task (cognitive) and understanding of each others' emotions (affective).

With respect to the Research Questions driving this study, the questionnaire by itself is meant to investigate RQ1, the learning and performance metrics allow for investigating RQ2, while all metrics together are used to investigate RQ3.

B. Participants

The pilot study was conducted with 96 children aged 9 to 12 years. Due to technical issues during the experiment, 18 participants are omitted from the analysis, resulting in a dataset of 78 children (41 females: $M = 10.3, SD = 0.75$; 37 males: $M = 10.4, SD = 0.60$). The experiment took place over the span of two weeks in two international schools in Switzerland and the participants participated in teams of two, in a session lasting approx. 50 minutes. The activity pipeline is summarised in Table I. There were always two experimenters available in the room but the system was fully automated to require the least intervention by the experimenters. While the participants were generally familiar with robots as a part of their curriculum and STEM activities, they did not have a prior experience with the robot platform used in this study which could introduce some novelty effect; however, that is a well-known HRI problem.

V. Analysis and Discussion

It is to be noted that the questionnaire and tests were done individually ($N = 78$ participants); however, in this section, we report values as a team average ($N = 39$ teams).

A. RQ1: On Participants' Self-assessment

In Fig. 5, we see the distribution of the team-averaged ratings for all questions in the questionnaire.

1) Engagement and Mutual Understanding: Participants rated themselves to be engaged highly at both task and social level (mean(1-8)= 4.43). Similar to engagement, the participants rated the understanding of their instructions and emotions by their partners as very high (12, 13). These results support H1(a).

2) Perception of the Robot: Despite the robot having a basic role in the current setup, the participants rate it very highly with regards to competence, intelligence, friendliness, and likeability (mean(14-17)= 4.78)—see 14-17 in Fig. 5. It must be noted that the interaction lasted 45 to 50 minutes; hence giving ample time for the participants to form an opinion on the characteristics of the robot (and their limitations). Also, despite a high number of teams not being successful in finding an optimal solution, we see that the majority of them think that the robot does not need to give more useful feedback (19). Furthermore, very few participants found the robot's behaviour as distracting (18). Hence, contrary to our expectations, H1(b) is rejected.

B. RQ2: On the Relation Between Performance and Learning Gain

We observe a spectrum of gains from negative to positive for the two learning gains described in Sec. IV-A.1. Fig. 6 shows the distribution of the performance and learning metrics. Specifically, 8 out of 39 teams have found an optimal solution.

To have an in depth view, we calculated Spearman's correlation between each pair of performance and learning metrics; however we did not find any significant correlations. In Fig. 7, we plot all the teams to see how they are scattered in the 2D space spanned by the last error and the relative learning gain. In line with Spearman's correlation results ($r_s = -0.08, p = 0.627$), we observe that the relative

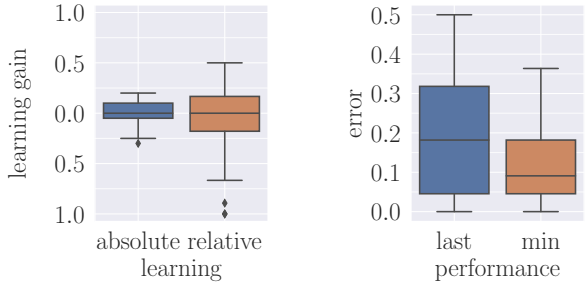


Fig. 6: Distribution of learning and performance metrics.

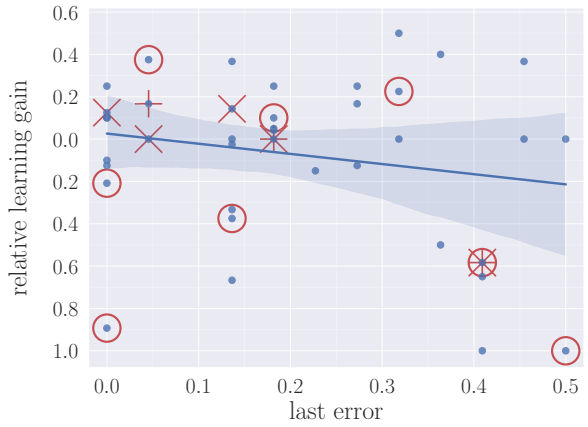


Fig. 7: Relative learning gain vs. last error plot for the teams ($N = 39$ teams). We denote the teams that felt stressed (with team average rating ≥ 4) by a circle 'O', those that said the robot was distracting (rating of 4 or above) by a cross 'X', and those that believed the robot should give more useful feedback (rating of 4 or above) by a plus '+', as rated for questions 11, 18 and 19, respectively. The line represents the linear regression line with a 95% confidence interval.

gain that a participant achieves is not proportional to their success in the game, which is an important observation for when providing interventions by the robot. In conclusion, participants who appear to be performing well (left side of Fig. 7) may not always be developing an understanding of the task, a finding which does not support hypothesis H2.

Lastly, a possible explanation for the observed low learning gains, as well as the lack of a relation between performance and learning gain, is that our pre- and post-tests rely on a high transfer between the task and the test, which is not spontaneous.

C. RQ3: On the Impact of Performance and Learning Gain on Participants' Self- and Robot assessment

In this section, we observe if performance or learning gain are related with participants' self-assessment on engagement, mutual understanding, perception of the robot, self-competence, stress, and especially the need for the robot to give more feedback or its assessment as a distraction. Spearman's correlation reports three medium correlations that are significant between last error and competence ($r_s = -0.369, p = .02$), minimum error and competence ($r_s = -0.417, p = .008$), and minimum error and mutual understanding ($r_s = -0.336, p = .03$). This indicates that 1) the higher the last error or the minimum error, the lower would the participants rate their self competence, and 2) the higher the minimum error, the lower mutual understanding would be rated.

It is important to note here that there were no significant correlations found between self-assessment metrics and the two learning gains: participants seem to have based their assessment of self-competence on apparent representations of learning and achievement, e.g. success-failure in the game, rather than the tests which are used to measure learning. A similar result was reported in [27] where the authors observed that "subjects who experienced success made significantly greater gains in positive self-assessments, and failure subjects made significantly greater gains in negative self-assessments". Note that the participants did not receive feedback on their scores in the tests.

We then performed a Kruskal-Wallis test to inquire if teams belonging to high and low performance groups report differently on the aforementioned questions. In line with our hypothesis H3(a), high performing teams (in terms of last error) rated their task engagement significantly higher than those who did not perform as well ($H = 5.669, p = .017$, Cohen's $d = 1.11$). Conversely, their perception of the robot is higher than that of low performing teams, but the result is not significant ($H = 2.785, p = .095$, Cohen's $d = 0.68$). For this reason, we deem H3(a) to be only partially supported by our findings, and specifically to be rejected concerning the perception of the robot.

Concerning H3(b), we see that there is no significant result neither with Spearman's correlation nor with Kruskal-Wallis test, meaning that low performance does not make the participants rate their stress higher, have a more negative opinion of the robot or, interestingly, wish the robot could give more useful feedback. Indeed, as shown in Fig. 7, teams that reported high levels of stress, the robot being distracting, or wished for more useful feedback are dispersed throughout the plot, regardless of their performance. As the figure shows actually most of the teams that perceived the robot to be distracting or wished for more useful feedback (marked in the figure by a cross and a plus sign, respectively) lie more on the top-left area of the plot, which denotes high learning and high performance (low error).

VI. Conclusions And Future Work

In this paper, we propose a novel robot-mediated collaborative educational activity that is evaluated in a user study involving 78 children aged 9-12. The user study aims at assessing various performance and learning metrics, alongside task and social engagement, mutual understanding between partners, self-perception of competence, stress, robot and robot behaviour. We report three key findings: 1) performance and learning are not correlated, and also do not correlate similarly with other metrics; 2) while affecting how a participant perceives their own competence, task

engagement, and mutual understanding with their partner, performance has no significant effect on the perception of the robot. Moreover, low performance has no correlation with wishing the robot to give more useful feedback; 3) despite its rudimentary behaviour, participants perceive the robot as highly competent, intelligent, friendly, likeable, not distracting, and report not feeling a need for more feedback from the robot.

Such findings allow for drawing conclusions which, albeit being far from definitive, provide insights for robot-mediated pedagogical activity design. Specifically: 1) the lack of correlation between learning and performance metrics highlights the importance of moving away from robot interventions that affect (and refer to) only superficial measures of students' learning, e.g. performance, rather focusing on behavioural patterns that more solidly indicate whether participants would end up learning or not; 2) the fact that the performance, low or high, did not have any effect on the perceived usefulness of the robot by the participants highlights the need for well-crafted domain specific metrics to truly assess the effectiveness of the robot and complement the general information provided by standard evaluation tools.

While the results are limited to the specific robot-mediated collaborative activity introduced in the paper; however, the conclusions drawn from them can be extended to other educational settings in highlighting the need for similar baseline studies and multi-dimensional evaluation metrics when assessing the impact of various robot strategies. As part of current and future effort along this research line, we are exploring and modelling the behavioural patterns that are indicative of higher understanding of the learning goal, and hence are indicative of engagement and mutual understanding which are beneficial in HRI educational settings. Furthermore, we plan to design and test various robot strategies that intervene based on such models in various roles and settings; for example, in the role of a mediator in a similar setting or as a peer playing the game directly with one child.

ACKNOWLEDGEMENTS

We would like to acknowledge the Swiss schools that helped us to make this study possible by giving us their valuable time and resources. We would also like to thank Gilles Raimond, Hala Khodr, Kevin Gonyop, and Thibault Asselborn for their help during the experiments. We are grateful to the SNSF for supporting this project through the National Centre of Competence in Research Robotics.

REFERENCES

[1] D. Menon, S. Bp, M. Romero, and T. Viéville, "Going beyond digital literacy to develop computational thinking in K-12 education,," in *Smart Pedagogy of Digital Learning*, L. Daniela, Ed. Taylor&Francis (Routledge), 2019.

[2] C. Chalmers, "Robotics and computational thinking in primary school," *IJCCI*, vol. 17, pp. 93–100, 2018.

[3] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science Robotics*, vol. 3, no. 21, p. eaat5954, 2018.

[4] A. Ioannou and E. Makridou, "Exploring the potentials of educational robotics in the development of computational thinking: A summary of current research and practical proposal for future work," *Education and Information Technologies*, vol. 23, no. 6, pp. 2531–2544, 2018.

[5] L. E. Hamamsy, W. Johal, T. Asselborn, J. Nasir, and P. Dillenbourg, "Learning by collaborative teaching: An engaging multi-party CoWriter activity," in *RO-MAN 2019*, 2019, pp. 1–8.

[6] J. Nasir, U. Norman, B. Bruno, and P. Dillenbourg, "You tell, I do, and we swap until we connect all the gold mines!" *ERCIM News*, vol. 2020, no. 120, 2020.

[7] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Empathic robots for long-term interaction," *IJSR*, vol. 6, no. 3, pp. 329–341, 2014.

[8] A. Ramachandran, C.-M. Huang, E. Gartland, and B. Scassellati, "Thinking aloud with a tutoring robot to enhance learning," in *HRI '18*, Feb. 2018, pp. 59–68.

[9] A. Ramachandran, C.-M. Huang, and B. Scassellati, "Give Me a Break!: Personalized Timing Strategies to Promote Learning in Robot-Child Tutoring," in *HRI '17*, Mar. 2017, pp. 146–155.

[10] C.-M. Huang and B. Mutlu, "Modeling and evaluating narrative gestures for humanlike robots," in *RSS 2013*, 2013, pp. 57–64.

[11] ——, "The repertoire of robot behavior: Designing social behaviors to support human-robot joint activity," *JHRI*, pp. 80–102, June 2013.

[12] M. Saerbeck, T. Schut, C. Bartneck, and M. Janse, "Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor," in *CHI '10*, 2010, pp. 1613–1622.

[13] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme, "Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions," in *Social Robotics*, A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, Eds. Springer, 2015, pp. 327–336.

[14] J. Kennedy, P. Baxter, and T. Belpaeme, "The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning," in *HRI '15*, 2015, pp. 67–74.

[15] E. Yadollahi, W. Johal, A. Paiva, and P. Dillenbourg, "When deictic gestures in a robot can harm child-robot collaboration," in *IDC '18*, 2018, pp. 195–206.

[16] P. Dillenbourg, M. Baker, A. Blaye, and C. O'Malley, "The evolution of research on collaborative learning," in *Learning in Humans and Machines: Towards an Interdisciplinary Learning Science*, H. Spada and P. Reimann, Eds. Oxford, Elsevier, 1996, pp. 189–211.

[17] P. Alves-Oliveira, P. Sequeira, F. S. Melo, G. Castellano, and A. Paiva, "Empathic robot for group learning: A field study," *ACM THRI*, vol. 8, no. 1, Mar. 2019.

[18] C. C. Chase, D. B. Chin, M. A. Oppezzo, and D. L. Schwartz, "Teachable agents and the protege effect: Increasing the effort towards learning," *J Sci Educ Tech*, vol. 18, no. 4, pp. 334–352, 2009.

[19] S. Chandra, P. Alves-Oliveira, S. Lemaignan, P. Sequeira, A. Paiva, and P. Dillenbourg, "Can a child feel responsible for another in the presence of a robot in a collaborative learning activity?" in *RO-MAN 2015*, 2015, pp. 167–172.

[20] E. Von Glasersfeld, "Cognition, construction of knowledge, and teaching," in *Constructivism in Science Education: A Philosophical Examination*, M. R. Matthews, Ed. Springer, 1998, pp. 11–30.

[21] A. Blaye, "Confrontation socio-cognitive et résolution de problèmes," Ph.D. dissertation, Centre de Recherche en Psychologie Cognitive, Université de Provence, 13261 Aix-en-Provence, France, 1988.

[22] M. Glachan and P. Light, "Peer interaction and learning: Can two wrongs make a right," in *Social cognition: Studies of the development of understanding*, ser. Developing body and mind. Harvester Press, 1982, no. 2, pp. 238–262.

[23] B. B. Schwarz, Y. Neuman, and S. Biezuner, "Two wrongs may make a right ... if they argue together!" *Cognition and Instruction*, vol. 18, no. 4, pp. 461–494, 2000.

[24] P. Dillenbourg, "What do you mean by collaborative learning?" in *Collaborative-learning: Cognitive and Computational Approaches*, P. Dillenbourg, Ed., 1999, pp. 1–19.

[25] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," *American psychologist*, vol. 55, no. 1, pp. 68–78, 2000.

[26] C. Bartneck, E. Croft, and D. Kulic, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *IJSR*, vol. 1, no. 1, pp. 71–81, 2009.

[27] P. S. Fry, "Success, failure, and self-assessment ratings," *Journal of Consulting and Clinical Psychology*, vol. 44, no. 3, pp. 413–419, 1976.