Heuristics for the Development and Evaluation of Educational Robotics Systems

Christian Giang, Alberto Piatti and Francesco Mondada

***Abstract—Contribution*: While previous research has studied the use of educational robotics in classrooms, there is still a lack of methods to support the development and evaluation of such tools. To this end, this paper presents an evaluation framework and a corresponding set of heuristics, which are specifically adapted to the needs and expectations in formal education settings.**

***Background*: The increased usage of educational robots in classrooms, as well as the steadily growing number of alternatives to choose from, raises the question of finding appropriate methods for the development and evaluation of such tools. Yet the current body of literature does not provide comprehensive frameworks that allow to adequately address this question.**

***Intended outcomes*: The work aims at providing an evaluation framework, which could support researchers and engineers, as well as educators and decision makers in taking informed decisions about educational robotics systems.**

***Application design*: This paper proposes to consider activities involving educational robotics systems as a kind of "educational augmented tabletop game". Within this framework, a set of fourteen heuristics was devised based on literature about games and learning tools. The validity of the devised heuristics was examined with a heterogenous group of twelve compulsory school teachers, who tested five different educational robotics systems.**

***Findings*: The experimental results illustrated high approval for the devised heuristics by the participating teachers. A heuristic evaluation based on the proposed framework appeared to be more appropriate to reflect the teachers' needs than conventional methods, i.e., the isolated comparison of system characteristics.**

***Index Terms*— Constructivist, design principles, educational robotics, games, heuristic evaluation, learning technology, STEM**

I. Introduction

IN RECENT YEARS, educational robotics (ER) have attracted a lot of attention from educators and researchers as a tool to support formal education [1]. The expected benefits of introducing robots into classrooms are manifold: it has been argued that ER promote students' interest in STEM (science, technology, engineering and mathematics) disciplines [1]–[5] and that they can be used to convey technical competencies, such as programming skills [6]–[8]. Moreover, it has been shown that by working with ER, students can acquire important transversal skills such as critical thinking, problem solving, decision making, communication or teamwork [9]–[12]. In addition, previous work has acknowledged its potential to foster the development of computational thinking skills [13], [14], a competence which has been popularized by Jeanette Wing in 2006, as a fundamental skill for modern societies along with reading, writing and arithmetic [15]. Other studies have demonstrated that ER can have positive effects on students' motivation, self-confidence and creativity [16], [17], hence facilitating a more joyful way of learning.

While most of the presented results seem very promising, there are still many open questions, due to the fact that scientific research in this field is still comparatively young [6], [18]. Previous work has classified educational robots into different types and analyzed their role and behavior during learning [19], explored the involved teaching domains and learning environments [9], [19], studied the implemented learning activities [9], and examined the acceptance of ER perceived by teachers and students [20]–[23]. However, it appears that the current body of literature does not yet provide comprehensive frameworks, which propose specific design principles for the development of ER tools. As a matter of fact, ER tools are usually developed by engineers and researchers, who often have none or limited experience with classroom teaching. These tools however, are often intended to be used by teachers, who perform learning activities with their students, in environments which are usually significantly different from controlled experimental conditions.

With respect to these circumstances, the question could be raised, whether current ER tools meet the expectations for the use in compulsory schools. Considering the large and steadily increasing number of alternatives to choose from [24], this question becomes even more pertinent. In this regard, a set of validated design principles could not only provide guidance to developers, but it could also serve as a support to identify appropriate tools for a given context. This paper introduces the "heuristics for the development and evaluation of educational robotics systems" (HEDEERS), a set of design principles devised specifically for ER tools. The presented approach aims at providing a holistic evaluation framework for ER tools, which is easy to use and applicable to a wide variety of different systems. The following sections describe the underlying model, the development of the heuristics and the experimental validation.

Manuscript received March 27, 2019; accepted April 7, 2019. (Corresponding author: Christian Giang)

F. Mondada and C. Giang are with the Laboratoire de Systèmes Robotiques (LSRO), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland (email : francesco.mondada@epfl.ch, christian.giang@epfl.ch).

A. Piatti and C. Giang are with the Department of Education and Learning (DFA), University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Locarno, Switzerland (email: alberto.piatti@supsi.ch, christian.giang@supsi.ch).

II. Background

The main motives of introducing ER tools into classrooms are based on Seymour Papert's *constructionism* [25]. Papert advocated for an active role of the learner, encouraging students to discover and form knowledge by self-exploration. One key aspect underlying his theory, is the manipulation of so-called learning artefacts, which allow students to actively build objects they can personally relate to. According to Papert, this is the most effective way of learning and in this context, educational robots appear to be a predestined embodiment of learning artefacts [9], [25]. As a matter of fact, ER tools provide many possibilities of active manipulation: from the assembly of the robot, to the programming of its behavior, up to the design of a personalized look – students can be involved in miscellaneous creative activities.

Over time, a countless number of various ER tools have been developed, each providing particular features and interaction methods (cf. Table I for examples). This diversity allows the end user to choose from a wide range of ER tools each with different behaviors and characteristics. However, the great variety makes it also difficult to systematically evaluate them, since the involved components, interaction methods and learning activities, as well as the foreseen learning objectives can differ significantly. For instance, some tools are specifically designed to be used by younger children, while others target older user groups. Taking this into consideration, it becomes clear, that an evaluation framework, which is generally applicable to a wide range of ER tools is not obvious.

There are three main approaches of using ER tools in classrooms: the theme-based curriculum, the project-based and the goal-oriented approach [1]. In the theme-based curriculum, the ER activities are part of a specific learning topic, which is explored by students within a given time span (e.g. during theme weeks). In the project-based approach, students work in groups to explore real-world problems and try to develop solutions to approach them. Finally, in the goal-oriented approach, students usually compete in extracurricular challenges (e.g. FIRST Lego League) to solve different kind of robotic tasks. The learning activities based on these approaches usually comprise a combination of various components: an educational robot, a programming/interaction interface, and one or several tasks to be solved. In some cases, the whole setting is embedded in a narrative, additionally providing a possibility for storytelling. Hence, when evaluating ER tools as a mean to perform classroom activities, it is essential to consider all these components together, as well as the interplay between them, since they always come as one entity. Therefore, this paper introduces a novel framework, called *educational robotics system* (*ERS*), which describes the combination of the educational robot, the programming/interaction interface and the presented tasks used for the classroom activities (Fig. 1). Consequently, this interconnection of multiple components also calls for an evaluation framework, which considers the entirety of the ERS, rather than assessing the system's components and their properties separately. Nevertheless, up until now, the segregated evaluation approach appears to be the most common practice for conventional evaluation methods. In previous studies, ER tools have mostly been evaluated based on selective characteristics of the robot (e.g. [8], [20], [26], [27]). However, the restriction to a few (mostly technical) evaluation criteria for only one component of the ERS (i.e., the robot), is arguably limiting the validity of such approaches. Moreover, no evidence has been presented regarding the validity of the selected evaluation criteria in such studies.

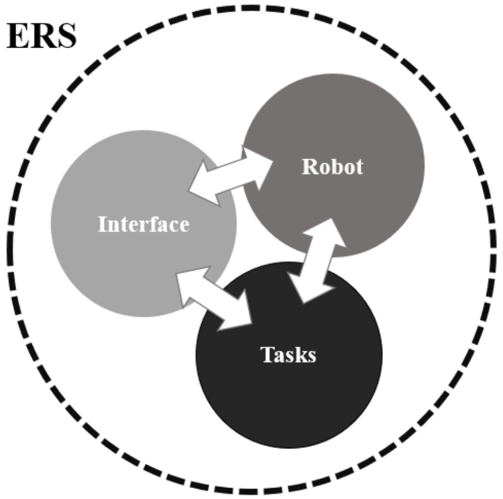

Fig. 1. Schematic representation of an educational robotics system (ERS). The white arrows indicate possible interaction effects between the components.

Capitalizing on the activity-based nature of ERS, this paper introduces a new perspective: indeed, many of the activities implemented using ERS comprise elements which are essential to classical tabletop games as well as to digital video games, such as interaction, enjoyment and challenge. Furthermore, many of these elements have been considered as core components for the design of so-called learning games [28]–[31]. These similarities suggest that activities involving ERS may be considered as a kind of "*educational augmented tabletop game*" which constitutes a combined entity of tangible tabletop games and digital learning games.

In usability research, heuristic evaluation approaches have proven to be a valuable and effective method for the evaluation and development of video, computer and board games [32], [33], as well as augmented tabletop games [34] and different kinds of learning tools [35]–[38]. Heuristics can be described as rule of thumbs, which can serve as design principles for the development and evaluation of products and are usually used by developers to identify a product's weak spots and limitations [39]. Drawing upon the presented interpretation of ERS as "*educational augmented tabletop games*", it seems justifiable, that heuristics, which has been successfully used for games and learning tools, can be combined and adapted for the use with ERS. The set of heuristics proposed in this paper consist of fourteen design principles, selected based on existing literature covering learning games, as well as augmented and classical tabletop games. The heuristics were selected by researchers with experience in classroom teaching and they were chosen with the aim to match the needs and expectations of compulsory school teachers. Table II summarizes the fourteen heuristics, the literature they were based on, and indicates whether the literature is game- or education-related.

TABLE I
EDUCATIONAL ROBOTICS SYSTEMS USED IN THIS STUDY

	Anki Cozmo[a]	**Calliope mini**[b]	**Lego WeDo 2.0**[c]	**Makeblock mBot**[d]	**Ozobot Evo**[e]
System type	Wheeled / social robot	Electronic kit	Construction kit	Wheeled robot	Wheeled / social robot
Programming	Graphical	Graphical	Graphical	Graphical	Tangible
User device	Tablet	PC	Tablet	Tablet	Pen and paper
Assembly	No	No	Yes	Yes	No
Extendable	No	Yes	No	Yes	No
Sensors	Camera with AI, ground sensors	Touch, buttons, light, temperature, gyroscope, compass, microphone	Distance, tilt	Distance, ground, light, button	Distance, RGB ground
Actuators	Motor, loudspeaker, color LED, screen, mechanical arm	Loudspeaker, color LED, 5x5 LED matrix	Motor	Motor, loudspeaker, LEDs	Motor, loudspeaker, LEDs

The information given in this table relates to one possible configuration of the respective system. All systems were presented using their standard available components. The interaction methods and presented tasks were chosen by the experimenters, while putting emphasis on the particularities of each system.
[a] https://www.anki.com/en-us/cozmo, [b] https://calliope.cc/, [c] https://education.lego.com/en-gb/product/wedo-2, [d] https://www.makeblock.com/steam-kits/mbot,
[e] https://ozobot.com/products/ozobot-evo

TABLE II
HEURISTICS FOR THE DEVELOPMENT AND EVALUATION OF EDUCATIONAL ROBOTICS SYSTEMS (HEDEERS)

No.	**Heuristic**	**Based on**	**GAME**	**EDU**
1	*Cognitive workload*: The system allows the user to maintain their sense of cognitive flow. Cognitive workload which is not related to the learning activities should be minimized.	[14,31,34]	X	X
2	*Challenge*: The system presents appropriate challenges tailored to the user. It should be “easy to learn, but hard to master”. The user’s fatigue is minimized by varying activities and pacing during the learning activities.	[31,34,35]	X	X
3	*Adaptability*: The system should be adaptable to the needs of the user. The system should be usable by all users of the target group regardless of their prior knowledge.	[34,35]	X	X
4	*Interaction*: The interaction method should satisfy the expectations of the user and follow the logic of the learning activities. The user interfaces should be compliant with industry standards and be usable in a very natural, easy and understandable way.	[31,34]	X	
5	*Level of automation*: The user should be able to execute all actions relevant to the learning activities by him/herself. All actions that are perceived as boring and rather unimportant to the learning activities should be performed by the system.	[34]	X	
6	*Collaboration and communication*: The entirety of the system should support interpersonal communication, collaboration and, if appropriate, competitiveness between users.	[31,34]	X	X
7	*Feedback*: The system should provide visual, acoustic or haptic feedback to help the user understand their performed actions and the resulting consequences.	[9,31,34]	X	X
8	*Comfort of the physical setup*: The physical setup should be fast and easy to assemble, comfortable to use and not require the user to take an awkward position.	[34]	X	
9	*Enjoyment and aesthetics*: The user should find the activities fun. The entirety of the system should be inviting and aesthetically appealing. It should quickly grab the user’s attention and facilitate the user’s concentration and immersion in the activities.	[31,33,35]	X	
10	*Transparency*: The system should provide a rich and open environment, allowing the inspection of all underlying mechanisms.	[9,14]		X
11	*Active learning*: The system encourages exploration, problem solving and enquiry. The user should feel safe in the knowledge that they can experiment without breaking the system.	[35, 38]		X
12	*Relevance*: The learning activities should be personally relevant to the user and allow him/her to relate the activities to the learning goals.	[9,31,35]		X
13	*Supports reflection*: The system should provide opportunities for reflection and debriefing on learning and highlight the process of learning to the user.	[31,35]		X
14	*Computational thinking*: The entirety of the system should support the development of computational thinking competences.	[2,24]		X

The last two columns of the table indicate whether the corresponding literature is game- or education-related.

TABLE III
PROFILES OF THE 12 PARTICIPATING TEACHERS

	Gender	Age	School	Background
T1	Male	30-50	Primary	Pedagogy
T2	Male	30-50	Primary	Psychology
T3	Male	> 50	Primary	Pedagogy
T4	Female	30-50	Primary	Pedagogy
T5	Female	> 50	Primary	Psychology
T6	Female	> 50	Primary	Pedagogy
T7	Female	> 50	Primary	Pedagogy
T8	Male	< 30	Lower sec.	Electrical Eng.
T9	Female	< 30	Lower sec.	Mathematics
T10	Male	30-50	Lower sec.	Aeronautical Eng.
T11	Male	< 30	Lower sec.	Mathematics
T12	Male	30-50	Lower sec.	Electrical Eng.

III. EXPERIMENTAL VALIDATION

HEDEERS was validated through experiments with twelve compulsory school teachers, who participated in an evaluation session, in which they tested five different ERS. During the evaluation session, they were asked to identify usability issues for each system. These issues were then mapped to HEDEERS in order to illustrate the applicability, completeness and orthogonality of the heuristics. Furthermore, based on the issues determined by the teachers, the systems were ranked regarding their suitability for classroom teaching. The results were then compared to two other rankings obtained from a questionnaire distributed at the end of the evaluation session: one ranking was based on the intuitive choices of the teachers, while the other one was based on a list of system characteristics and technical features (e.g., type of sensors/actuators, connection method etc.). The latter was a representation of conventional evaluation methods and the list of relevant characteristics was determined using the teachers' answers given in the questionnaire. Finally, the questionnaire also provided information about the acceptance of HEDEERS by the teachers and their satisfaction with the testing procedure.

A. Participants

To validate the matching between the selected heuristics and the expectations of compulsory school teachers, a heterogeneous group of 12 teachers (different gender, age, school level and professional background) was selected to participate in an evaluation session (Table III). At the time of the study, all teachers were in service and enrolled in a two-year training program for a certificate of advanced studies (CAS) in educational robotics. This CAS is the first of its kind in Switzerland, and participants are considered to be so-called early adopters, taking a pioneering role among their peers.

B. Educational Robotics Systems

Five ERS were tested by the participating teachers during the evaluation session. All systems were promoted for educational purposes by their manufacturers and none of them were known to the teachers before the study. Aiming at validating the heuristics for a wide range of ERS, the selection was based on the characteristics presented in Table I. The goal was to include a selection of systems comprising a great variety of the shown characteristics, in order to achieve a diverse representation of ERS. It is important to note that most of the information given in Table I relate to one possible configuration of the respective system. For instance, some systems allow the use of various programming interfaces and user devices or can be extended with additional sensors and/or actuators. For this study, all systems were presented using their standard available components. Moreover, the interaction methods and proposed tasks were chosen by the experimenters, while putting emphasis on the particularities of each system. Since the aim of this study was the evaluation of the devised heuristics rather than the assessment of the selected ERS, not presenting all possible configurations to the teachers is justifiably not a relevant problem.

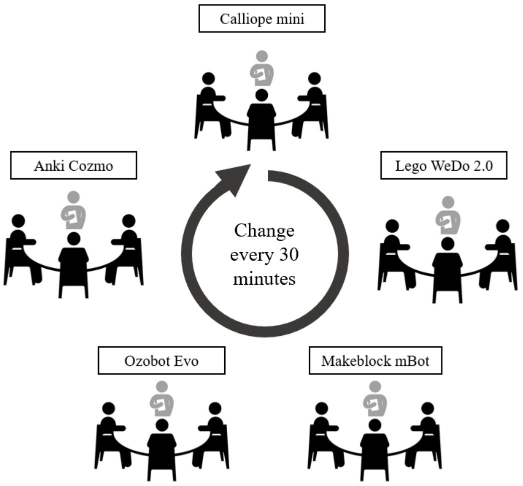

Fig. 2. Schematic overview of the evaluation procedure. Each group of 2-3 teachers (black) was followed by one experimenter (gray). On a rotational basis, each group tested each robotic educational system.

C. Testing procedure

At the beginning of the session, all participants were introduced to HEDEERS, ensuring that there were no unclarities about the meaning of each heuristic. For the evaluation study, the teachers formed five groups of 2-3 people, a recommended group size for ER activities [9]. A schematic overview of the testing procedure is depicted in Fig. 2.

Each group had 30 minutes to test a system by working through a worksheet prepared by the experimenters. Each test was introduced by a short video made by the manufacturer highlighting the features and functionalities of the presented ERS. This was followed by a user tutorial, showing the group how to get started with the system. Finally, the group was given time to further explore the system by either following proposed activity suggestions or by following their own ideas and interests. On a rotational basis, each group tested all five systems and was asked to identify as many usability issues as possible for each system while testing. A printout with the set of heuristics was available on all tables. However, the teachers were not obliged to use them, and they could identify issues not related to the heuristics. Moreover, the teachers were instructed to put emphasis on identifying usability issues related to the possible use of the systems in class, which also includes difficulties that their students might encounter when working with the presented ERS.

Throughout the tests, each group was followed by one experimenter, who supervised the heuristic evaluation by taking the role of a so-called “observer” [39]. The main function of the observers was to record the usability issues determined by the teachers, using written reports for later analysis. Furthermore, they provided technical support in order to facilitate the testing procedure.

D. Weighting and mapping of usability issues

The written reports obtained from the observers provided a summary of all usability issues for every ERS identified by each group. However, since some issues may have a stronger impact on the user experience than others, it was not sufficient to only consider the quantity of the identified issues for each system. Instead, the teachers were also asked to assign a weight to each issue, in order to obtain a meaningful weighting among the issues. The weights were based on the Nielsen severity scale [34], [39] for usability issues:

- 0 - *Not a usability problem at all*.
- 1 - *Cosmetic problem only*: It does not have a profound impact on the activity.
- 2 - *Minor problem*: It has a slight impact on the activity and influences the experience a bit.
- 3 - *Major problem*: This problem has a severe impact on the activity and negatively influences the user experience.
- 4 - *Usability catastrophe*: This problem has to be fixed in order to allow for a decent user experience.

Assuming that different groups would identify different usability issues [39], all the identified issues were merged into one list for each system by the experimenters at the end of the study. Subsequently, each issue was mapped to one of the heuristics. The information about the aggregate of all issues identified for a system and the heuristics they were mapped to, was hence only available to the experimenters and not to the teachers. The final ranking among the systems was determined based on the number and the severity of the issues identified for each system. The systems were ranked by always giving a higher importance to more severe issues (e.g., one usability catastrophe is always worse than many minor problems). If two systems had the same number of issues for one severity class, the next lower class was considered.

E. Questionnaire

At the end of the evaluation session, all participants were asked to complete a questionnaire comprising four subsections: first, they were asked to provide a personal ranking of the five systems based on the following question: “*For the use in my class, I would choose the systems in the following order”*. This information was used by the experimenters to determine a ranking based on the teachers’ intuitive choices. In the next subsection, the teachers were asked to assess the relevance of each heuristic using a 4-point Likert scale (strongly agree / agree / disagree / strongly disagree). Moreover, they were asked to propose amendments to the heuristics, in case they thought that some criteria were missing. The answers were used to analyze the teachers’ acceptance of the devised heuristics. Subsequently, the teachers were asked to rate the importance of different system characteristics of ERS (e.g., type of sensors/actuators, connection method etc.) using the same 4-point Likert scale. Additionally, they were given the possibility to indicate features and components that were not listed. This information was used by the experimenters to determine a ranking based on system characteristics. Finally, in the last subsection of the questionnaire, the participants were asked to provide their opinion about the testing procedure.

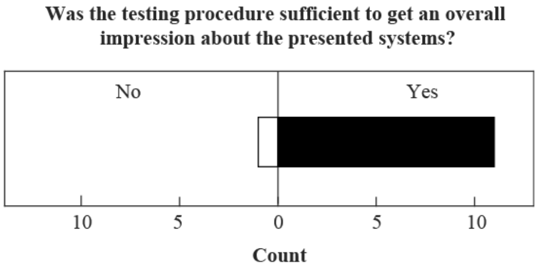

Fig. 3. Satisfaction with the testing procedure reported from the questionnaire. Almost all participants (11 out of 12) considered the testing procedure as sufficient for getting an overall impression about the presented systems.

IV. Results and discussion

Since most of the results presented in this section rely on the assumption that all participants were able to form an opinion about the presented systems, it was important to ensure that the evaluation session allowed them to explore the systems sufficiently well. As reported in the last subsection of the questionnaire, it appears that the testing procedure indeed allowed the teachers to adequately discover the systems within the 30 minutes of testing (Fig. 3). A crucial point might have been the prepared worksheets, which facilitated the systematic exploration of the systems. Furthermore, it can be assumed that also the presence of the observers might have contributed to the efficacy of the testing procedure, since technical issues were resolved quickly, allowing the participants to remain focused on the main tasks.

A. Acceptance of heuristics

As a first mean to examine the validity of HEDEERS, the teachers were asked to assess the relevance of each heuristic using a 4-point Likert scale (see section *Questionnaire*). Demonstrating a general acceptance of the devised heuristics from a heterogenous groups of compulsory school teachers would provide evidence for their validity, since ultimately, teachers are the ones who select, adapt and create learning activities involving ERS.

The results illustrated that a large majority of the teachers agreed on the validity of HEDEERS as design principles for the development and evaluation of ERS (Fig. 4). For some of the heuristics (e.g., *level of automation* and *comfort of physical setup*) the acceptance was lower compared to others (e.g., *active learning* and *cognitive workload*). However, as a matter of fact, approval was dominating rejection for all the devised heuristics,

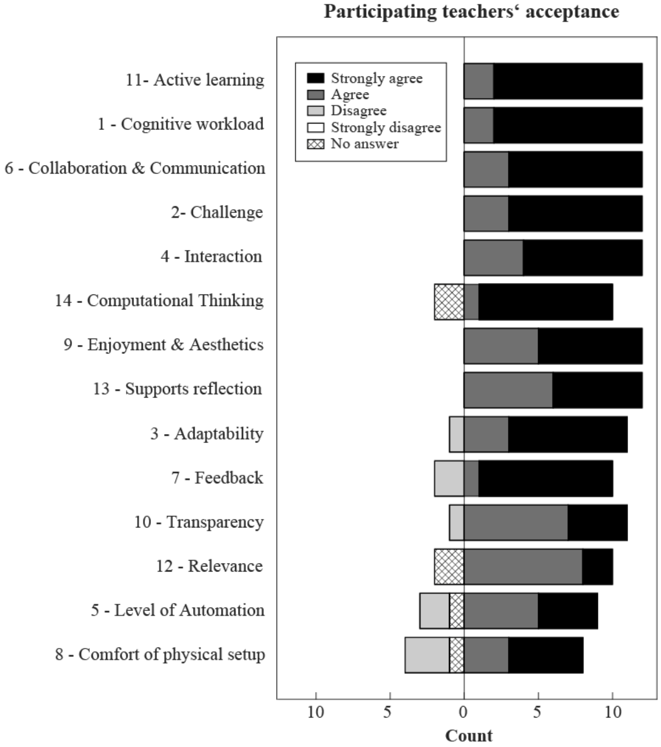

Fig. 4. Acceptance of each heuristic by the teachers participating in the study (n = 12). A large majority agreed on the validity of HEDEERS as design principles for the development and evaluation of ERS.

demonstrating a general acceptance by the participating teachers. Interestingly, also heuristics based only on game-related literature (e.g., *interaction* and *enjoyment & aesthetics*) found wide acceptance, supporting the proposed approach to consider ERS as a kind of "*educational augmented tabletop game*". Nevertheless, it should be mentioned that the two heuristics that found the least acceptance (i.e., *level of automation* and *comfort of physical setup*) were also based only on game-related literature. Yet more research would be needed to investigate whether all game-related (and education-related) heuristics apply equally well to ERS. For some of the heuristics (e.g., *computational thinking* and *relevance*) a few participants did not respond, indicating that these heuristics may need a more precise description to prevent a lack of clarity.

Finally, the teachers were also asked to propose amendments to the heuristics, in case they considered that some criteria were missing. However, none of the participating teachers suggested any amendments, indicating that HEDEERS comprises all the criteria the teachers considered to be important for the use of ERS in classrooms.

B. Ranking based on usability issues

A total of 63 usability issues (1 usability catastrophe, 26 major, 22 minor and 14 cosmetic problems) were identified by the participating teachers for all systems (Fig. 5, bottom right panel). The results illustrated that almost half (31) of the identified usability issues could be associated to three heuristics: *interaction* (12), *adaptability* (11) and *comfort of the physical setup* (8). This is not surprising, since usability issues related to these heuristics are often easily noticeable. Some of the usability issues associated to these heuristics were for example: "*Programming interface not intuitive*" *(mapped to interaction)*, "*Limited number of actuators constrains possibilities*" *(adaptability)* or "*Setting up the system takes too much time*" *(comfort of the physical setup)*.

Usability issues were found for all heuristics except for one (i.e., *supports reflection*) and all the identified issues could be clearly mapped to one of the fourteen heuristics by the experimenters. Moreover, there were no issues identified by the teachers which could not be associated to a heuristic. This highlights the completeness and orthogonality of HEDEERS, while emphasizing its applicability as a holistic guiding tool for the development and evaluation of a wide range of different ERS. Nevertheless, it seems that usability issues related to some of the heuristics would need more extensive testing for identification, since they could be less evident to discover. Extending the testing time could therefore reveal more usability issues for heuristics where none or only few issues have been identified (e.g. *supports reflection*, *level of automation* or *feedback*), and hence provide a more exhaustive evaluation of the presented ERS.

The individual analysis for each of the five systems presented during the evaluation session revealed considerable differences (Fig. 5, first five panels). While most usability issues found for *Makeblock mBot* did not have a severe impact on the user experience (0 usability catastrophes (Ca) / 1 major problem (Ma) / 7 minor problems (Mi) / 3 cosmetic problems (Co)), the issues found for *Anki Cozmo* (Ca-0 / Ma-5 / Mi-0 / Co-6), *Ozobot Evo* (Ca-0 / Ma-5 / Mi-7 / Co-3) and *Lego WeDo 2*.0 (Ca-0 / Ma-8 / Mi-5 / Co-1) had a larger influence. The only usability catastrophe identified during the evaluation study was found for *Calliope mini* (Ca-1 / Ma-7 / Mi-3 / Co-1). The final ranking based on the identified usability issues, is presented in the second column of Table IV. Although the usability catastrophe found for *Calliope mini* (i.e., "*Only German user language" (interaction)*) is strongly related to the non-German-speaking participants of this study, it seems like user language is a non-negligible factor for the use of ERS in classrooms. Albeit some elements of *Calliope mini* were available in English, most teachers reported that they preferred a translation to their mother tongue (Italian). Under the present circumstances, the teachers believed they were not able to use this ERS in class and hence, decided to assign the most severe weight (i.e., *usability catastrophe*) to this usability issue.

As hypothesized, only few overlaps were found for the issues identified by the different groups. As typical for heuristic evaluations involving a limited number of evaluators [39], each group identified different usability issues for each system. Nevertheless, it can be assumed that the aggregate of the usability issues identified by the five groups is a reasonable representation of the systems' limitations, providing an extensive list of weaknesses and limitations for each system. As shown by Nielsen and Landauer [40], five evaluators are usually sufficient to uncover almost 75% of the known usability issues of a product.

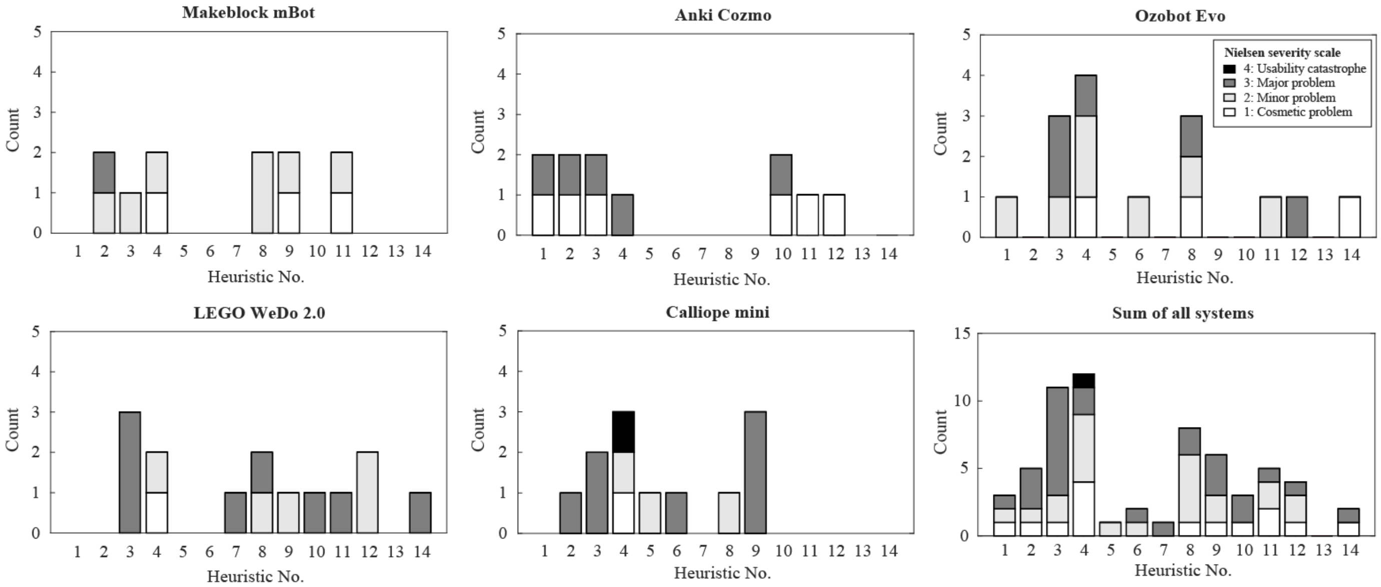

Fig. 5. Usability issues identified by the participating teachers during the evaluation session for each ERS (first five panels) and for all systems together (last panel). The bars indicate the number and severity of the issues associated to each heuristic.

TABLE IV
RANKINGS FOR THREE DIFFERENT APPROACHES

Rank	By usability issues	By intuitive choice	By system characteristics
1	Makeblock mBot (Ca-0/Ma-1/Mi-7/Co-3)	Makeblock mBot (85%)	Ozobot Evo (77%)
2	Anki Cozmo (Ca-0/Ma-5/Mi-0/Co-6)	Anki Cozmo (63%)	Makeblock mBot (69%)
3	Ozobot Evo (Ca-0/Ma-5/Mi-7/Co-3)	Ozobot Evo (58%)	Calliope mini (62%)
4	Lego WeDo 2.0 (Ca-0/Ma-8/Mi-5/Co-1)	Lego WeDo 2.0 (33%)	Anki Cozmo (54%)
5	Calliope mini (Ca-1/Ma-7/Mi-3/Co-1)	Calliope mini (10%)	Lego WeDo 2.0 (38%)

Ranking by usability issues: the values in parentheses indicate the number of issues identified (usability catastrophes (Ca) / major problems (Ma) / minor problems (Mi) / cosmetic problems (Co)). More severe issues are always given more importance (e.g., one usability catastrophe is always worse than many minor problems). *Ranking by intuitive choice and ranking by system characteristics*: the values in parentheses indicate the percentage of the maximum achievable score.

C. *Ranking based on intuitive choices*

By means of the questionnaire distributed at the end of the session (see section *Questionnaire*), each teacher was asked to provide a personal intuitive ranking of the five presented systems regarding their usability in class (Fig. 6). In order to quantify an overall result, a scoring system was introduced: a system was assigned 4 points every time it was selected as a first choice (3 for second, 2 for third, 1 for fourth and 0 for fifth choice). The final ranking (Table IV, third column) was established based on the ratio between the points obtained by each system and the maximum achievable score (48 points).

Remarkably, the results yielded a consistent match with the ranking based on the identified usability issues. The dominant lead of *Makeblock mBot* in the personal ranking (85% of the maximal score) is equally reflected by the low number of severe usability issues identified by the teachers (Ca-0 and Ma-1). It is followed by *Anki Cozmo* (63%) and *Ozobot Evo* (58%), which convinced a similar number of teachers. Coherently, an equal number of severe issues was found for both systems (Ca-0 and Ma-5 for both). The higher number of low severity issues for *Ozobot Evo* account for its lower positioning, which is in accordance with the results based on the personal choices. *Lego WeDo 2.0* (33%) and *Calliope mini* (10%) ranked on the last two places of the personal ranking. A similar result was obtained from the ranking based on the usability issues,

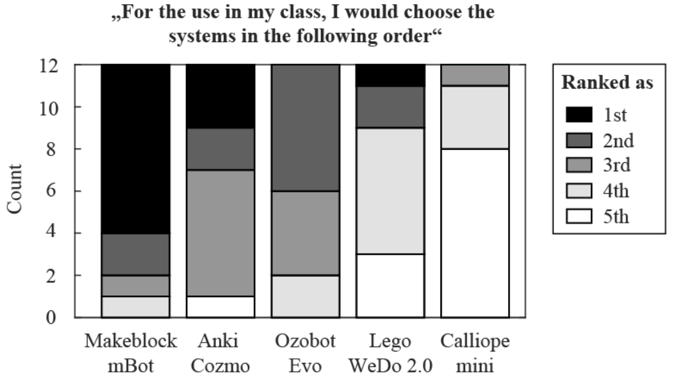

Fig. 6. Personal intuitive choices of the participating teachers. Each bar indicates the number of teachers who ranked the corresponding system as their first, second, third, fourth or fifth choice for the use in class.

reflected by the high number of severe issues identified for both systems (Ca-0 and Ca-1, Ma-8 and Ma-7, respectively). The low rating for *Calliope mini* might again be related to the language barrier, caused by the German user language. As a result, most of the teachers could not explore the system in the way a German-speaker could, which led to a limited user experience, and thus, to the low positioning of this ERS.

In summary, the consistency between both rankings can be interpreted as another indicator for the validity of HEDEERS as an evaluation framework for ERS. The fact that the personal intuitive choices of the teachers matched with the ranking based on the usability issues, which could all be successfully mapped to the heuristics, indicates that there are no criteria outside the ones listed in HEDEERS, that the teachers considered important enough to impact their opinion about the presented ERS. It could be argued that the intuitive choices of the teachers could have been biased by the usability issues they identified before, and a matching between both rankings would therefore not surprise. However, as mentioned in the previous section, each group identified different issues, and thus, none of the teachers was aware of the aggregate of all issues when they indicated their intuitive choices.

D. Ranking based on system characteristics

Aiming at evaluating the effectivity of HEDEERS compared to conventional methods, a third ranking was established. For this purpose, the presented systems were rated by different characteristics and features, determined by means of the answers given by the teachers in the third subsection of the questionnaire. From a list of different system characteristics, the teachers had to indicate which ones they considered relevant using a 4-point Likert scale (see section *Questionnaire*). For the ranking, the five characteristics with the highest values of approval (more than 83% of the teachers agreed or strongly agreed on their relevance) were included. Additionally, the teachers were asked to indicate the sensors and actuators they considered as most relevant for their teaching. For the ranking, the four most commonly mentioned sensors and actuators were included. The complete list with the most preferred system characteristics, sensors and actuators was then used by the experimenters to evaluate the presented ERS: for each item on the list that a system complied with, it was given 1 point (Table V). The final ranking (Table IV, fourth column) was then established based on the ratio between the points for each system and the maximum achievable score (13 points).

The results illustrate that the ranking obtained following this approach did not provide a coherent match with the intuitive choices of the teachers. While the intuitive ranking determined *Makeblock mBot* as the teachers' clear preference, this approach ranks it in the second place (69% of the maximum score). Instead, *Ozobot Evo* (77%) was determined as the leader of the ranking, which has only been third based on the teachers' intuitive choice. Another remarkable difference is given by the good rating of *Calliope mini* (62%), which came in last far behind in the intuitive ranking. It is followed by *Anki Cozmo* (54%) *and Lego WeDo 2.0* (38%), which ranked on the two last places following this approach.

TABLE V
PARTICIPATING TEACHERS' PREFERRED SYSTEM CHARACTERISTICS

	AC	CM	LW	MM	OE
Wireless connection	X	X	X	X	X
Didactic material		X	X		X
System extendable		X		X	
Usable on a desk	X	X	X		X
Preprogrammed	X	X		X	X
Distance sensor			X	X	X
Input button		X		X	
RGB ground sensor					X
B/W ground sensor				X	X
Motor	X		X	X	X
LED	X	X		X	X
Lifting arm	X				
Loudspeaker	X	X		X	X

List of participating teachers' preferred system characteristics and components determined from the questionnaire. Crosses indicate whether a system complies with a desired property (*AC = Anki Cozmo, CM = Calliope mini, LW = Lego WeDo 2.0, MM = Makeblock mBot, OE = Ozobot Evo*).

The obtained results underline the presumed weakness of conventional evaluation approaches for ERS. Indeed, the isolated evaluation of system characteristics and components of ERS appears to be insufficient to appropriately represent the needs of compulsory school teachers. This is particularly interesting, since for this study, the teachers were given the possibility to specify the characteristics and components they considered as relevant. However, formerly such criteria have been selected and applied for the evaluation of ERS without demonstrating any evidence about their relevance to the user. Moreover, such approaches mostly focused on the properties of the robot, while the other components of the ERS, namely the programming/interaction interface and the proposed tasks, were often neglected. Based on the results of this study, it seems like a more holistic approach, which considers the entirety of the system (i.e., the robot, the programming/interaction interface and the involved tasks), could provide more appropriate representations of what is really desired in formal education settings.

V. CONCLUSION, LIMITATIONS AND FURTHER RESEARCH

This paper introduced a framework to support the development and evaluation of ER tools targeted to the use in formal education settings. The underlying idea of the proposed approach is to consider the entirety of an educational robotics system (ERS, i.e., the combination of the robot, the programming/interaction interface and the proposed tasks), rather than the system's components and their properties separately. Drawing upon this holistic approach and the activity-based nature of ERS, this study proposes to consider activities involving ERS as a kind of "*educational augmented tabletop game*". Within this framework, ERS can be evaluated by heuristic evaluation, a methodology which has been proven useful for the evaluation of games and learning tools. Existing heuristics coming from those fields were combined and adapted to devise HEDEERS, a set of fourteen heuristics, providing design principles for ERS that are specifically aimed at meeting the needs and expectations in formal education settings. An evaluation study with a heterogenous group of twelve compulsory school teachers validated the applicability, completeness and orthogonality of HEDEERS and highlighted its characteristics with respect to conventional evaluation methods. The results showed that the heuristics embodied a good representation of the teachers' needs regarding the use of ERS in classrooms. The proposed framework could therefore guide researchers and engineers in the development of ER tools, providing design principles which are coherent with the needs and expectations of compulsory school teachers. In this context, HEDEERS may not only be useful for the development of the robot, but also for the design of the programming/interaction interface and the creation of the proposed tasks. However, in any of these cases it is imperative that developers consider the entirety of the intended ERS, since the corresponding learning activities usually involve a combination of all the three components. The results also showed that HEDEERS could potentially be applied for the evaluation of existing ERS. This could be particularly interesting for educators and decision makers interested in selecting ER tools for the use in formal education. In this regard, a heuristic evaluation using HEDEERS could help to identify limitations and weaknesses of a given ERS and moreover, represent a resource-efficient alternative to costly and time-consuming user studies.

Due to the small sample size, the results of this study should rather be interpreted as a proof of concept for the proposed model for ERS and the corresponding set of heuristics. Therefore, studies with larger sample sizes should be considered, in order to draw more substantial conclusions. However, it can also be argued that previous studies on heuristics for digital and tabletop games involved similar sample sizes for the validation (e.g., [32], [34]) and yet yielded recognized results. Another limitation of this study is given by the mapping of the identified issues to the heuristics, performed by the researchers and hence involving an implicit risk of bias. Further research should therefore involve the application of the heuristics by independent evaluators, in order to consolidate the validity and applicability of HEDEERS. Ideally, future work would include real development scenarios where the heuristics are used for the design of new ERS components. Moreover, this study mainly focused on the validation of the heuristics by teachers from primary and lower secondary schools. It could be interesting to extend the testing audience to teachers from higher education levels to investigate whether HEDEERS also applies for more advanced teaching scenarios.

Finally, it should be noted that this first version of HEDEERS was developed based purely on existing literature and the classroom experiences of the experimenters. Reaching out to the main stakeholders involved in ER (i.e., developers, teachers and students) could provide valuable input to refine the heuristics and allow a more detailed description for each heuristic. Especially students seem to be too little involved so far, yet they are the ones who ultimately interact with these tools. In this context, qualitative research methods, such as observational field studies or focus groups could be considered. Especially focus groups have been shown to be useful for the creation of design science research artefacts [41]. Performing focus groups with developers, teachers and students could therefore provide insight into their perceptions about ER, possibly reveal conceptual differences among those groups and hence support the development of future ERS.

Acknowledgment

We would like to thank Morgane Chevalier and Mordechai (Moti) Ben-Ari for their contributions to this study. Moreover, we would like to thank all the teachers who participated in the study and all the members of SUPSI-DFA and IDSIA who supported us in conducting the evaluation session.

References

[1] D. Alimisis, "Educational robotics: Open questions and new challenges," *Themes Sci. Technol. Educ.*, vol. 6, no. 1, pp. 63–71, 2013.

[2] S. Atmatzidou and S. Demetriadis, "Advancing students' computational thinking skills through educational robotics: A study on age and gender relevant differences," *Rob. Auton. Syst.*, vol. 75, pp. 661–670, 2015.

[3] A. Vollstedt, M. Robinson, and E. Wang, "Using Robotics to Enhance Science , Technology , Engineering , and Mathematics Curricula," *Am. Soc. Eng. Educ. Pacific Southwest Annu. Conf.*, 2007.

[4] M. Mataric, N. Koenig, and D. Feil-Seifer, "Materials for Enabling Hands-On Robotics and STEM Education.," *AAAI Spring Symp. Semant. Sci. Knowl. Integr.*, no. March, pp. 99–102, 2007.

[5] R. B. Levy and M. M. Ben-ari, "Robotics Activities–Is the Investment Worthwhile?," in *Informatics in Schools. Curricula, Competences, and Competitions*, 2015, vol. 9378, pp. 22–31.

[6] D. Alimisis and C. Kynigos, *Teacher Education on Robotics-Enhanced Constructivist Pedagogical Methods*. 2009.

[7] S. Atmatzidou, I. Markelis, and S. Demetriadis, "The use of LEGO Mindstorms in elementary and secondary education : game as a way of triggering learning," *Work. Proc. Int. Conf. Simulation, Model. Program. Auton. Robot.*, pp. 22–30, 2008.

[8] I. R. Nourbakhsh, K. Crowley, A. Bhave, E. Hamner, T. Hsiu, A. Perez-Bergquist, S. Richards, and K. Wilkinson, "The robotic autonomy mobile robotics course: Robot design, curriculum design and educational assessment," *Auton. Robots*, vol. 18, no. 1, pp. 103–127, 2005.

[9] F. B. V. Benitti, "Exploring the educational potential of robotics in schools: A systematic review," *Comput. Educ.*, vol. 58, no. 3, pp. 978–988, 2012.

[10] S. Blanchard, V. Freiman, and N. Lirrete-Pitre, "Strategies used by elementary schoolchildren solving robotics-based complex tasks: Innovative potential of technology," *Procedia - Soc. Behav. Sci.*, vol. 2, no. 2, pp. 2851–2857, 2010.

[11] S. Atmatzidou and S. N. Demetriadis, "Evaluating the role of collaboration scripts as group guiding tools in activities of educational robotics: Conclusions from three case studies," *Proc. 12th IEEE Int. Conf. Adv. Learn. Technol. ICALT 2012*, pp. 298–302, 2012.

[12] M. Petre and B. Price, "Using Robotics to Motivate 'Back Door' Learning," *Educ. Inf. Technol.*, vol. 9, no. 2, pp. 147–158, 2004.

[13] A. Repenning, D. Webb, and A. Ioannidou, "Scalable game design and the development of a checklist for getting computational thinking into public schools," *Proc. 41st ACM Tech. Symp. Comput. Sci. Educ. - SIGCSE '10*, p. 265, 2010.

[14] I. Lee, F. Martin, J. Denner, B. Coulter, W. Allan, J. Erickson, J. Malyn-Smith, and L. Werner, "Computational thinking for youth in practice," *ACM Inroads*, vol. 2, no. 1, p. 32, 2011.

[15] J. M. Wing, "Computational Thinking," vol. 49, no. 3, pp. 33–35, 2006.

[16] D. Miller, I. Nourbakhsh, and R. Siegwart, "Robots for Education," *Springer Handb. Robot.*, pp. 1283–1301, 2008.

[17] A. Khanlari, "EFFECTS OF ROBOTICS ON 21st CENTURY

SKILLS," *Eur. Sci. Journal, ESJ*, vol. 9, no. 27, pp. 26–36, 2013.

[18] M. J. Matarić, "Robotics Education for All Ages," *Proceedings, AAAI Spring Symp. Access. Hands-on AI Robot. Educ.*, 2004.

[19] O. Mubin, C. J. Stevens, S. Shahid, A. Al Mahmud, and J.-J. Dong, "a Review of the Applicability of Robots in Education," *Technol. Educ. Learn.*, vol. 1, no. 1, 2013.

[20] M. Chevalier, F. Riedo, and F. Mondada, "Pedagogical Uses of Thymio II," *IEEE Robot. Autom. Mag.*, vol. 23, no. 2, pp. 16–23, 2016.

[21] S. Kradolfer, S. Dubois, F. Riedo, F. Mondada, and F. Fassa, "A sociological contribution to understanding the use of robots in schools: The thymio robot," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8755, pp. 217–228, 2014.

[22] M. Fridin and M. Belokopytov, "Acceptance of socially assistive humanoid robot by preschool and elementary school teachers," *Comput. Human Behav.*, vol. 33, pp. 23–31, 2014.

[23] K. Kim, H. Choi, and J. Baek, "A Study on the Teachers ' Perception of School Curriculum Implementation about Robot-based Education in Korea Concept of Robot-Based Education," *Adv. Sci. Technol. Lett.*, vol. 59, no. Education, pp. 105–108, 2014.

[24] L. Mannila, V. Dagiene, B. Demo, N. Grgurina, C. Mirolo, L. Rolandsson, and A. Settle, "Computational Thinking in K-9 Education," *Proc. Work. Gr. Reports 2014 Innov. Technol. Comput. Sci. Educ. Conf. - ITiCSE-WGR '14*, pp. 1–29, 2014.

[25] S. Papert, *Mindstorms: Children, computers and powerful ideas*, vol. 1. 1980.

[26] E. B. B. Gyebi, M. Hanheide, and G. Cielniak, "Affordable Mobile Robotic Platforms for Teaching Computer Science at African Universities," *Robot. Educ.*, no. May, p. 7, 2015.

[27] T. Pachidis, E. Vrochidou, V. G. Kaburlasos, S. Kostova, M. Bonković, and V. Papić, "Social Robotics in Education: State-of-the-Art and Directions," in *Advances in Service and Industrial Robotics*, 2019, pp. 689–700.

[28] M. Prensky, "Computer games and learning: Digital-based games," *Handbook of Computer Game Studies*. pp. 97–124, 2005.

[29] R. Garris, R. Ahlers, and J. E. Driskell, "Games, motivation, and learning: A research and practice model," *Simul. Gaming*, vol. 33, no. 4, pp. 441–467, 2002.

[30] T. W. Malone, "What makes things fun to learn? heuristics for designing instructional computer games," *Proc. 3rd ACM SIGSMALL Symp. first SIGPC Symp. Small Syst. - SIGSMALL '80*, vol. 162, pp. 162–169, 1980.

[31] S. Boller and K. M. Kapp, *Play to learn: Everything you need to know about designing effective learning games*. Alexandria, VA: ATD Press, 2017.

[32] H. Desurvire, M. Caplan, and J. A. Toth, "Using heuristics to evaluate the playability of games," *Ext. Abstr. 2004 Conf. Hum. factors Comput. Syst. - CHI '04*, p. 1509, 2004.

[33] P. Sweetser and P. Wyeth, "GameFlow: A Model for Evaluating Player Enjoyment in Games," *Comput. Entertain.*, vol. 3, no. 3, pp. 3–3, 2005.

[34] C. Köffel and M. Haller, "Heuristics for the Evaluation of Tabletop Games," *Eval. User Exp. Games, Work. 2008 Conf. Hum. Factors Comput. Syst.*, no. March, 2008.

[35] M. B. Barbosa, A. B. Rêgo, and I. de Medeiros, "HEEG : Heuristic Evaluation for Educational Games," pp. 224–227, 2015.

[36] H. Mohamed and a. Jaafar, "Development and potential analysis of Heuristic Evaluation for Educational Computer Game (PHEG)," *Comput. Sci. Converg. Inf. Technol. (ICCIT), 2010 5th Int. Conf.*, pp. 222–227, 2010.

[37] N. Jerzak and F. Rebelo, "Serious games and heuristic evaluation - The cross-comparison of existing heuristic evaluation methods for games," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8517 LNCS, no. PART 1, pp. 453–464, 2014.

[38] M. Kölling and F. McKay, "Heuristic Evaluation for Novice Programming Systems," *Trans. Comput. Educ.*, vol. 16, no. 3, p. 12:1--12:30, Jun. 2016.

[39] J. Nielsen, "Heuristic evaluation," in *Usability Inspection Methods*, John Wiley & Sons, New York, NY, 1994.

[40] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in *CHI '93 Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, 1993, pp. 206–213.

[41] M. C. Tremblay, A. R. Hevner, and D. J. Berndt, "The Use of Focus Groups in Design Science Research," in *Design Research in Information Systems: Theory and Practice*, Boston, MA: Springer US, 2010, pp. 121–143.

Christian Giang received a M.Sc. in Electrical Engineering and Information Technology from the Swiss Federal Intitute of Technology in Zurich (ETH Zürich). Since 2017 he is a PhD student in the joint doctoral program of the École Polytechnique Fédérale de Lausanne (EPFL) and the University of Applied Sciences and Arts of Southern Switzerland (SUPSI). His research focus is on the use of educational robotics in formal education.

Alberto Piatti is director of the Department of Education and Learning (DFA) at the University of Applied Sciences and Arts of Southern Switzerland (SUPSI) and senior teacher-researcher in Mathematics and Mathematics education at SUPSI. He holds a master's degree in Mathematics from ETH Zürich and a PhD in economics from the University of Lugano. After a Postdoc and advanced research activities in the field of human knowledge modelling at the Dalle Molle Institute for Artificial Intelligence (IDSIA) and teaching activities at SUPSI and at the University of Lugano, he joined SUPSI-DFA in 2010 as a member of the Executive Board and teacher-researcher in Mathematics Education.

Francesco Mondada is professor at the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland and director of the Center for Learning Sciences at EPFL. After a master and a PhD received at EPFL, he led the design of many miniature mobile robots, commercialized and used worldwide in thousands of schools and universities. He co-founded several companies selling these robots or other educational tools. He is author of more than a hundred publications in the field of robot design. He received several awards, including the Swiss Latsis University prize, as best young researcher at EPFL and the Credit Suisse Award for Best Teaching as best teacher at EPFL.