EvalMe

Citation for published version:
Lechelt, S, Brudy, F, Marquardt, N & Rogers, Y 2021, EvalMe: Exploring the value of new technologies for in situ evaluation of learning experiences. in *CHI '21 : Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.*, 59, ACM, The ACM CHI Conference on Human Factors in Computing Systems 2021, Virtual Conference, Japan, 8/05/21. https://doi.org/10.1145/3411764.3445749

Digital Object Identifier (DOI):
10.1145/3411764.3445749

Document Version:
Peer reviewed version

Published In:
CHI '21

EvalMe: Exploring the Value of New Technologies for In Situ Evaluation of Learning Experiences

Susan Lechelt
Institute for Design Informatics, University of Edinburgh, Edinburgh, United Kingdom
susan.lechelt@ed.ac.uk

Frederik Brudy
University College London, London, United Kingdom & Autodesk Research, Toronto, Canada
frederik.brudy@autodesk.com

Nicolai Marquardt
University College London, London, United Kingdom
n.marquardt@ucl.ac.uk

Yvonne Rogers
University College London, London, United Kingdom
y.rogers@ucl.ac.uk

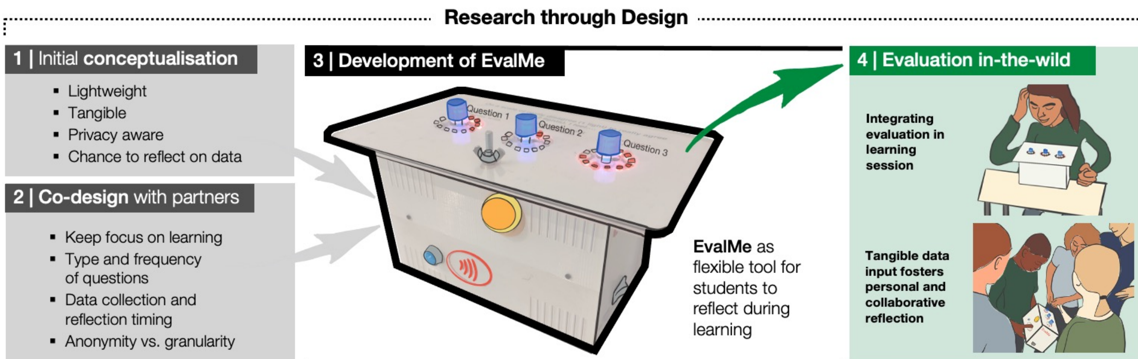

Figure 1: Through a Research through Design approach, we conceptualized and co-designed the EvalMe system, a flexible tangible tool that can be used during workshops or in class settings to elicit reflection on learning "in the moment".

ABSTRACT

Tangible interfaces have much potential for engendering shared interaction and reflection, as well as for promoting playful experiences. How can their properties be capitalised on to enable students to reflect on their learning, both individually and together, throughout learning sessions? This Research through Design paper describes our development of EvalMe, a flexible, tangible tool aimed at being playful, enjoyable to use and enabling children to reflect on their learning, both in the moment and after a learning session has ended. We discuss the insights gained through the process of designing EvalMe, co-defining its functionality with two groups of collaborators and deploying it in two workshop settings. Through this process, we map key contextual considerations for the design of technologies for in situ evaluation of learning experiences. Finally, we discuss how tangible evaluation technologies deployed throughout a learning session, can positively contribute to students' reflection about their learning.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); Human computer interaction (HCI); Interaction devices; • **Applied computing** → Education.

KEYWORDS

Education/learning, Tangible, Schools, Educational Setting, Children/Parents

ACM Reference Format:
Susan Lechelt, Frederik Brudy, Nicolai Marquardt, and Yvonne Rogers. 2021. EvalMe: Exploring the Value of New Technologies for In Situ Evaluation of Learning Experiences. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445749

1 INTRODUCTION

Asking students to reflect on their learning experience and their understanding of pedagogical content is an integral aspect of the learning process. Reflection can help students to think more deeply about what they are learning, assess their own understanding of what is being taught, and simultaneously, help instructors understand whether the pedagogical methods they are employing are effective. In practice, reflection in learning environments, like classrooms and educational workshops, is often done through asking students to fill out surveys or post-tests at the end of a session by themselves, or through assessing their understanding of a topic with questions asked through classroom response systems (e.g., clickers). However, surveys and post-tests can miss important aspects of the learning experience, as children can forget how they felt or what they learned at specific points in time (e.g., [5, 22]). Moreover, children are not often given the opportunity to discuss together how their experience and understanding has changed throughout the course of the learning session. Having the opportunity to communicate what they have experienced with peers, can provide them with different perspectives on their learning experiences, especially relative to others.

Might there be alternative approaches, that are both lightweight and enjoyable to use, and which can support more in-depth self-reflection and shared reflection amongst the participants who provide the data? If so, what are the practical and pragmatic considerations of using them in real learning environments? To address these questions, we report on the Research through Design process we followed when developing a new, tangible in situ evaluation tool, called EvalMe. The tool was designed for children 10 years of age and older to: (i) provide a means for lightweight data collection throughout a learning session, rather than just at the end and (ii) make the data collected transparent to those providing it to give them an opportunity to reflect on it together. Importantly, we wanted to build on the potential of tangible interfaces for engendering shared interaction and reflection while also providing a more playful experience that could facilitate reflection in the moment and collectively afterwards. Furthermore, a goal was to support shared reflection as a two-way process, where teachers/instructors and students can discuss together specific activities and experiences encountered throughout a session.

This paper reports on our Research through Design approach (Figure 1). We describe how we initially conceptualized EvalMe, then how we developed and adapted it for use in two different educational workshop settings through collaborating with different stakeholders, and finally how we evaluated it. Our contribution is two-fold. First, we describe the contextual factors that are important to consider when designing new tangible evaluation technologies that are intended to be used in classroom and workshop settings, and by doing so contribute to research on the use of tangible devices in the context of evaluating learning experiences. Second, we describe the expected and unexpected benefits of using a more transparent and tangible interface as a means of evaluating and reflecting upon participants' experiences. Using this lens, our paper explores what can be gained from opening up the design space to enable new forms of technologies to be used, in order to evaluate both the success of learning activities and children's experiences when engaging with them.

2 RELATED WORK

In this section, we discuss research on reflection and playfulness in children, classroom response systems and tangible data input devices. This provides the basis for informing our approach of designing a tangible system for *lightweight, enjoyable data collection for classroom and workshop settings, which also supports children in reflecting on their experience individually and together.*

2.1 Reflection and Playfulness In Children's Learning

Reflection in learning can be defined as an "*activity in which people recapture their experience, think about it, mull it over and evaluate it*" [4]. Reflecting during the learning process can serve a diversity of purposes for the learner, including gaining a deeper understanding of learning material, understanding how one learns best, as well as self-development and empowerment, more broadly [10]. Reflection can be either individual or shared [18]. Through individual reflection, a learner might assess whether the approach they are taking is successful, and what steps to take to improve how they are learning. In turn, shared reflection, involving discussing experiences with peers or an instructor, can expose learners to the perspectives of others and by doing so, help them challenge their assumptions. It is especially important for young learners, given that the social context of learning plays a central role in how children make sense of the world [31]. Moreover, shared reflection can also benefit instructors, by informing them of their students' progress and helping them evaluate the efficacy of their pedagogical approach [16].

However, it is known that reflection does not always come naturally, and needs to be encouraged [10]. This is especially true for students, as it has been suggested they tend to wait for a teacher to initiate the evaluation or discussion of a learning experience, rather than doing it themselves [18, 20]. This challenge opens up the design space for thinking about whether various kinds of technology could help support both individual and shared reflection in learning. If so, how? Researchers have explored how technology can support specific dimensions of reflection, for example, how it can be used to encourage learners to revisit content, prompt explanations, or even challenge their assumptions or perceptions [10]. Approaches to doing so have varied widely; from verbally answering reflective questions during a learning session using classroom response systems [16], to writing responses to reflective prompts in virtual learning environments [26], through to receiving peer feedback about a shared story online [17].

A design principle that has been demonstrated to be effective for fostering reflection in children is playfulness. Play is viewed as being central to how humans engage with and learn about the world, and also fosters curiosity [13]. For example, Rogers et al [29, 30] pioneered playful pervasive computing – where a diversity of mobile and tangible technologies was designed to enable children to move in and out of overlapping physical, digital and communicative spaces to facilitate their understanding and reflection in situ. These included sensing and recording aspects of the local environment while exploring it (e.g., measuring pollution levels); looking up information using augmented reality that is of relevant interest when wandering through a city centre (e.g., historical sites) and texting and sending photos to others of what is being observed when in different parts of a physical environment. In the domain of personal informatics, researchers have also demonstrated how technologies that have been designed to encourage playful competition between children, for example, when running during breaktime, can help them reflect about their experiences both individually in situ and together afterwards [12, 19]. This has been used as a strategy to foster reflection about the experience of using physical activity trackers [12] as well as about more abstract mathematical concepts [19].

A benefit of these playful, experiential approaches are increased motivation and engagement that promote students' interest in their learning activities (e.g., [25, 32]). Being able to articulate what one is learning, seeing and doing in a verbal modality can also lead to a deeper understanding [8]. Next, we turn to discussing two distinct forms of technologies designed for reflection, to motivate our approach: classroom response systems and tangible interfaces.

2.2 Classroom Response Systems

Classroom response systems (CRS), which are now frequently used in both higher education contexts and in classrooms with children, typically comprise handheld devices together with screen-based visualizations. The dominant method for CRS is to ask students to respond to multiple choice questions during a learning session [16], often as a means of formative assessment of target learning outcomes [21]. Most commonly, questions asked with CRS are related to factual recall, conceptual understanding or knowledge application [21]. After all students have answered a given question individually, the aggregated data that they have entered is displayed for all in the classroom to see and then used by instructors to support discussion about a given concept or topic. Teachers can also use the data later to gain insight into whether the students understand the learning material [21]. The benefits of using CRS in this way include: supporting sustained attention; improving students' engagement with learning material; encouraging discussion about the learning outcomes [3] and enabling anonymity so that students can engage without fear of being judged [16].

However, little is known about whether and how CRS can be used to support students to be able to reflect on their experience of a particular lesson, beyond just their understanding of a specific question. CRS are also not designed to track how students' responses change over the course of a learning session; instead, they offer a means largely for in-the-moment reflection and discussion of a topic. This limits how they can be used; for example, they are not suitable for responding to questions that take different forms (i.e., not just multiple choice).

Currently, CRS input devices take the forms of either purpose-built clickers (e.g., [7]) or apps for smartphones or laptops (e.g., [24]). They provide a ready-to-hand approach to data entry, meaning that the technology itself remains at the periphery of attention rather than being present-at-hand (see [23]); the act and means of inputting the data are not framed as an experience in itself, e.g., as a means of engendering playfulness, shared attention or reflection about the questions being asked. Next, we turn to related work in the domain of tangible interfaces, to motivate the potential of other forms of technologies for promoting a more "present-at-hand" experience.

2.3 Tangibles Interfaces for Reflection

A variety of tangible devices have been developed as a way of eliciting reflection from people after an event or learning experience. The novelty of coupling physical activity with digital effects has been found to support reflection in children [2, 23, 29]. They have also been found to support playfulness in learning, which in turn sustains interest and creativity [29]. The tangible questionnaire, VoxBox, was designed to encourage people to be playful when giving feedback and reflecting on various aspects of an event, proving to be highly effective at eliciting a wide range of responses [14]. A similar kind of tangible device, called SmallTalk, was specifically developed as a tangible survey system to capture what young children thought of a live performance that they had just seen [11]. A set of colourful tangible boxes were designed which the children approached to answer a set of questions about a theatre performance they had experienced - through tangible interactions with a set of buttons, dials, and spinners. It can be difficult obtaining feedback from young children about something they have just experienced, but SmallTalk was able to put them back into the production and helped them talk about specific experiences, suggesting that its properties were able to encourage even young children to reflect, remember and think more deeply in situ.

One possible way that these forms of tangible devices for gathering responses encourage reflection is by slowing down interaction [1]; the process of playfully interacting with intriguing sliders and dials compared to, for example, quickly ticking a box on a paper when answering a question, might afford more time for users to consider their responses. The principle of deliberately slowing down interaction by embedding a device with surprising or thought-provoking modes of input and output, has long been used to encourage reflection more broadly [27, 28].

Tangible interfaces, therefore, can make the process of responding to questions engaging, playful and helpful for scaffolding discussion and reflection. They have also been demonstrated to be noticed from afar, resulting in a "honey-pot effect", enticing people to interact with them. They become objects of curiosity; where users are intrigued by their functionality, especially when they see others also using them [6]. Researchers have capitalised on these benefits to transform traditionally 'dull' evaluation and feedback tasks into ones that are more engaging and inviting. For example, VoxBox [14], mentioned earlier, was designed to attract passers-by by being inclusive and approachable, leveraging well-known affordances of buttons, dials, and other everyday input devices. In another project that explored alternative form factors for answering questions, and how they can encourage reflection, Jennett et al. [15] developed an installation of "Squeezy Green Balls" which provided playful 'stress balls' that people squeezed in relation to how they felt about specific environmental issues. The evaluations of both VoxBox and the Squeezy Balls, demonstrated how their novelty together with the easy interaction they afforded, was able to encourage participation, and led to much discussion by passers-by about the questions being asked.

Many of these tangible devices have been used in "one-off" sessions, typically after an event. Here, we were interested in whether tangible devices could be designed and used to elicit reflection after different learning activities during a multi-hour workshop, in order to get feedback about specific activities, rather than only a generalised summary of the session. This way the instructors and students can have more entry points into a learning experience by which to reflect on. However, increasing the frequency of using tangible devices may give rise to a tension, whereby it can become disruptive to the ongoing flow of the session, and thereby distracting the students. The question we address in our research, is whether and how playful modes of tangible data input can be designed for classroom contexts to promote reflection without being distracting from the ongoing learning activities. More generally, the goal was to investigate whether a tangible device could elicit reflective responses from children about their learning experiences throughout a class activity – be it a workshop or other – that they could discuss with their peers and instructor – so it becomes more of a form of communication rather than being only used primarily as an evaluation instrument.

3 RESEARCH THROUGH DESIGN APPROACH

The work presented in this paper follows a Research through Design (RtD) approach, which broadly defined refers to "practice-based inquiry that generates transferable knowledge" [9]. Rather than being a solutionist approach where the core end goal is a designed artefact that solves an articulated problem, RtD, at its core, involves the practice of continuously reframing and reinterpreting the problem space through the process of making and reflecting on artefacts that serve as potential solutions [33]. In this way, the very process of RtD is a means of better articulating the problem space and deriving new knowledge to inform potential designs that transform the world towards a "preferred state".

Here, we chose to employ an RtD approach as a means of gaining insight into the problem space of novel evaluation methods for workshops and classrooms. In line with this conceptualization of RtD, the goal of the research was not just to develop and introduce a new technology artefact, but also to provide new understandings of the challenges of carrying out in situ evaluation in workshops and classrooms, and in this way, to inform the design space for new technologies that aim to meet this challenge. Through this approach, we aimed to achieve a 'preferred state' where students would be empowered and motivated to reflect about the learning process throughout a learning session individually and together, as well as to make the process of doing so useful and easy to facilitate for their instructors.

The RtD process described here comprised four phases: (i) the initial conceptualisation of our tangible device, EvalMe, within our research team; (ii) collaborative work with stakeholders to understand their perceptions of the design concept, when co-defining EvalMe's functionality and changing the design; (iii) the development of EvalMe and (iv) the evaluation of EvalMe in two different workshop settings. Each of the phases led to distinct insights about the problem space and reformulations of design principles and opportunities. It also allowed us to gather data from a range of sources (instructors and children) in order to gain a deep understanding of how it was used. Below, we present each of the four RtD phases, together with the findings and insights they gave rise to.

4 INITIAL CONCEPTUALISATION AND COLLABORATIVE WORK

In this section, we report on the first two phases of our RtD process, that is, formulating the initial design concept for EvalMe, followed by co-defining its functionality with two groups of collaborating partners who are experts in educational workshops for children. The outcome of this section is to define core requirements for the form factor and functionality the envisioned EvalMe system, together with practical considerations for how it might be used to best effect in workshop settings.

4.1 Formulating the Initial Design Concept and Core Requirements

The starting point for designing EvalMe was to arrive at initial design criteria, by reflecting on the challenges we ourselves have faced, when evaluating new artefacts and pedagogical methods with children. Often, within our research and practice, we have found most evaluation methods inappropriate for our purposes. For instance, when we have wanted to gather reflective feedback from children about specific aspects of learning activities during a learning session, we have found asking them to fill out a survey can be obtrusive and distracting. At the same time, we have found that the richness of children's in-the-moment reflections can be lost, when they are asked to fill out a survey or otherwise reflect on their learning, after a session has finished. Equally, while we frequently use qualitative approaches including observation and video-based analysis of interaction, these fall short in terms of understanding children's own perspectives on their learning. To overcome these shortfalls, we discussed how to come up with a new method, that would provide a less obtrusive means of data collection during a learning session, and that could be reflected upon by the participants later through group discussion.

As a starting point, all the authors met to discuss related work and how it informs potential solutions to the given problem space. The goal of this meeting was to derive a set of criteria for a system that might be successful in enabling students to reflect about the learning process throughout a learning session and making the process of doing so useful for their instructors.

We first discussed the body of work on CRS. We found they are typically used as a means of formative assessment, while being less well suited for encouraging children and students to reflect on their experience, and on how it develops over time. We discussed what other types of methods for asking questions could be used—from those related to learning outcomes (as are most often used with classroom response systems) to those related to the experience of a workshop or class session (for example, self-reported levels of interest and engagement). We also discussed the potential of designing a tangible artefact that could gather data in a visible way. Based on previous work with tangible interfaces for supporting reflection [11, 14], we agreed that this form of interaction has the potential to encourage more curiosity while enabling the process of inputting data to be either private to an individual or public to others around them. We also considered based on the positive findings of previous tangible user interface research, that using a tangible device, instead of a purely screen-based one, has the potential of making the experience of 'evaluation' more playful, encouraging participation, and fostering discussion.

Thus, based on our discussion and reflections, as a starting point we produced a set of criteria for the design of an in-situ evaluation system. These were to:

1. Devise a system that would enable lightweight data gathering at different points in a classroom or group workshop session.
2. Make the system tangible and able to be passed around the classroom at different points in time.

3. Ensure that the process of submitting data could be either private or public—in terms of how visible it was to others around the person submitting the data.
4. Present the data collected about the children's learning experience to both the instructors and to the children; in doing so, enable the children themselves to reflect on their own learning and elaborate on their data through group discussion.

These criteria informed the development of the form factor and functionality of EvalMe, which was decided to be: (1) a tangible, hand-held device, that would be able to be passed around at different points in a learning session and would be flexible in terms of the questions it asked of the children/students together with (2) an application to visualize the gathered data and present it back to the participants. However, the envisioned system needed to be suitable and accepted by practitioners—i.e., educational workshop leaders and class instructors—to use as part of their practice. Before developing our system, we discussed the initial design ideas with two groups of collaborators, who were experts in designing and leading educational workshops. They were able to provide us with different perspectives on teaching in workshops and the challenges with evaluating them.

4.2 Developing the Concept with Collaborating Partners

We involved two groups of collaborating partners in the design process and held two collaborative meetings with each group. Both groups were leading educational workshops with 12–14-year-old students and were enthusiastic about trialling our proposed tangible system in their workshops. We chose these groups for their interest in the research domain, and also because they were interested in trying out our system in their workshops in a variety of schools, making for a diverse population in which to test the system. This process informed aspects of the design of EvalMe, as well as providing insights into the practical considerations of deploying the envisioned system in real workshop settings.

4.2.1 Context. Both groups of collaborating partners, who we call G1 and G2, comprised practitioners in the arts and in education whose work entails using innovative methods and novel technologies to teach computing to children in school settings. Both had substantial prior experience with running and evaluating workshops with children. Both also were planning to deploy new workshops to introduce computing topics using creative pedagogical methods and a range of materials including physical computing artefacts. Their workshops were planned to be held with 12–14-year-old students within schools, but outside of the students' typical classrooms. Thus, the workshops were expected to give rise to different implicit social rules compared with being used in a typical classroom setting. While the workshops themselves were a platform for designing and studying EvalMe, rather than the objective of the research, it is important to provide detail about their structure and aims to contextualise the RtD process and outcomes. In Table 1, left, we provide a summary of G1 and G2's planned workshops, in terms of their aims, structure and their evaluation goals.

In sum, G1 planned to introduce students to computing topics through the medium of dance, in line with the English Computing Curriculum. The instructors' goals for trying out our in-situ evaluation system were to assess whether, and to what extent, the pedagogical methods they adopted improved the students' understanding of the target computing concepts. G2 planned to use crafts and physical computing toolkits in their workshop and to use EvalMe to enable students to reflect on the data they share about themselves, as well as being able to view data as a creative material which can be experimented and played with. The instructors' goals for using EvalMe, therefore, were largely to evaluate the students' experience of the workshop, and to determine whether it was able to trigger changes in their perceptions of data over time. In sum, the key dimensions in which G1 and G2 differed were the extent to which their workshops planned to contribute to specific learning outcomes (i.e., the workshop led by G1 had explicit links to a national computing curriculum) and relatedly, their evaluation goals (i.e., G1 had the goal of evaluating the extent to which the workshop contributed to the students' understanding, while G2 had the goal of evaluating their experiences and perceptions of data).

4.2.2 Collaborative Meetings. We held two meetings with each group of collaborators to introduce the EvalMe concept, gather feedback from them and to plan how the envisioned EvalMe could be used in their workshops. The first meeting comprised a discussion to understand the workshop goals and the instructors' needs for the in-situ evaluation system. In the second meeting, we began by describing the design concept to the collaborators, and together iterated the design concept to adapt it to their workshop needs. Specifically, we asked the collaborators to discuss in detail how they would envision using the proposed system within their workshop, taking detailed notes of their decisions and perspectives, and proposing how the system could be adapted to suit these. The meetings were not rooted in specific participatory design methods, but were researcher-led, structured conversations. We met with each group of collaborators separately, to enable them to reflect on how the envisioned system would work in their context, without being influenced by the perspectives of the other group. In the next section we provide a descriptive account of what was discussed and how it influenced the envisioned EvalMe design.

4.3 Findings from the Collaborators

Important concerns that were raised included the extent to which our collaborators would be willing to dedicate attention to EvalMe during the workshops; the types of questions they would value asking of the students, and at what points in time; and when best to consider presenting and discussing the collected data with the students.

4.3.1 Ensuring the System Does not Divert the Instructors' Attention. A topic of discussion was whether using a new artefact for data collection would be distracting for the instructors leading the workshop. A concern for both groups was that the planned workshops would already be using various technologies—e.g., physical computing artefacts—alongside other paper and craft materials. Therefore, it was important to ensure that the instructors' attention was not overshadowed by the novel EvalMe device and was able to remain in the periphery of the instructors' attention throughout a session. We agreed that one of the ways to ensure that using EvalMe would

Table 1: The left side of this table describes the context of the workshops led by G1 and G2, focusing on their aims, set up, and evaluation goals. The right side of the table details how EvalMe was adopted in practice, including the questions that were asked of the students, and the contexts in which they were asked.

Group/ WS	Workshop Context and Goals of Using EvalMe		Use of EvalMe in Practice	
	Workshop (WS) details	**Goals of Using EvalMe**	**Configuration**	**Answer Rounds**
G1	**Workshop leaders:** A choreographer and a computing education specialist **Aim:** Introduce students to computing topics through dance, in line with the national computing curriculum. **Material/Activity:** Paper-based activities, dance, sensor-based physical computing artefacts **Duration**: 5 hours **Participants**: 16 (12-14 years old)	Assess whether, and to what extent, the pedagogical methods improved the students' understanding of computing concepts; gain more granular insight into which methods of instruction work best.	Questions asked: Q1. I have a strong understanding of this topic Q2. The movement has helped to increase my understanding of this topic Q3. The interactive tools have helped to increase my understanding of this topic Authenticated Mode: Yes	1) At the beginning of the workshop (only answered Q1) 2) During a break in learning activities 3) Coming back from lunch 4) At the end of the workshop.
G2-W1	**Workshop leaders:** A theatre producer and a creative technologist **Aim:** Enable students to reflect on the data they share about themselves, and to view data as a creative material which can be experimented and played with. **Material/Activity:** Physical computing blocks, craft materials, an iPhone app, paper-based materials **Duration**: 2 hours **Participants**: 15 (12-14 years old)	Evaluate the students' experience of the workshop and whether the pedagogical methods helped the students view data as valuable, and as a creative material.	Questions asked: Q1. I'm having fun! Q2. Data is something I can play with Q3. Data is like. . . [marble, plasticine, chalk, paint] Authenticated Mode: No	1) At the start of the workshop 2) After a group discussion 3) After first being introduced to the physical computing kit 4) At the end of the workshop
G2-W2	Second instance of G2-W1 with 30 participants	Same as G2-W1	Same as G2-W1 with Q3 changed to [On a scale of 1-12] "Information about me is valuable"	Same as G2-W1

not require substantial time and effort for the instructors during the workshops, would be to design it so that it could be configured by them before the workshop took place. This led to us formulating a new requirement for EvalMe: the tangible artefact, the questions to be asked, and the means of visualizing the data should all be configured before the session.

4.3.2 Deciding on the Questions. After deciding that all the questions to be asked of the students should be configured before a session, we considered how EvalMe could be designed to best support this. G1 initially considered whether it would be possible to ask the students different questions at different points in time. However, through further discussion it was agreed that changing the questions during a session might be more difficult to configure and increase the amount of time the instructors would have to allocate to EvalMe during the workshops. Therefore, in both groups, it was decided that a good strategy would be to use the same questions throughout the workshop. While this reduced the flexibility of the questions asked at different points in time, it also reduced the amount of time the instructor would be required to spend themselves, configuring the interface between sessions.

4.3.3 What Kinds of Questions to Ask? The decision to use the same questions throughout the workshop at different intervals in time, meant that the collected data would be able to show any changes in the students' responses over time throughout the workshops. This raised another challenge: what questions to choose to ask that would be informative in terms of the evaluation aims. We suggested to both groups that considering asking three questions might be a good number, as this would enable different questions to be asked, but also not be too time consuming for the students to complete. We also encouraged them to think about different response formats—for example, multiple choice or Likert scale responses.

G1 suggested for their workshops evaluating whether specific learning outcomes were fulfilled and whether the pedagogical methods employed were suitable. Their workshop was subdivided into three different topic areas. They articulated that they would like to evaluate the students' understanding of each topic, but also, to gather a sense of how their overall understanding of all the topics changed between the start and end of the workshop. They decided on three general questions, using scale-based responses, that would ask students to reflect about their understanding of a given topic (see Table 1, right). Phrasing the questions to include the word "topic", rather than specifying the topic that they were referring to, provided some flexibility to the questions, without requiring the instructors to reconfigure them during the workshop.

G2, in turn, decided to ask two questions that were intended for students to reflect on their experience of the workshop and perception of data, which they could respond to using a scale (see Table 1, right). They also decided to experiment with a more open-ended final question, which was to ask the students to compare the data collected to suggested creative materials (e.g., water, air). The reason for this was that they hoped the question would trigger a more creative discussion at the end of the workshop about what data means to the participants.

4.3.4 When to Collect the Data? Balancing the Value of the Data with Potential Distraction. Much time was spent in the collaborative meetings discussing at which points to gather the data from the students. The discussions centred around balancing how much data would be valuable, with how distracted the students could be expected to become by EvalMe. It was pointed out that because EvalMe would be new to the students, we expected there to be a novelty effect when first introduced. Specifically, we expected that upon encountering it for the first time, each student would spend some time looking at it and exploring it before submitting their response. Both G1 and G2 decided that the strategy that they would use to reduce this, would be to familiarise the students with the device at the beginning of the workshop by explicitly introducing it to them and enabling each student the chance to submit responses to the questions.

Beyond agreeing on how to collect the first set of data, G1 and G2 had differences of opinion about the contexts in which the subsequent sets of responses should be gathered. This was tied to the extent to which they were concerned that EvalMe would be distracting to the students' learning. G1 felt that even momentarily taking attention away from the learning activity to submit responses might be detrimental and decided that they would pass EvalMe around during the planned breaks (morning, lunch, afternoon) set for the workshop. In contrast, G2 was not worried about students being distracted when using it; they decided to pass the device around in parallel with other ongoing activities.

G1 decided that they would use EvalMe sparingly, after each learning session had been completed. In contrast, G2 decided that they would try to collect data with EvalMe "as much as possible" at regular intervals, without planning specific timings for doing so. They chose to do this because they thought that collecting a high number of sets of responses from the students would lead to more granular temporal data and were interested in seeing how this would help them reflect on their workshop structure.

4.3.5 When to Present the Data? We had initially envisioned the data that was collected being displayed using a projector onto a wall throughout the whole of a workshop, assuming the instructors could use it when wanting to ask the students to verbally reflect on the evaluation data collected. However, in contrast to our proposal both groups decided that they would reveal the collected data to the students only at the end of the workshop. The reason for this was that they were worried that if the data was constantly being displayed it might distract both themselves and the students from the other learning activities being carried out.

4.3.6 Anonymity of the Responses and Granularity of Revealed Data. Much discussion ensued about how to track how each student's responses changed over time, and to balance this with the anonymity of the responses. G1 and G2 considered it important that the data presented back to the students would not reveal their identities to the rest of the class; however, for the purposes of their own analysis, they also wanted to be able to track how each student responded in different answer rounds. Based on this, we agreed to design EvalMe so that instructors have the option of collecting pseudonymous data, using RFID tags. While G1 initially asked if each ID could be mapped to a student's name, we postulated that the practicality of setting up each RFID tag to correspond to the students' names at the beginning of a workshop would be too time-consuming. It was also pointed out that not being anonymous anymore could potentially make the students feel less comfortable with providing negative responses as the instructor would be able to see who had given them.

G1 were also concerned about how the data could be presented back to the students while ensuring their anonymity. They did not want them to feel singled out, especially when asked about how well they understood certain topics. We suggested designing visualizations that would present only aggregated data, for example through bar charts. However, they pointed out that even if the data was aggregated, there might be instances when students with outlying data might feel singled out. For instance, if only one student in a group indicated that their understanding was very low in response, e.g., by submitting the value of "1" on a number scale, while most others indicated that they had a high level of understanding, this could lead to the student who responded with the lowest score feeling uncomfortable being the only one who did not understand what was being taught. To overcome this dilemma, it was decided that for rating scale responses, the students would be able to submit from a wide a range of responses (e.g., using a 1-12 scale), but that the data displayed back to them would be less granular - showing only ranges, for example as "low", "medium" and "high" levels of understanding.

5 EVALME: SYSTEM OVERVIEW

Based on the discussions and concerns raised, we began generating ideas for a tangible system that would fit our initial criteria and be appropriate for the collaborators' needs. After a number of iterations, we designed a device that could collect and show user data using three interconnected components: 1) *EvalMe Sender*: a tangible, hand-held box that could be passed around the classroom, allowing students to respond to a small number of questions; 2) *EvalMe Receiver*: that could be plugged into the teacher's computer via USB to wirelessly receive the responses; and 3) *EvalMe Desktop*: an application running on the computer, that could collect and visualize the received data. The circuit diagrams, CAD models and code—for both the tangible components and the node.js EvalMe

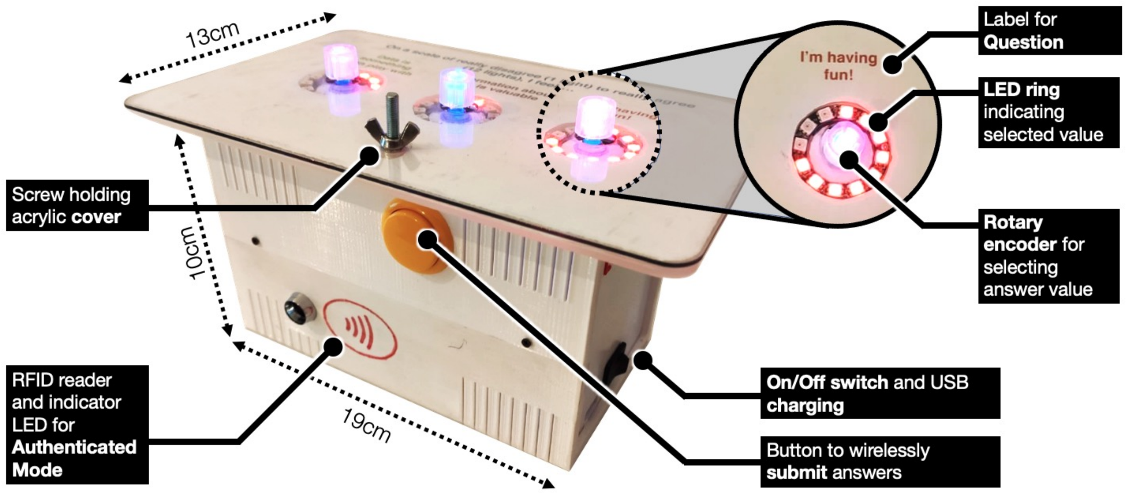

Figure 2: The EvalMe Sender is passed between students to collect answers from students to a 3–question questionnaire. Answer types can be categorical or numerical. LEDs indicate the selected response.

Desktop application—are openly available in the project's GitHub repository at https://github.com/frederikbrudy/EvalMe.

5.1 EvalMe Sender

EvalMe Sender consists of a custom-made 3D printed box, intended to be passed around a room whenever the instructor wants to collect responses. The Sender contains an Arduino Uno microcontroller powered through a rechargeable Li-Ion battery (Figure 2). On the top are three rotary encoders (without a fixed start or end point) which allow for flexible increase/decrease by rotating them. These are surrounded by 12 LEDs to provide feedback as to which dial had been selected. Students can then rotate each rotary encoder to select a response for the question written above it. As noted, G1 wanted to use scale-based responses while G2 wanted to use both scale-based and multiple-choice responses. Using the rotary encoders with LEDs, allows for flexibility in answer types, i.e., questions can either ask for a categorical (e.g., multiple choice) or numerical response (e.g., rating scale), corresponding to the 12 LEDs.

5.1.1 Instructor Set Up And Use. To prepare the Sender for a class, the instructor needs to set the questions and the answer options in advance using the Desktop app (described shortly). They then print a paper template with the questions and the answer options on it and place this on top of the Sender. G1 and G2 had both stated that they wanted to focus on a single set of questions during a session. For different workshop/teaching needs, we also envision multiple templates that could be easily exchanged during class. An instructor can initiate multiple answer rounds for the same questionnaire throughout a learning session, using the Desktop application (see later). An answer round marks a period in the class when each student gives their individual responses to the three questions.

5.1.2 Interaction by the Student. When interacting with EvalMe, students turn the encoders to select a response. Once satisfied, they press a button on the front to submit their answer wirelessly to the EvalMe Receiver; to avoid biasing the next student, the selected answer is immediately reset, meaning the circular LEDs turn off. They then pass the Sender on to the next student.

One of our requirements was to design the interface so that the process of submitting a response could be made either private or public. We therefore positioned the questions on top of the box, with the answer LEDs slightly inset to the frame. We chose to place the rotary encoders on the top of the box, so that the selected answer is only visible when a person leans over the box, allowing the student to stay in control of who can see the answer or whether they want to keep it private.

The Sender can work with or without requiring students to authenticate their identity. When working in *Authenticated Mode*, each student is given an individual RFID enabled card at the beginning of class. They then use this card to "log in" by tapping it on the RFID symbol on the side of the Sender before they can give their response. The RFID cards serve as unique identifiers but are, by design, pseudonymous; they are not associated with students' names or other personally identifiable information. In the *Authenticated Mode*, a student cannot submit an answer without tapping the RFID on EvalMe, and during each question round only a single answer can be submitted per student. As discussed earlier, we included *Authenticated Mode* as G1 wanted to enable instructors to understand development of *individual responses over time.*

5.2 EvalMe Receiver

The EvalMe Receiver is the counterpart of the Sender box. It is plugged into the instructor's computer via USB and collects the submitted answers wirelessly (through a low-power wide-area network—LoRa) for the EvalMe Desktop app. As with the Sender, the Receiver consists of an Arduino Uno microcontroller with a LoRa shield for the wireless communication. Using a LoRa network allows for long ranges (signals easily pass through a 5-story building), without requiring network infrastructure (unlike WIFI),

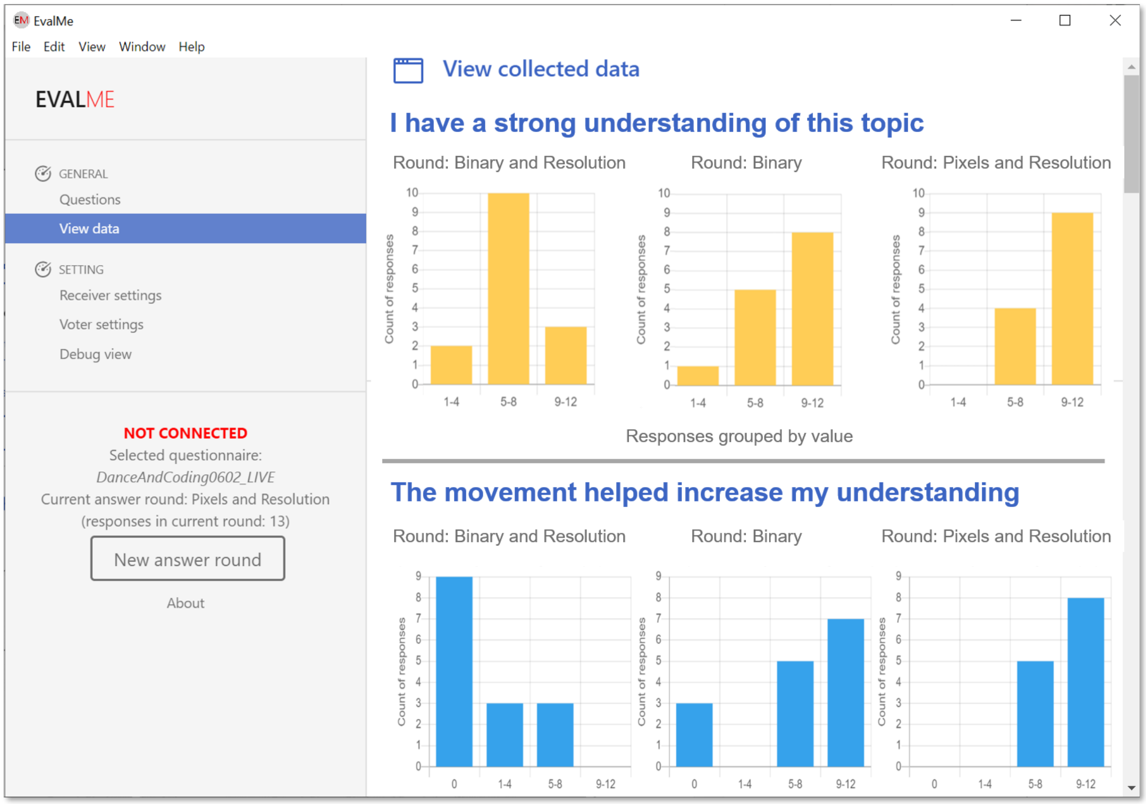

Figure 3: The EvalMe Desktop application visualizes the collected data for each question-and-answer round in bar charts. The visualization shows the data that was collected in G1, where each round took place after a particular workshop part.

or a pairing step (unlike Bluetooth). The Arduino's software is programmed to forward messages received through the LoRa shield via USB to the computer's serial port, and to send control messages it receives from the computer via LoRa to the EvalMe Sender. Sending messages to the Sender is used to change the settings for requiring authentication and the type of answer options.

5.3 EvalMe Desktop

The EvalMe Desktop application (Figure 3) was developed using node.js and the Electron framework. It is used to (i) create a new set of questions, (ii) record responses for an answer round; and (iii) visualize responses for the use of the instructor and optionally for discussion in the learning session. When starting the application, the software automatically connects to the Receiver via a serial connection. The instructor can then switch between different sets of questions listed on the app or create a new one. Each response is associated with the current *answer round*. As mentioned previously, an answer round is used to group responses to compare development over time and can be named by the teacher, for example to indicate a particular activity.

The Desktop software was also designed to be used to control the parameters for the Sender: whether or not to require authentication as well as the maximum value for selection using the rotary encoder (ranging from 3–12). The number of LEDs that light up for each rotary step is then adjusted to be in line with the maximum number that can be selected. For example, if the maximum selectable value is set to 3, each increment step lights up 4 LEDs. New settings are sent wirelessly to the Sender, once received all LEDs on the Sender flash green to confirm that the settings have been saved.

As a starting point, we designed a bar chart visualization for the Desktop application to visualize the data, in which each answer round is represented in the form of a canonical bar graph (Figure 3). We decided on this format to convey the answer data for each question as we considered it simple and hoped it would be understandable by children 10 years and up as well as instructors. Consecutive answer rounds are placed side by side, to enable the viewer to see how responses changed over time (Figure 3). As discussed previously the granularity of data displayed often needs balancing with anonymity. Therefore, the Desktop application enables instructors to select values to be grouped together in the visualization. This can be seen in Figure 3, where values of 1–4, 5–8 and 9–12 are grouped together. This option ensures instructors have granular data to analyse and reflect on, but that the data presented to the students obscures individual responses. Raw and aggregated data can be exported as a spreadsheet if further analysis or sharing is required.

When using *Authenticated Mode* during response collection, instructors can select an individual user from a dropdown menu and analyse their responses over time.

6 DEPLOYING EVALME

Having built and tested EvalMe we then deployed it in 3 workshops, 1 of which was led by G1 and 2 of which were led by G2. The purpose of doing so was to understand, in detail, the patterns of interaction that would occur when using EvalMe in educational workshops, focusing on the benefits and limitations of using EvalMe in G1 and G2's specific workshop contexts. The method chosen for this was to observe how the children interacted with EvalMe during the workshops, and to subsequently interview the workshop leaders about their own observations and perceptions of EvalMe.

6.1 The Workshops

The workshops were held in 3 different schools (G1, G2–W1 and G2–W2). While both G1 and G2 had planned to use EvalMe in more workshops, these were cancelled due to the onset of COVID–19 in March 2020. One device was used during each workshop for 4 answer rounds.

6.1.1 The Role of the Researcher. During each workshop, a member of the research team was present and set up the software and hardware together with the workshop instructors (i.e., the collaborators). The researcher observed the workshop from the side without active intervention; they only controlled the desktop application. The workshop instructors ensured that EvalMe was passed to every student as part of the class activity. At the end of the workshop, the researcher also interpreted the data visualizations for the students.

6.2 Data Collection And Analysis

At each of the three workshops, the researcher who was present wrote down observations of how the students interacted with the EvalMe Sender at different points in time. As only one EvalMe was passed around in each session, the researcher was able to observe each student's interaction with it and the context in which this occurred. At the end of each workshop the researcher interviewed the G1 and G2 about their perceptions of how well EvalMe had worked, and discussed with them their written observation notes, asking them to share any other observations they had made. Through this process, the observations were expanded and developed into more expansive, descriptive accounts. The research team then met to discuss these descriptive accounts and collectively analysed the findings in terms of the pros and cons of EvalMe with respect the preferred state. Specifically, we discussed the descriptive accounts through the following questions: to what extent, and in what contexts did EvalMe enable the students to reflect upon the learning process? Was using EvalMe useful and easy to facilitate for the instructors? In sum, our approach is to provide a qualitative, descriptive account of what took place at the workshops, which triangulates the researchers' perspective with that of the workshop leaders.

6.3 Findings

When EvalMe was first introduced, there was much excitement and curiosity observed amongst the students. They spent time looking in detail at the components and repeatedly turning the dials, before submitting their first round of responses to the questions. During each subsequent answer round, the students became more aware of the need to pass it on to the student next to them. Having one EvalMe box was sufficient for groups of 15–16 participants. However, for G2–W2, which had 30 participants, it involved waiting that was time-consuming. We noted that in this case, for each of the four answer rounds, the act of passing EvalMe around took up to 8 minutes; in two instances it was perceived to be taking too long when switching between learning activities.

Next, we present detailed findings for a set of identified themes: (i) the extent to which using EvalMe fitted into the workshop structure; (ii) how it gave the students the opportunity to choose between hiding and sharing their responses from/with those around them; (iii) how the students were found to use it tactically as a break from learning; (iv) the extent to which the students were able to interpret and discuss the data that was being collected and (v) the perceived value of the collected data for the workshop instructors.

6.3.1 Taking Turns to Pass EvalMe Around. The students took it in turns to pass the EvalMe Sender around to the next student in the classroom — with the last student in a group getting up to take it to the next table after asking if everyone at their table had already used it. However, at times, in all three workshops, it would get "stuck" at a particular table or with a particular student. In these instances, it was up to the instructor to nudge the student to move the box along. An issue that was encountered when EvalMe was being passed around while other learning activities were taking place, was that this nudging momentarily diverted the instructor's attention away from teaching. In contrast, when EvalMe was used between learning activities this was not found to be an issue.

6.3.2 Submitting Responses Privately vs. Publicly. Both "private" and "public" uses of EvalMe were observed. Some students chose to submit their responses by shielding it from others, tilting the input face of EvalMe towards their bodies when turning the dials so as to obscure what they were choosing from those around them. Other times, more "public" responding was evident. This was most frequent, in G1, when EvalMe was used during the breaks. Groups of students were observed to crowd around together and when it was their turn, talked aloud with their peers when inputting their answers—which in turn prompted them to discuss together what they had learned. This was also observed to an extent in G2-W1 and G2-W2. However, in the setting where EvalMe was passed around during a learning activity, the amount of discussion that the students engaged in about their responses was more limited. A question that brought about much shared discussion in G2-W1 was when comparing data to other creative materials (i.e., "Data is like. . . [marble, plasticine, chalk, paint]"). For example, one student asked their peer what they thought the question meant, and another talked through their reasoning for choosing "chalk". In sum, using the EvalMe device between learning activities promoted more talk amongst the students where they reflected on the questions being asked together, than when it was passed around during an ongoing learning activity.

6.3.3 Holding on to Control: Using EvalMe to reflect vs. As a break from learning. When used during a break, the students sometimes took their time answering the questions, as if to extend the break time. Especially after lunch, we observed them taking longer using EvalMe than during other times. In G2, in contrast, when EvalMe was used in the middle of an ongoing activity, the students responded and passed the device on more quickly.

We took note of how those students who held onto EvalMe for a longer period (>30s) when it was their turn to respond, interacted with it. We found that most who did so seemed to be focused — reading the questions, hesitating before inputting their answers, and carefully twisting the dials to make sure that the number of lit lights was the final value they wanted to submit.

In a small number of instances, some students were observed to use EvalMe as a reason to divert their attention, away from the learning at hand. For example, in G2-W2, a group of 3 students spent almost a full minute twisting and turning the dials and repeatedly submitting responses, until asked by the instructor to pass the device on. In G2-W1, we observed a student who did not make eye contact with the instructors and had seemed disengaged throughout the workshop (as observed by both the researcher and the workshop leader), repeatedly turn the dials and submit responses until asked to stop. In these cases, our observations indicated that the students had been already distracted prior to using EvalMe; nevertheless, it provided them with another opportunity to be distracted from the learning activity (albeit for a short period of time).

6.3.4 Reflecting on the Data After The Session: Mediating the Interpretation Of Results. The aggregated data collected over the course of the three workshops was presented at the end of each session back to the students who were asked to discuss why they responded in the ways they did, and how their responses changed over time. The students, however, were reticent to speak in public and in front of their peers about what they thought the graphs meant. When asked specifically to say what they thought had changed in the data, none of them raised their hands to respond.

To help begin a discussion, the researcher took the initiative to describe what the different bar graphs represented. For example, in G2-W1, she described how one of the graphs indicated that the students started having much more fun after a given activity. She then asked the participants to reflect on why this was the case. Similarly, in G1, she pointed to the bar chart that represented the activity which made the most difference in the students' understanding and asked them why this was the case. When the researcher noted how the answers to the question *"data is like... [marble, plasticine, chalk, paint]"* were randomly distributed during each answer round (in G2-W1), there was much giggling amongst the students. They seemed to find it funny. Then one student volunteered to say they had picked their answer at random to this question to which others nodded in agreement. This moment of acknowledgement appeared to 'break the ice' after which there was more of a discussion where more students volunteered comments, suggesting how the "creative" attributes of data compared with the other art and craft materials. Hence, the tactic of the instructor taking the initiative to get a discussion going enabled the students to feel more comfortable before explaining why a specific activity was helpful to their understanding (G1) and why their ratings changed over the course of the workshop (G2-W1 & G2-W2).

6.3.5 The Value of Data at Set Points in Time. The workshop leaders in G2 had mentioned that they initially wanted to use EvalMe *"as many times as possible"* throughout the workshop, constantly passing the device around. However, in practice they only used it for four answer rounds. A reason for this was that they realized that having it constantly being passed around did not add much to the value of the dataset. Instead, they decided that a better strategy for finding out which activities were successful was to choose key times during the workshop at which to use EvalMe to collect feedback, and to annotate the data using the EvalMe Desktop software, for example, describing which learning activity that had just taken place.

6.3.6 Value for Instructors: Rapid Turnaround of Immediate Feedback. When interviewing the workshop instructors, we found that they perceived the data collected using the EvalMe device to be beneficial in several ways. Firstly, G1 found the learning outcome-based questions useful for providing them with insight about the extent to which each of the different learning activities used for a given topic helped the students deepen and integrate their knowledge. Secondly, they noted how having immediate student feedback helped them understand whether and why a workshop was successful. Thirdly, G2, saw how ratings increased over the day for the students' responses to statements like "data about me is valuable" and "data is something I can play with" was helpful in letting them know the workshop was meeting its goals (see Table 1).

6.3.7 Value for Instructors: Reflection on Teaching. G1 and G2 also noted that the feedback data EvalMe provided them gave useful insights and reassurance about their practice of running the workshop. For example, the lead instructor from G2 noted how they observed that, after introducing physical computing artefacts to the students, their response to the statement, "I'm having fun" substantially increased on the 12-point scale. As a result of seeing this data, in the next iteration of the workshop, the instructor decided to introduce the physical computing artefacts earlier on and cut down on the introductory activities that they had originally planned for the beginning.

7 DISCUSSION

The key question we addressed in our research was whether and how playful modes of tangible data input can be designed for classroom contexts to promote reflection. The goal was to investigate whether a tangible input device could elicit reflective responses from children about their learning experiences throughout learning sessions, that they could discuss with their peers and instructor. Our research demonstrated the value of using a tangible device to collect answers to a small set of questions—enabling the students to reflect in the moment, both individually and together. The act of turning a dial to answer the questions asked of them, rather than ticking a box on a paper-based or screen-based survey, also seemed to be enchanting, allowing for fine tuning while contemplating. Why was this simple act so impactful?

7.1 The Value of Tangibility For Individual And Shared Reflection In Learning Settings

One of the reasons was that the physical form factor of EvalMe transformed the task of data input into one where the technology itself was more "present-at-hand". Turning the dial and seeing the LEDs light up when answering a question was more embodied and evocative than making a mark on a piece of paper or clicking a radio button on a screen. Moreover, it was beneficial for promoting both *individual* and *shared* reflection. At an individual level, we observed many of the students spending considerable time moving the dial slowly back and forth before finally submitting their answer — seemingly thinking about their response.

The act of passing the physical box around to other students also made the process of evaluation a shared, connected activity that the students completed together, rather than something to be done by themselves hurriedly at the end of a learning session. When the EvalMe was passed around, its physical presence provided a place around which they could congregate; this in turn prompted each to answer questions and sometimes talk spontaneously about what scale they had selected and why. A few students, however, appropriated the box to their own individual advantage, by appearing to take their time answering the questions and shielding themselves from the others when doing so, using it as a delay tactic to prolong a break.

Our findings corroborate with other research on how tangible interfaces can be used to support shared attention, and both individual and collaborative interactions (see [1, 23]). In particular, the physical presence of the artefact supports a more embodied form of action when answering a question compared with ticking boxes with a pen or when simply clicking a button for a CRS. The combination of moving a dial around 12 points on a scale and seeing the LEDs correspondingly light up or down provided a richer and more visible kind of affordance and feedback when reflecting in the moment on how to answer a question. In so doing, it could have slowed the students down as they watched their hand moving the dial and may have resulted in them giving each question more thought when answering it.

7.2 The Value of Displaying the Data Gathered In Situ

The data that was collected over the duration of the workshop was useful for the instructors, themselves, helping them to think about how to improve the pedagogical structure of their workshops. It also provided more detailed breakdown of their evaluation to share, for example, with external funding bodies - which was perceived to be very valuable to them. However, the screen-based visualisations that were displayed at the end of the workshops, were less successful at eliciting a discussion amongst the students. It seemed that the aggregate bar charts summarising the data were too 'distant' from how the students had answered them, for them to interpret them at the end of the day in a public forum. Part of their reticence to speak about the responses shown could also have been to do with the social norm of feeling embarrassed saying the wrong thing when in a public setting — the students did not know the workshop leader or researcher in the same way they get to know their teachers. Hence, when prompted by the workshop leader, it may have felt awkward for the students to speak out their interpretation of the bar charts. Only after the researcher stepped in and interpreted the trends in the data for the students, did they volunteer more to speak up and draw conclusions about their feedback data.

A question this raises is how can the questions as well as visualizations of the data be designed to be more appealing and intriguing to the students? If the first type of data to be shared with them was about something innocuous, such as their hunger levels correlated with their tiredness levels over the day, this might act as an icebreaker while also giving them the opportunity to understand the structure of the graphs/bar charts or other representation being used. Another strategy could be to ask more creative questions that challenge students to reflect on their change of perceptions over the day, for example, about the data (e.g., questions 2 and 3 of G2-W1; see Table 1). Hence, it might be useful to provide teachers with different ideas and ways of questioning the responses collected they can choose from rather than having to construct the questions, themselves.

7.3 The Research Through Design Process for Contextualizing Design

The RtD process followed helped us to reinterpret and formulate a deeper understanding of the problem space of in-situ evaluation, and to arrive at design decisions which were agreed on by our collaborators. However, as is common when conducting RtD, several tensions surfaced when presenting our design ideas to the collaborators and when they raised their concerns about our research. One example was the decision as to whether to show the participant's feedback at the end of a workshop or throughout it. Our idea was to project a shared visualization on a wall that could be looked at throughout the day by the students and instructor – with the aim of being able to trigger spontaneous discussions and reflection. G1 and G2, however, were opposed to this idea, as they thought it might distract the students too much from the ongoing learning activity. In another example, G1 asked if they could track children's identities alongside their responses, which we were opposed to, as part of our ethics approval was to say that collected data should be pseudonymous. These tensions were important for raising privacy and pedagogical concerns, that might otherwise have been overlooked. In terms of the examples provided, they highlighted the collaborators' perceived importance of maintaining children's attention to the learning activities, as well as the need to understand each child individually, rather than just as part of aggregated data.

The process of negotiation, however, was often not straightforward; while we wanted the prototype to be novel and playful, our collaborators were at times more concerned and cautious about how it could and would be used. The overarching strategy was to make the EvalMe system flexible, so that it could be adapted for a specific workshop. This involved working out how to make it "seamless" for the instructor to use and not get in the way of their pedagogical practice. Strategies for this included ensuring that the system was configured before the start of a learning session, did not call for instructors' attention during the session (for example, through the changing of questions or switching between software on a computer), and minimized the extent to which the instructors had to intervene in disruptive or distracting patterns of interaction.

7.4 The Extensibility of EvalMe to Other Contexts

While we envision EvalMe as a tool for use in both classroom and workshop contexts, we have so far only tested it in the latter setting. Classrooms are clearly different to workshop spaces; for example, in some classrooms, children may not be encouraged to speak with others during a learning activity or be encouraged to stay sat at their desks until a teacher gives them the permission to move around. Thus, it is possible that some of the specific patterns of interaction observed in this study, like taking turns to pass the EvalMe box around, or discussing responses with others while answering questions, might not be possible in a classroom setting. While our initial findings suggest that EvalMe can lead to productive in situ reflection for workshop settings, where there is more flexibility in how they are run and more time for collecting responses for reflection, it remains to be seen as to whether EvalMe would be as effective in a classroom.

8 CONCLUSION

Our Research through Design study demonstrated how to design an effective tangible device that students can use to provide feedback about their learning experiences during a workshop. We showed how transforming the act of selecting an answer from being an automatic one (e.g., pushing a clicker button) to being a more embodied one resulted in more reflection about how to answer a question in the moment. Designing an interaction to be tangible and which at the same time provides corresponding LED feedback was able to reify the selection process, transforming what is normally a cursory task into something that is more engaging, and even magical. However, simply presenting this back as aggregated data in the form of a canonical representation (e.g., bar charts) will not spontaneously trigger a discussion amongst the students; it needs the instructor to think of asking questions that are easy to reflect upon in the moment and will not embarrass the students. Changing the process of evaluation through using a tangible device and shared visualizations offers much scope for thinking differently about how feedback is collected and shared, but consideration is needed as to how much of this to reveal to the students in the presence of their teachers.

ACKNOWLEDGMENTS

This research has been made possible by the EPSRC IAA Knowledge Exchange Dance and Coding Project (IAA 2017-20) and the AHRC Creative Informatics Project (AH/S002782/1). We would like to give a profound thanks to our collaborators for their time, insight and enthusiasm: Clare Duffy and the Civic Digits team, Rebecca Evans of Pell Ensemble and Su Adams of U Can Too. We also thank Emma McFarland of Studio EMC and Jasmine Wilson of Studio Wayne McGregor for their contributions to the study and Ethan Wood for his invaluable support in assembling the EvalMe prototypes. Finally, we thank all of the children who participated in the workshop deployments.

REFERENCES

[1] Alissa N. Antle and Alyssa F. Wise. 2013. Getting down to details: Using theories of cognition and learning to inform tangible user interface design. *Interact. Comput.* 25, 1 (2013), 1–20.

[2] Ester Baauw and Panos Markopoulous. 2004. A comparison of think-aloud and post-task interview for usability testing with children. In *Proceedings of the 2004 conference on Interaction design and children: building a community* (IDC '04), Association for Computing Machinery, New York, NY, USA, 115–116. DOI: https://doi.org/10.1145/1017833.1017848

[3] Gerald Bergtrom. 2006. Clicker Sets as Learning Objects. *Interdiscip. J. E-Learn. Learn. Objects* 2, 1 (January 2006), 105–110.

[4] David Boud, Rosemary Keogh, and David Walker. 2013. *Reflection: Turning experience into learning.* Routledge.

[5] Karen Brennan and Mitchel Resnick. 2012. New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the American educational research association, Vancouver, Canada*, 25.

[6] Harry Brignull and Yvonne Rogers. 2003. Enticing people to interact with large public displays in public spaces. In *Proceedings of INTERACT*, Brighton, UK, 17–24.

[7] Jane E. Caldwell. 2007. Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sci. Educ.* 6, 1 (2007), 9–20.

[8] M. Chi. 1997. Why is self explaining an effective domain general learning activity. *Adv. Instr. Psychol. Lawrence Erlbaum Assoc.* (1997).

[9] Abigail C. Durrant, John Vines, Jayne Wallace, and Joyce S. R. Yee. 2017. Research Through Design: Twenty-First Century Makers and Materialities. *Des. Issues* 33, 3 (July 2017), 3–10. DOI: https://doi.org/10.1162/DESI_a_00447

[10] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, 216–223.

[11] Sarah Gallacher, Connie Golsteijn, Yvonne Rogers, Licia Capra, and Sophie Eustace. 2016. SmallTalk: Using Tangible Interactions to Gather Feedback from Children. In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction* (TEI '16), Association for Computing Machinery, New York, NY, USA, 253–261. DOI: https://doi.org/10.1145/2839462.2839481

[12] Andrew Garbett, David Chatting, Gerard Wilkinson, Clement Lee, and Ahmed Kharrufa. 2018. ThinkActive: Designing for Pseudonymous Activity Tracking in the Classroom. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, New York, NY, USA, 1–13. DOI: https://doi.org/10.1145/3173574.3173581

[13] William Gaver. 2002. Designing for homo ludens. *I3 Mag.* 12, June (2002), 2–6.

[14] Connie Golsteijn, Sarah Gallacher, Lisa Koeman, Lorna Wall, Sami Andberg, Yvonne Rogers, and Licia Capra. 2015. VoxBox: A Tangible Machine that Gathers Opinions from the Public at Events. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction* (TEI '15), Association for Computing Machinery, New York, NY, USA, 201–208. DOI: https://doi.org/10.1145/2677199.2680588

[15] Charlene Jennett, Ioanna Iacovides, Anna L. Cox, Anastasia Vikhanova, Emily Weigold, Layla Mostaghimi, Geraint Jones, James Jenkins, Sarah Gallacher, and Yvonne Rogers. 2016. Squeezy Green Balls: Promoting Environmental Awareness through Playful Interactions. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (CHI PLAY '16), Association for Computing Machinery, New York, NY, USA, 389–400. DOI: https://doi.org/10.1145/2967934.2968102

[16] Robin H. Kay and Ann LeSage. 2009. Examining the benefits and challenges of using audience response systems: A review of the literature. *Comput. Educ.* 53, 3 (November 2009), 819–827. DOI: https://doi.org/10.1016/j.compedu.2009.05.001

[17] Hyeonjin Kim and Michael J. Hannafin. 2008. Grounded design of web-enhanced case-based activity. *Educ. Technol. Res. Dev.* 56, 2 (2008), 161.

[18] Külli Kori, Margus Pedaste, Äli Leijen, and Mario Mäeots. 2014. Supporting reflection in technology-enhanced learning. *Educ. Res. Rev.* 11, (2014), 45–55.

[19] Victor R. Lee and Joel Drake. 2013. Quantified recess: design of an activity for elementary students involving analyses of their own movement data. In *Proceedings of the 12th International Conference on Interaction Design and Children* (IDC '13), Association for Computing Machinery, New York, NY, USA, 273–276. DOI: https://doi.org/10.1145/2485760.2485822

[20] Äli Leijen, Ineke Lam, Liesbeth Wildschut, P. Robert-Jan Simons, and Wilfried Admiraal. 2009. Streaming video to enhance students' reflection in dance education. *Comput. Educ.* 52, 1 (2009), 169–176.

[21] Linda J. Collins. 2007. Livening up the classroom: Using audience response systems to promote active learning. *Med. Ref. Serv. Q.* 26, 1 (January 2007), 81–88. DOI: https://doi.org/10.1300/J115v26n01_08

[22] Panos Markopoulos, Janet C. Read, Stuart MacFarlane, and Johanna Hoysniemi. 2008. *Evaluating children's interactive products: principles and practices for interaction designers.* Elsevier.

[23] Paul Marshall. 2007. Do tangible interfaces enhance learning? In *Proceedings of the 1st international conference on Tangible and embedded interaction*, 163–170.

[24] S. McLoone, S. O'Keeffe, R. Villing, and C. Brennan. 2014. Evaluation of a Smartphone-based Student Response System for Providing High Quality Real-time Responses in a Distributed Classroom. (January 2014), 210–215. DOI: https://doi.org/10.1049/cp.2014.0687

[25] D. Metcalf, M. Milrad, D. Cheek, S. Raasch, and A. Hamilton. 2008. My Sports Pulse: Increasing Student Interest in STEM Disciplines through Sports Themes, Games and Mobile Technologies. In *Fifth IEEE International Conference on Wireless, Mobile, and Ubiquitous Technology in Education (wmute 2008)*, 23–30. DOI: https://doi.org/10.1109/WMUTE.2008.38

[26] Pantelis M. Papadopoulos, Stavros N. Demetriadis, Ioannis G. Stamelos, and Ioannis A. Tsoukalas. 2011. The value of writing-to-learn when using question prompts to support web-based learning in ill-structured domains. *Educ. Technol. Res. Dev.* 59, 1 (2011), 71–90.

[27] Sara Price, Taciana Pontual Falcão, Jennifer G. Sheridan, and George Roussos. 2009. The effect of representation location on interaction in a tangible learning environment. In *Proceedings of the 3rd International Conference on Tangible and Embedded Interaction*, 85–92.

[28] Yvonne Rogers and Henk Muller. 2006. A framework for designing sensor-based interactions to promote exploration and reflection in play. *Int. J. Hum.-Comput. Stud.* 64, 1 (2006), 1–14.

[29] Yvonne Rogers and Sara Price. 2006. Using ubiquitous computing to extend and enhance learning experiences. (2006).

[30] Yvonne Rogers, Sara Price, Cliff Randell, Danae Stanton Fraser, Mark Weal, and Geraldine Fitzpatrick. 2005. Ubi-learning integrates indoor and outdoor experiences. *Commun. ACM* 48, 1 (January 2005), 55–59. DOI: https://doi.org/10.1145/1039539.1039570

[31] Lev Vygotsky. 1978. Interaction Between Learning and Development. *Read. Dev. Child.* (1978), 34–40.

[32] T. Wu, T. Yang, G. Hwang, and H. Chu. 2008. Conducting Situated Learning in a Context-Aware Ubiquitous Learning Environment. In *Fifth IEEE International Conference on Wireless, Mobile, and Ubiquitous Technology in Education (wmute 2008)*, 82–86. DOI: https://doi.org/10.1109/WMUTE.2008.9

[33] John Zimmerman and Jodi Forlizzi. 2014. Research Through Design in HCI. In *Ways of Knowing in HCI*, Judith S. Olson and Wendy A. Kellogg (eds.). Springer, New York, NY, 167–189. DOI: https://doi.org/10.1007/978-1-4939-0378-8_8