Wearables for Learning: Examining the Smartwatch as a Tool for Situated Science Reflection

Brittany Garcia  
The StoryLab  
Texas A&M University  
brinni@tamu.edu

Sharon Lynn Chu  
The StoryLab  
Texas A&M University  
sharilyn@tamu.edu

Beth Nam  
The StoryLab  
Texas A&M University  
namieux@tamu.edu

Colin Banigan  
The StoryLab  
Texas A&M University  
cbanigan@tamu.edu

ABSTRACT
Relatively little research exists on the use of smartwatches to support learning. This paper presents an approach for commodity smartwatches as a tool for situated reflection in elementary school science. The approach was embodied in a smartwatch app called ScienceStories that allows students to voice record reflections about science concepts anytime, anywhere. We conducted a study with 18 fifth-grade children to investigate first, the effects of ScienceStories on students’ science self-efficacy, and second the effects of different motivational structures (gamification, narrative-based, hybrid) designed into the smartwatch app on students’ quality and quantity of use. Quantitative results showed ScienceStories increased science self-efficacy especially with a motivational structure. The gamified version had the highest quantity of use, while narrative performance performed worst. Qualitative findings described how students’ recordings related to science topics and were contextualized. We discuss how our findings contribute to understanding of how to design smartwatch apps for educational purposes.

Author Keywords
Wearables; Smartwatch; Education; Children; Science learning; Situated cognition; Everyday cognition.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
A key problem in the learning of science, especially at elementary levels, is that children do not understand how science is relevant to them. This lack of perceived usefulness often comes from a lack of personal or cultural relevance of the content to be learned; the science is decontextualized with respect to actual contexts of use [1]. Yet, the child’s experiences in the world outside the classroom are rich and full of experiences that implicitly or explicitly relate to science [2, 3]. These represent tremendous missed opportunities to help children understand more about their physical and material environment.

This paper proposes the use of commodity smartwatches as a tool to motivate children to reflect on how learned science concepts relate to everyday scenarios. Various forms of mobile and wearable computing technologies have been studied as means to support science learning in formal and informal contexts. Systems such as Kitchen Chemistry [4], Zydeco [5], and StoryKit [6] have been proposed at proof-of-concept levels, but PDAs, smartphones and tablets are the main successfully commercialized mobile devices that have so far been investigated in the educational milieu. Relatively little formal discussion has yet occurred on the potential of smartwatches, as the new player on the market, to support children’s learning. Smartwatches have for the most part been explored in the healthcare and fitness domains (e.g., [7, 8]). The proceedings of CHI in 2016 included 9 papers on smartwatches, among which 6 addressed usability aspects of the smartwatch, 2 related to multi-device interaction, and 1 related to fitness. None explored smartwatches for learning.

Characteristics of the smartwatch in terms of portability, glanceability and ability to receive and capture information straight on the user’s wrist [9] positions it as a unique platform for learning and attitude change that takes place at micro-moments on-the-go and in-the-wild. A key challenge with the use of smartwatches to such effects, however, is that the usage of smartwatches once they are obtained tends to be low. Prior research on smartwatch use has consistently noted that the “sustained engagement with these devices” begins to drop after initial use. The drop can be attributed to a number of factors: Dibia [10] lists user boredom and decreased novelty, and resource constraints of the smartwatch. Shih et al. [8] state that the low intrusiveness of the smartwatch may lead to problems with users being mindful about the device and any tasks that it presents. For smartwatches to be effective as a platform to support in-situ science reflections in the education domain, the issue of continuous use has to be addressed.

Our work addresses two main questions: first, we investigate whether smartwatch apps can function as tools for in-situ reflection to positively affect elementary school-aged students’ self-efficacy towards science. And second, we explore different motivational frameworks (specifically,
gamification and the use of storytelling) by which smartwatch apps may be structured to catalyze engagement in smartwatch use within an educational context.

Following, we begin by summarizing the literature on smartwatch use in education and its associated problem of low use. We then describe our approach of smartwatches as situated reflection tools and the theoretical framework in which the approach is grounded. The rest of the paper covers the mixed-design study that we conducted with 18 students in a fifth-grade class using a custom-made smartwatch app.

BACKGROUND AND RELATED WORK

Smartwatch Use in Education

Smartwatches are one of the most common wearable devices, defined by Barfield and Caudell [11] as “fully functional, self-powered, self-contained computer[s] that [are] worn on the body... (and) provide access to information and interaction with information, anywhere and at anytime”. The main focus of research with smartwatches has been on how they can help people to track, manage and improve health and fitness [8]. Literature on the use of smartwatches for learning is quite sparse. We found that existing literature on the topic can be grouped into three types: theoretical work, design proposals, and empirical work.

Theoretical work discusses the affordances, trends and challenges of using the smartwatch in educational settings. Bower and Sturman [12] conducted a study with educators self-rated as well versed on using smartwatches, and discovered 14 affordances and 13 issues that can be grouped into three themes with regards to wearables in education: pedagogical uses (in-situ contextual information, recording, etc.); educational quality (engagement, greater efficiency, presence, etc.) and logistics (hands-free access, cost, etc.). Similarly, based on a survey of Chemistry undergraduate students, Adediwura et al. [13] found that while more than half of the respondents were not aware of wearable technologies, they were able to point to potential benefits and challenges of this class of devices for learning. Others [12, 14-16] have also pointed out challenges with smartwatch use in education, including security and privacy issues with persistent data collection; classroom limitations (e.g., disruptions, cheating); affordability; multitasking; and outside dependency (e.g., constant access to the internet).

A set of papers put forth design proposals for how the smartwatch may be used to support learning. For example, smartwatches have been proposed as devices to monitor class and student performance to maximize knowledge-building processes, and to understand students interests and state-of-being at school by de Arriba-Perez et al. and Llorente and Morant [17, 18]. Similarly, Park et al. [19] implemented the iBadge wearable system that records the interaction between student and teacher, and later compiled this data for the teacher to assess children’s learning progression. Outside of the classroom, smartwatches have been proposed as a platform to allow students to interact with the teacher and peers for help on class assignments [20, 21].

Finally, there is little work that is empirical relating to smartwatches in education. Scholl, Willie and Laerhoven [22] implemented a smartwatch app coupled with Google Glass that tracked the movement of a student conducting a science experiment in the wet lab. They found that the smartwatch was able to successfully guide the student through the experiment without disrupting their workflow to take handwritten notes. Esakia et al. [23] have incorporated smartwatches into computer science education, and found that programming for smartwatches positively helped students to understand abstract concepts and device constraints. Empirical work on smartwatches are much more common in other areas: smartwatches for activity tracking (e.g., [8, 24-26]); general user perceptions on smartwatches (e.g., [27-30]); designing for smartwatches (e.g., [31-37]), interaction with smartwatches (e.g., [38-42]); and how people use smartwatches in everyday life (e.g., [43, 44]).

One particularly relevant work is that of Pauw et al. [45], although it makes use of mobile technologies instead of wearables. They call their approach ‘life-relevant science learning’ and designed a mobile app ScienceKit that allows children to “capture moments of interest in their daily lives with multimedia and connect them to science inquiry by making claims, posing questions, and designing experiments”. However, their study was conducted with only 7 learners in a summer camp called ‘Kitchen Chemistry’ focusing on cooking, whereby the children could record moments of interest only during the time that they were in the sessions. Moreover, the children’s activity during the camp was moderated by facilitators.

Although smartwatches are projected to increase in penetration over the next few years [46], sustained adoption of the watches have been questionable. Harrison et al [47] and Ledger [48] reported that users begin to lose interest in their smartwatch devices after 6 to 12 months. Some research has been done on how to improve the sustained use of smartwatches. Kumar et al. [30] and Jung et al. [33] suggested that tackling the issue requires us to begin understanding user perceptions about smartwatches and user ideas of how they work. Our research addresses the use of smartwatches in elementary science education, focusing on its affordance to support situated reflection, and design structures that can motivate its use for learning.

THEORETICAL FRAMEWORK

Situated Reflection and Self-Efficacy

We approach the use of the smartwatch as a tool for situated reflection. The approach rests on the understanding that learning is socially and culturally determined. Learning is not just knowing a concept and being able to recite its definition. Knowing a concept means that the knowledge can be wielded in everyday cognition. This is strongly echoed in the theory of situated cognition that posits the importance of the “integration of subject-matter knowledge and everyday
knowledge…for children’s conceptual development” [49]. Palincsar [50] isolates three central aspects of situated cognition: i) the grounding of education in a practical world of experience; ii) relationship of a concept to a broader body of knowledge; and iii) inculcation of a sense of agency in children that learning is not just for school but also for the purpose of understanding life events.

Thus in practice, situated learning requires the identification and use of example scenarios that is practice-based, generative in concepts, and empowering in that children feel that they have a role to play in them. However, opportunities for such scenarios are limited in the classroom. Young [51] lamented that solving “the task of selecting situations that combine children’s everyday knowledge with subject-matter knowledge is a very serious problem”. Herein lies the benefit of the smartwatch. The smartwatch opens up a vista of scenarios from the child’s own life for her to deconstruct, classify, exemplify, relate to, and extend, with respect to formal science concepts within the constraints of her mental capacity. This is illustrated in Figure 3. The persistent presence of the smartwatch functions as both enabler, allowing children to capture science reflections whenever these occur to them in everyday life (practice-based, empowering), and as reminder, prompting the child to reflect on daily experience (generative).

Reflecting on a concept entails at least three stages: i) **Breakdown.** The individual realizes something that is incoherent with her current understanding of a concept; ii) **Inquiry.** The surprising realization leads to a reappraisal of previous knowledge; and iii) **Transformation.** The individual re-conceptualizes her understanding of the concept [52]. In contrast to instructivist learning, which focuses on the acquisition of knowledge, in the reflective process feelings and cognition are closely interrelated [53]. The recalibration of the individual’s mental model may not be accurate, and the newly gained understanding may still be incorrect, but the individual, having gone through the transformation process, would still feel that they know more than before – thus a possible increase in self-efficacy.

Self-efficacy has been defined as “people’s beliefs about their capabilities to exercise control over their own level of functioning and over events that affect their lives” [54]. The construct has often been equated with self-confidence or self-esteem. As a carrier of meaning and valence that shapes one’s perception of external events, self-efficacy influences people’s feelings, cognition, and course of actions [55]. Perceived self-efficacy can trigger self-regulatory processes such that “a person with the same knowledge and skills may perform poorly, adequately, or extraordinarily depending on fluctuations in self-efficacy thinking” [54]. The ability of self-efficacy to influence one’s motivation to engage in specific behaviors and tasks is generally accepted [56].

**Motivating Engagement in Reflection**

In considering the smartwatch as a situated reflection tool for science learning, two points have to be addressed: a) such informal educational technologies tend to be self-initiated and learner-led [57], relying actively on the child’s motivation to engage; and b) wearable technologies tend to recede into the background over time leading to a drop in use. The overlaying of motivational structures in educational smartwatch apps may help to address these issues. Various forms of motivational structures have been used in the design of apps and systems. Invoking gameful experiences through gamified interfaces, for example, have been used with some success to catalyze behavioral engagement in educational systems [58]. Gamification is the use of game design elements, such as points, leaderboards, and badges, in non-game contexts [59]. The embedded elements provide the user with a sense of challenge and reward her with rewards if the challenge is met. This may potentially lead her to engage in the underlying task more frequently. An example would be the well-known MOOC (Massive Open Online Course) platform Khan Academy, which rewards learners with badges and trophies for their progress in watching content videos and taking quizzes [60]. In our approach, gamifying the smartwatch reflection task would entail posing science reflection as the challenge, which if fulfilled brings the user some kind of reward (Figure 1).

Another motivational structure that has often been used in the design of educational interfaces is that of storytelling. Narrative-centered learning environments contextualize the learning task within a story that the user co-constructs [61]. They make use of the inherent structure of narrative and the meaning-making potential of storytelling to motivate the user to engage in the learning tasks more frequently and with more care. The main difference between narrative- and game-based structures is that narrative motivates by prompting users to connect to an interesting fictional scenario, while the game motivates by promising the

---

**Figure 1.** Smartwatch as a situated reflection tool with game-based motivational structure

**Figure 2.** Smartwatch as a situated reflection tool with narrative motivational structure
acquisition of something that the user would like to have. Examples of narrative-centered educational technologies include *Crystal Island*, a virtual environment developed for an eighth-grade microbiology course to immerse students in a world where they have to solve cases on what pathogens have poisoned the characters. An evaluation of the environment showed that students had some learning gains, but gains in motivational aspects such as self-efficacy and interest in science were more significant [62]. In a narrative-centered version of our smartwatch approach, the reflection task would be modified to encompass a story scenario as a third factor in the dynamic interplay of reflection (Figure 2).

**RESEARCH QUESTIONS**

Our work asked the following questions:

**RQ 1:** Does the use of the smartwatch as a tool for situated reflection affect children’s self-efficacy in science?

**RQ 2:** What kind of motivational structure improves the use of the smartwatch as a tool for situated science reflection?

**RQ 3:** How is the smartwatch used as a tool for situated science reflection?

We hypothesized that children who reflect on how science relates to different scenarios in their everyday life through the smartwatch will have a greater sense of self-efficacy in science. But sufficient reflection of adequate quality must be done for the impact to be expected. Figure 4 shows our proposed theoretical framework. We posited that the use of *ScienceStories* will increase students’ science self-efficacy as opposed to only regular class instruction (Hypothesis H-A). The more effective use of *ScienceStories*, the greater the sense of science self-efficacy (Hypothesis H-B), where use effectiveness is determined by both the quantity and quality of science recordings done. We also hypothesized that while *ScienceStories* with a gamified structure will increase the quantity of recordings (Hypothesis H-C), *ScienceStories* with a narrative structure will increase both the quantity and quality of recordings (Hypothesis H-D).

**THE SCIENCESTORIES SMARTWATCH APP**

**Design Process**

To investigate our research questions, a smartwatch app called *ScienceStories* was developed for children aged 8 to 11. Built using the Android Wear platform, the app utilizes the built-in microphone and speakers in a smartwatch to enable voice recording and playback of reflections recorded by the child. The design process was as follows: i) A basic app design to allow reflections capture was created following basic usability principles for smartwatches (e.g., prominent text, minimal number of screens, swiping interactions). This base app design followed the flow shown in Figure 5 (excluding the screens with an *); ii) Science concepts to be addressed were obtained from the participating school teacher, and prompts were developed for each concept according to the following logic: (Stage 1) What is the [science concept about?]?, (Stage 2) What happens with the

---

**Figure 4. Theoretical framework**

**Figure 5. Base smartwatch app (* denotes varying screens in different app versions)**

[science concept] in the real world?, (Stage 3) What would happen if the [science concept] did not exist?. For example, the prompts for the science concept ‘food chain’ were as follows: (Stage 1) “Who are the producers, consumers, and decomposers around you?”, (Stage 2) “How does energy flow around you?”, (Stage 3) “What would happen if there was no more grass to eat?”. A stage was unlocked only when the previous one had at least one saved recording. Only three stages were available. Figure 6 shows examples of screens from the base app: (left) a list of keywords related to the focus science topic – ‘Carbon dioxide-Oxygen cycle’ in the screenshot; (middle) a prompt related to the science topic; and (right) the recording screen. Recordings could be saved or discarded after recording; iii) Three different versions of the base app were developed (gamified, narrative-based, and hybrid combining game and simple narrative elements). These versions had other screens beyond the base app to embed different kinds of motivational structures; and iv) a usability test was conducted with the three app versions.

**Motivational Structures Design**

To ensure that the various versions were comparable, only the stage selection screen, the setup screen and the feedback screen were altered for each version (marked with an * in Figure 5). Figure 7 shows the variations for the stage selection screen. Stages in the narrative version were called ‘episodes’ (left), and ‘levels’ in the game and hybrid versions. Figure 8 shows the variations of the setup screen. The ‘setup’ screen provided some context for the prompt. The narrative version setup presented a first-person immersive fictional story that led into the science prompt for that stage. The story was presented in both text and audio voice-over to avoid the possibility of a child’s ability or discomfort with reading long texts to be a factor in
interaction. The three stages formed a narrative arc such that the first stage set the story context, the second stage related a climatic action, and the third stage led towards story resolution. The game version setup screen displayed simple game instructions. The hybrid version setup was similar to the narrative version for the setup screen, but the story was presented in a short third-person expository style. The feedback screen variations are shown in Figure 9. Feedback in the narrative app was conveyed by a character in the story. The game and hybrid versions gave a typical ‘good job’ feedback.

Usability Testing

Eighteen child participants (11 boys and 7 girls with mean age = 9) participated in the test. Participants were recruited through an announcement to the university student, staff and faculty listservs. Children were scheduled on a first-come, first-served basis. The child watched a 5-minute-long video about a science concept, and was shown how to use the app. Then each child was given a smartwatch with a version of ScienceStories and asked to complete recordings over two days. On Day 3, participants completed the IBM PSSUQ scale (Post-Study System Usability Questionnaire) that we adapted for smartwatch apps. Overall, participants rated the usability of the narrative version of ScienceStories an average of 4.40, the gamified version 4.43, and the hybrid version 4.38. No significant differences were found in usability across the versions. Minor issues that emerged from the usability test with the app (e.g., confusing titles, holding the recording button while recording) were fixed.

STUDY DESCRIPTION

After usability testing of the ScienceStories smartwatch apps, we worked with a local elementary school to conduct a study that addressed our research questions. The study involved 18 participants in total (7 boys and 11 girls with mean age = 10.5) from a fifth-grade science class. Their demographics were as follows: White (n = 6); Hispanic (n = 11); and Mixed Races (n= 1). Parental consent forms and minor assent forms were distributed to the parents and children prior to the study. The commercially available ASUS Zenwatch 2 was used for the study. The Zenwatch has a battery life of approximately 48 hours with continuous use.

The study followed a nested design (see Figure 10), whereby the top-level factors are ‘intervention’ and ‘baseline’. Intervention consisted of smartwatch use on top of classroom instruction, and baseline was only regular classroom instruction. The sub-level factors of intervention were ‘with-structure’, the use of the smartwatch app that has a motivational structure (narrative, gamified, hybrid app versions) as opposed to ‘no-structure’, the smartwatch app with no motivational structure (base version of app only). The ‘with-structure’ factor had two further sub-factors consisting of the smartwatch app with the three varying motivational structures of ‘narrative’, ‘game’ and ‘hybrid’.

The study was conducted over a 5-week period (see Figure 10). At school, the students cover one science topic or concept each week. In the first 3 weeks, the participants used ScienceStories with the 3 motivational structures as follows: a participant was assigned a different app version at the beginning of each week such that she has used all of the 3 versions at the end of the 3 weeks. The order of use of the three versions was counter-balanced among all the participants. In week 4 of the study, all participants used the base version of ScienceStories that had no motivational structure. And in week 5, all participants filled in pre- and post-questionnaires, and only underwent regular class instruction without using the smartwatch. Although we are aware that weeks 4 and 5 were tied to specific science topics, the study was designed as such so as to be able to fit the constraints of the schedule of an authentic functioning public school classroom, and we were not able to change it.

Briefly, the protocol for the intervention conditions each week of the study was as follows:

Day 1: i) The students filled in a set of pre-questionnaires; ii) A short presentation was given demonstrating the functionalities of the watch (e.g. turning the watch on and off, swiping functions); iii) The participants were grouped based on the app version that they were assigned to that week, and each child was given a smartwatch with the respective app version installed to bring home; vi) Each
group was briefed on their assigned app version; vii) All the students were given the study task to record stories from their daily life related to the science concept covered in class that week over the course of Day 2 and Day 3. The children were also told that there were no right or wrong answers for the recordings, and that they could use the app as little or as much as they wanted.

Day 2: The students used the ScienceStories smartwatch app to record reflections however many times they wanted. No interaction between the researchers and the students took place.

Day 3: After two and half days of watch use, the smartwatches were retrieved from the students, and the students filled in post-questionnaires.

At the end of week 4, each student filled in another questionnaire comparing the app versions, and participated in a half-hour individual interview. We note that although the 5 weeks of the study addressed different topics or concepts according to the school science curriculum, the topics were similar in nature (Week 1: Food chains; Week 2: Adaptation; Week 3: Inherited and learnt traits; Week 4: Environmental changes; Week 5: Carbon dioxide-Oxygen cycle).

DATA COLLECTION AND MEASURES
The pre- and post-questionnaires assessed, on a likert scale of 1 (Strongly disagree) to 5 (Strongly agree), self-efficacy in science using the Morgan-Jinks Student Self-Efficacy Scale (MJSES) [63] applied to the science topic addressed each study week. The MJSES scale was chosen because the scale focused on younger students in school settings. The MJSES consists of 3 subconstructs: talent, which consists of items such as ‘I am a good science student in [science topic]”; ‘I am smart in [science topic]”; effort, with items such as ‘I work hard in [science topic] at school’, ‘I always get good grades in [science topic] when I try hard”; and context, with items like ‘Most of my classmates like to do [science topic] in science because it is easy’, ‘No one cares if I do well in [science topic] in school’. For the purpose of analysis, a difference score was computed for each component of the scale by taking the pre-test score and subtracting it from the post-test score.

The post-questionnaires also measured usability of the smartwatch apps using the adapted IBM PSSUQ scale (Post-Study System Usability Questionnaire). All recordings were retrieved from the students’ smartwatches and were fully transcribed. The following were obtained from the smartwatches: the total number of recordings captured by each child, and total number of recordings for each app version. All quantitative data were input into the SPSS statistical analysis software package. One-way between-subjects ANOVAs were run to see whether there were any differences between the participants’ reported usability scores again for the various app versions. No significant difference was found between the various app versions, showing comparable usability.

Figure 10. Study design

Figure 11. Results graphs
FINDINGS AND DISCUSSION
We report our findings below grouped according to the research questions:

RQ1 Science Self-Efficacy: A repeated measures ANOVA was conducted to compare the effects of the different app versions (gamified, narrative-based, hybrid) on participants’ MJSES science self-efficacy difference scores. No significant effects were found for these overall MJSES scores, but a repeated measures ANOVA on the separate subconstructs of the MJSES showed a significant effect of app version for the subconstruct ‘effort’, $F(2, 34) = 4.191, p = .024$ (see Figure 11A). Pairwise comparisons showed that the significant differences were between the narrative and hybrid versions, and the gamified and hybrid versions. Mean effort scores for both the narrative ($M = -.278, SD = .545$) and gamified ($M = -.167, SD = .345$) versions showed a negative change, while the hybrid version showed a positive change ($M = .144, SD = .387$).

Self-efficacy difference scores for the narrative, gamified and hybrid app versions were grouped and averaged out for each participant as difference scores for 'with-structure' condition. A repeated measures ANOVA was conducted to compare the effects of presence of a motivational structure (app with structure, base app with no structure) on participants’ MJSES science self-efficacy difference scores. There was a significant effect of motivational structure, $F(1, 15) = 5.958, p = .028$ (see Figure 11B). The ‘with-structure’ factor had a positive increase ($M = .034, SD = .111$), while mean difference scores for the app with no structure had a negative mean ($M = -.121, SD = .199$).

Finally, a repeated measures ANOVA was conducted to compare the effects of smartwatch intervention (smartwatch use; baseline regular classroom instruction) on participants’ MJSES science self-efficacy difference scores. The ‘with-structure’ scores were not aggregated with the ‘no-structure’ scores although both consisted of a smartwatch intervention because we deemed the discrepancy between the two conditions to be too large. There was a significant effect on self-efficacy change level, $F(2, 28) = 3.977, p = .030$ (see Figure 11C). For the intervention condition with motivational structures, science self-efficacy showed a positive change ($M = .048, SD = .100$). Science self-efficacy showed negative changes for both the intervention condition with no structure ($M = -.118, SD = .206$) and the baseline condition ($M = -.081, SD = .151$).

We hypothesized that the use of ScienceStories will increase students’ science self-efficacy as opposed to only regular class instruction (hypothesis H-A). This hypothesis was supported. Students felt as if they knew more about science when using ScienceStories while their degree of confidence decreased with regular classroom instruction. Part of the findings from the post-interviews (see section later) provides a rationale for this. The children felt empowered and became mindful of their science knowledge. We propose that making connections to the science topics outside of the classroom led them to feel that they are able to use knowledge learned and that they can wield the knowledge for personal situations. It is also possible, although unlikely, that because the students saw the smartwatch activity as simply more fun than classroom work, heightened general positivity led to an increase in science self-efficacy in that condition.

RQ2 Motivational Structures and Usage Profile: The average number of recordings made by a child (called his/her ‘quantity score’) was calculated by week and by app version. By week, the averages were more or less similar as follows: Week 1 ($M = 2.78$), Week 2 ($M = 2.17$), Week 3 ($M = 3.22$) and Week 4 ($M = 2.56$). Figure 11D shows the averages by app version. The gamified version had the highest means ($M = 3.00$), followed by the narrative version ($M = 2.67$) and the no-structure version ($M = 2.56$), and the hybrid version had the lowest means ($M = 2.50$).

All recordings were coded to identify concepts that related to the focus science topic in the following manner. For each science topic, we identified the list of concepts that needed to be mastered by students from the state-mandated fifth-grade science curriculum provided to us by the teacher. This list was used as benchmark to identify concepts that were relevant to the science topic in each recording. A ‘quantity score’ was generated for each participant by summing the total number of science topic-relevant concepts across all of the participants’ recordings from each app version. Figure 11D shows the average quantity score for each app version per child. The no-structure version had the highest means ($M = 4.06$), followed by the hybrid version ($M = 3.83$) and the gamified version ($M = 3.56$), and the narrative version had the lowest means ($M = 2.72$).

Further, each participant was given a summary ‘use effectiveness score’ through the summation of the ‘quantity’ and ‘quantity’ scores. A Pearson product-moment correlation coefficient was computed to assess the relationship between the use effectiveness scores and science MJSES science self-efficacy scores. There was a significant correlation between the two variables only for the narrative version of ScienceStories ($r = .507, p = .032$) (Figure 11E).

We hypothesized that the more effective the use of ScienceStories, the greater the sense of science self-efficacy (hypothesis H-B). The hypothesis held true only for the narrative app version. For the other app versions, the level of science self-efficacy was independent of degree of app use. Yet, the narrative version was used the least effectively, having the lowest scores for both quantity and quality of recordings. This means that for the narrative app, students had to use the app increasingly more effectively so as to increase in self-efficacy. On the contrary, using the other app versions even poorly – i.e., the mere fact of using the app – led to an impact on science self-efficacy. This could possibly be because narrative structures require a certain degree of engagement for users to feel that they have a stake in the experience, while gamified structures function in a more direct manner to affect state of mind.
Our hypothesis H-C was that the gamified version of the watches would have the highest quantity of stories. This hypothesis was supported with the gamified version having the highest mean number of stories per child, although the difference with the other app versions was not large.

Hypothesis H-D stated that the narrative app version would have the highest scores for quantity and quality of recordings. Our results, however, showed that conversely it performed the worst. This is corroborated by the significantly low effort scores for the narrative app relative to the gamified and hybrid versions. A possible explanation is that the narrative steered focus away from the science concept. Thus, neither did the students attempted to relate their reflections intimately with the science concept, nor did they perceive that they had more to say. This was unexpected because prior literature on storytelling showed that storytelling causes students to be more interested in educational topics [64] by experiencing empathy and empowerment through co-construction of the story [65, 66].

**RQ3 Smartwatch as a Situated Reflection Tool**

**Smartwatch Recordings – Relation to Science:** We were interested in whether and how the students applied the science concepts in their smartwatch recordings. A qualitative coding process was done by two coders on all the recordings using a coding scheme based on Graesser and Person’s [67] framework of question types that students ask in classroom situations. Graesser and Person’s framework groups questions into two main categories: short-answer and long-answer. Short-answer questions have 5 subtypes (verification, disjunctive, concept completion, features specification, and quantification), and long-answer questions have 10 subtypes (e.g., definition, example, comparison, interpretation, causal consequence, …). The framework was used as a coding scheme by considering recordings as responses expected from the various question types.

The students made 191 recordings stories in total over the 3 study days. One coder first coded the recordings to be either short-answer or long-answer. Short-answers took mostly the form of direct responses to a prompt given in the watch app, akin to the child responding to a test question in class (e.g., Recording A in Table 1). Short-answers accounted for 49.74% of all the recordings. Long-answers were mostly recordings that added other information to short-answers (e.g., Recording B). Long-answers made up 50.26% of the recordings. The recordings were further coded by each of the two coders independently with the most relevant subtype. Intercoder reliability between the two coders was satisfactory at 87.23%. All the recordings could be coded with at least one of 6 subtypes. Note that a recording could have more than one code, if relevant.

In the short-answer category, recordings coded with Concept Completion accounted for 56.84% of all recordings. In Graesser and Person’s framework, concept completions refers to “who? what?” questions. In the smartwatch recordings, concept completion recordings were direct, factual answers to the prompt from the app (e.g., Recording A in Table 1). Feature Specification are questions of the type “What qualitative attributes does entity X have?”. Feature specification recordings (36.84%) related specific attributes related to the science concept (e.g., Recording D). Irrelevant recordings were those that had no relation to the science topic. 6.32% of recordings were irrelevant in the short-answer category (e.g., Recording K).

In the long-answer category, recordings coded as Definition made up 7.63%. Definition questions are of type “What does X mean?”. The Definition smartwatch recordings provided a direct definition or description of the science concept (e.g., Recording B). Example questions ask “What is an example label or instance of the category?”. Smartwatch recordings coded as Example (68.64%) gave examples of the science concept (e.g., Recording C). Interpretation questions are of type “What concept or claim can be inferred from a static or active pattern of data?”. In recordings coded as Interpretation (20.34%), students inferred something about the science concept from observation and analysis (e.g., Recording E). And finally, Judgment questions relate the following: “What value does the answerer place on an idea or advice?”. Smartwatch recordings of type Judgment (9.9%) had the student making some kind of subjective judgment about the science concept (e.g., Recording F). Irrelevant recordings were 2.54% in the long-answer category.

<table>
<thead>
<tr>
<th>ID</th>
<th>Science Concept</th>
<th>[App version] Prompt</th>
<th>Example Science Recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Food Web</td>
<td>[Narrative] “Who are the producers, consumers, decomposers around me?”</td>
<td>“The only consumer I found was a cow.”</td>
</tr>
<tr>
<td>B</td>
<td>Adapta-tions</td>
<td>[Hybrid] “What does it mean to adapt to a habitat?”</td>
<td>“To adapt to an environment is to change the way you do things...A chameleon camouflages by changing color into whatever it’s near or on, that way it does not get eaten and it cannot be seen by the things it is trying to eat.”</td>
</tr>
<tr>
<td>C</td>
<td>Food</td>
<td>[Narrative] “Who are the producers, consumers, decomposers around me?”</td>
<td>“There is a flower in my backyard. A flower is a producer because a flower makes its own food from the sunlight.”</td>
</tr>
<tr>
<td>---</td>
<td>------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>D</td>
<td>Inherited/ Learned Traits</td>
<td>[Narrative] “What are inherited traits that you have as a dog?”</td>
<td>“I have inherited a long tail and a nose for sniffing things out.”</td>
</tr>
<tr>
<td>E</td>
<td>CO2 Cycle</td>
<td>[Base] “Record stories from you daily life that you think are related to the oxygen-carbon dioxide cycle.”</td>
<td>“Today I was playing outside and I was thinking of the oxygen and carbon dioxide cycle. I realized that when I breathe (makes breathing sound) carbon dioxide I give it to the plants and the plants give me oxygen back. It’s all a process that goes back and forth.”</td>
</tr>
<tr>
<td>F</td>
<td>CO2 Cycle</td>
<td>[Base] “Record stories from you daily life that you think are related to the oxygen-carbon dioxide cycle.”</td>
<td>“Another time was in the car and I stucked my head out the window and I can smell all of the pollution. I believe pollution can stop the carbon dioxide and oxygen cycle and can affect the cycle really badly.”</td>
</tr>
<tr>
<td>G</td>
<td>Food Webs</td>
<td>[Gamified] “What would happen if there was no more grass to eat?”</td>
<td>“What would happen if there were no more grass to eat? The animals would have no food so they would die and we would die ‘cause we need animals.”</td>
</tr>
<tr>
<td>H</td>
<td>CO2 Cycle</td>
<td>[Base] “Record stories from you daily life that you think are related to the oxygen-carbon dioxide cycle.”</td>
<td>“Now for a pop quiz, what type of gas release, is released from plants? If you guess carbon dioxide you’re right! Now for the next question is what type of gas is released from animals? If you guessed oxygen then you are correct!!”</td>
</tr>
<tr>
<td>I</td>
<td>CO2 Cycle</td>
<td>[Base] “Record stories from you daily life that you think are related to the oxygen-carbon dioxide cycle.”</td>
<td>“When I go to the park I usually see a bunch of tree and plants and flowers and the uhh produce oxygen for us and they breathe in carbon dioxide. We breathe in oxygen and breathe out carbon dioxide.”</td>
</tr>
<tr>
<td>J</td>
<td>CO2 Cycle</td>
<td>[Base] “Record stories from you daily life that you think are related to the oxygen-carbon dioxide cycle.”</td>
<td>“Today me and my neighbors were playing basketball. I looked up plants and thought ‘Woah! I’m breathing in-no plants breathe in carbon dioxide and I breathe out oxygen.”</td>
</tr>
<tr>
<td>K</td>
<td>Food Webs</td>
<td>[Gamified] “How does the energy flow around you?”</td>
<td>“How does the energy flow around? Light energy.”</td>
</tr>
</tbody>
</table>

We also coded the recording types by app version. There were 48 recordings in total from the narrative app; 53 from the gamified app; 58 from the hybrid app, and 54 from the base app. Figure 12 shows the breakdown of our coding. Recordings from the narrative app were mostly of type Example (39.58%), followed by Feature specification (29.17%). Gamified app recordings had almost equivalent number of recordings of type Example (30.19%) and Concept completion (30.19%). The hybrid app recordings were predominantly of type Example (44.83%). And the base version had a majority of type Example (35.19%) and Interpretation (31.48%) types of recordings. These results suggest that somewhat deeper science reflections occurred only with the base app at the interpretive level. The base app had a high percentage of recordings type interpretation this means that students were able to interpret science concepts through long-answers when using the base app.

Narrative and gamified apps had the highest recordings as feature specification. This could mean that connections to science concepts were expressed more easily through short-answers by giving specific attributes of the science concept. A low percentage of definitions was seen across all app versions. Students chose instead to record examples of the science concepts. The base app was the only version to have a student make a judgment about the science concept. Although there was overall a low percentage of irrelevant recordings, the narrative and gamified apps had the most number of unrelated recordings.

**Smartswatch Recordings – Recording Contextualization:**

We coded the recordings for whether they were contextualized or decontextualized. A contextualized recording was one that was couched into a certain context or scenario, whereas a decontextualized recording had no context and only answered the science prompt given in *ScienceStories*. Approximately the same number of recordings were decontextualized across the app versions – 36.7% for the narrative version, 46.2% for the gamified version, 44.4% for the hybrid version, and 39.1% for the version with no motivational structure.

The contextualized recordings were further coded as being contextualized either by scenarios from the child’s everyday life or by a fictional scenario. A fictional scenario was defined as details in the recording that set up the context of a story but did not relate to real-life. We found that the fictional scenarios were of two types: entirely imaginary coming from the child, or coming from the setup of the prompt given in the app. Figure 11F shows the percentage of recordings that had everyday scenarios and fictional scenarios for each app version. The ‘with-structure’ app versions all had higher percentages of recordings with fictional scenarios (narrative 87.1%, gamified 67.9%, hybrid 83.3%) than the base version (17.9%). The base app had much higher percentage of recordings with everyday scenarios (71.4%).

Fictional scenarios were of three types: i) roleplaying (48.61%) – the child took on a fictional persona and verbally enacted the science concept (e.g., Table 1D); ii) hypothetical contexts (9.72%) – ‘what if’ situations related to the science concept and what would happen in them (e.g., Table 1G); and iii) formal contexts (41.67%) – the child set up a scenario using a formal structure like a pop quiz or an interview to talk about the science concept (e.g., Table 1H). Everyday scenarios were also of three types: i) physical presence (33.33%) – the child related the science concept to a specific object while being at a specific location (e.g., Table 1C); ii) physical but impersonal (57.14%) – the child related the science concept to an object in a general place. The child may
have experienced being physically present at the stated location but he/she was not actually there during recording (e.g., Table 11); and iii) physical activity (9.52%) – the child related the science concept to what they were doing without specifying a location (e.g., Table 11). We note that the app did not allow for GPS tracking, and thus our analysis does not cover where and when the children did the recordings.

**Children’s Perceptions:** The students’ post-interviews were transcribed, and an open coding process was performed on the responses pertaining to their opinions on whether they learned more about science using the different versions of the app. Categories that emerged were: i) Providing additional information. Children perceived that the narrative version of the app provided them with extra information that helped them understand the science topic better. A response that was coded as such was: “Because whenever we uh sometimes at school they might say it’s something different and over here they can say something a little bit more about it and so that and so yeah it makes me learn something new”; ii) Metacognitive awareness. Some children felt that the watch app led them to realize what they knew and they did not know, and sometimes to question that state of knowledge. “I just think it just gives you a chance to for you to know what you’ve learned”; and iii) Empowerment. Students felt empowered when using the watch because they were able to say what they thought about the subject, thus allowing for personal interest to be imbued in the topic. For example, a student’s response was: “This one I think I taught it because this thing asked me the question so I guess I got to say what I think about it so I really liked this one”.

Our qualitative analysis of the recordings elicited the ways by which the students reflected on the science concepts, and by which the recordings were contextualized. We generalize the qualitative findings in the models shown in Figure 13. The gamified version had satisfactory reflections (spread of concept completion, feature specification and example recording types) of the science concepts (SC) contextualized in both fictional scenarios (FS) and everyday scenarios (ES). This is represented by the equally weighed arrows ‘SC→FS’ and SC→ES’. ‘Although that version did not embed a narrative, the students at times made up their own fictional contexts for the recordings. Science reflections using the narrative app, conversely, were heavily contextualized in fictional scenarios, with little emphasis on relating to everyday experiences and surface-level connection to the science concept. This is represented by the heavier weighed arrow ‘SC→FS’ and larger dot at FS. The hybrid version was similar to the narrative version in the sense that reflections were heavily couched in fictional scenarios, but the recordings related more to the science concept (heavier weighed arrow ‘SC→FS’). And finally, in reflections with the base version, everyday experiences were pertinently related to the science concepts, but had very few recordings.

**STUDY LIMITATIONS AND FUTURE WORK**

Our study has a number of limitations that we point out below. First, although all the prompts in each app version followed the same template, they related to different science topics. This was inevitable for two reasons: i) we worked in an authentic school context and aligned our study with the curriculum schedule at our partner elementary school; and ii) the learning effects from children re-engageing with the same science topic several times would be hard to overcome. Second, because of the fixed timeline that authentic school contexts present, we managed to counterbalance the use of only the three app versions with motivational structures. The use of the base app and the baseline conditions were done last by all students. And third, the children in our study used each app version over three days. This is longer than snapshot studies, but it is possible that over even longer use, the effects on science self-efficacy would be muted or enhanced. Future work involves the study of the smartwatch as a situated reflection tool with more complex game-based structures, and with longer period of use.

**CONCLUSION**

We proposed an approach that frames the smartwatch as a tool for situated reflection to support informal science learning in everyday life for elementary school students. We incorporated different kinds of motivational structures into the design of the ScienceStories educational smartwatch app, to explore ways to promote app use over time. Our study showed that the smartwatch as a situated reflection tool positively affects science self-efficacy and elicited the richness of students’ in-situ smartwatch science recordings. Much research has shown that self-efficacy is an important predictor of academic performance as well as amount of persistence in school work [68]. However, none of the app versions excelled in all dimensions. The gamified app appeared to be the best compromise in terms of generating sufficient recordings with some connection to the science concept, and resulting in increased science self-efficacy as compared to traditional classroom instruction.

More work has to be done to investigate smartwatch effects on actual learning or knowledge increase. As of now, we have barely begun to scrape the surface of the untapped potential of smartwatches for education. With its new set of characteristics focusing on persistent presence on the wearer, the smartwatch may allow teachers in the future to address and scaffold students’ learning in spatial and temporal contexts way beyond the classroom.

**ACKNOWLEDGMENTS**

We thank all children and the teacher who participated in our study. This work has been supported by National Science Foundation grant # CRII-1566469.

---

![Figure 14. Emphasis in the various smartwatch app versions](image-url)
REFERENCES
17. de Arriba-Pérez, F., M. Caeiro-Rodriguez, and J.M. Santos-Gago. Towards the use of commercial wrist wearables in education. in Experiment@ International Conference (exp. at’17), 2017 4th. 2017. IEEE.
32. Xu, C. and K. Lyons. Shimmering smartwatches: Exploring the smartwatch design space. in Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction. 2015. ACM.
59. Hamari, J., J. Koivisto, and H. Sarsa. Does gamification work?--a literature review of empirical studies on gamification. in System Sciences (HICSS),


