Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp

Yvonne Rogers[1,2], Kay Connelly[2], Lenore Tedesco[3], William Hazlewood[2], Andrew Kurtz[2], Robert E. Hall[3], Josh Hursey[2], and Tammy Toscos[2]

[1] The Open University, Computing Department
Milton Keynes, MK7 6AA, UK
y.rogers@open.ac.uk
[2] Indiana University, School of Informatics,
Bloomington, IN 47405, USA
{connelly,whazlewo,ajkurtz,jjhursey,ttoscos}@indiana.edu
[3] Indiana University~Purdue University, Indianapolis,
Center for Earth & Environmental Science Indiana 46202, USA
{ltedesco,bhall}@iupui.edu

Abstract. How should Ubicomp technologies be evaluated? While lab studies are good at sensing aspects of human behavior and revealing usability problems, they are poor at capturing context of use. In-situ studies are good at demonstrating how people appropriate technologies in their intended setting, but are expensive and difficult to conduct. Here, we show how they can be used more productively in the design process. A mobile learning device was developed to support teams of students carrying out scientific inquiry in the field. An initial in-situ study showed it was not used in the way envisioned. A contextualized analysis led to a comprehensive understanding of the user experience, usability and context of use, leading to a substantial redesign. A second in-situ study showed a big improvement in device usability and collaborative learning. We discuss the findings and conclude how in-situ studies can play an important role in the design and evaluation of Ubicomp applications and user experiences.

Keywords: In-situ studies, design, evaluation, user experience, usability, mobile learning.

1 Introduction

Evaluation is central to the design process when developing a new product, system or application. As ubiquitous computing technologies (aka Ubicomp) mature, it will become increasingly important that they, likewise, are evaluated to meet usability and user experience goals. However, Ubicomp applications are inherently difficult to evaluate due to their context of use. Traditional evaluation methods and metrics, designed for controlled laboratory settings, fail to capture the complexities and richness of the real world in which the applications are placed. For example, task completion times and usability errors say little about how an Ubicomp application

J. Krumm et al. (Eds.): UbiComp 2007, LNCS 4717, pp. 336–353, 2007.

engenders a novel user experience, such as collective story telling through distributed photography [27]. A new approach to capturing more of the context of use has been to create 'living' laboratories that attempt to simulate a particular environment, such as the home, that is instrumented to sense and measure all manner of human behaviors [e.g., 16,17].

An alternative paradigm has been to push the research out of the lab into the real world [see 29]. In-situ studies (also known as 'in the wild' studies) are beginning to appear that evaluate the situated design experience of Ubicomp, resulting in *understandings* of how novel pervasive technologies are appropriated in real world settings. These are quite different from the *results* of lab-based studies and include how: visitors engage with installations in museums [14]; people play mixed reality games in city streets and online [2, 4]; spectators record and communicate large-scale events [27]; biologists capture and analyze environmental field work observations [33] and students share and use a public display situated in their common room [7].

Kjeldskov et al., have argued, however, that in-situ studies provide *little added value,* being difficult and more expensive to conduct than lab studies and question whether "it is worth the hassle" [18]. While they can be labor-intensive and more costly to run than a lab study, it is increasingly accepted within the Ubicomp community that the rich and varied data that can be obtained *in situ* provide quite different insights into people's perceptions and their experiences of using, interacting or communicating through the new technologies in the context of their everyday and working lives. In addition, studies can be designed to obtain data about the usability of the technology, in terms of what functions are used, which are not and the difficulties encountered when used in a particular context.

The potential costliness and difficulty of running *in situ* studies, however, raises research questions as to how to make them effective. Utmost in many researchers' minds is how long should they last? Is a day, a week, a month or a year optimal? This obviously depends on the goals of a study, but the debate is most pertinent when evaluating mobile devices and applications that are explicitly designed to change people's habits that take time (e.g., exercising more [9, 31]) versus those that are designed to support and enhance an existing activity (e.g., brainstorming, scientific inquiry [25]). Another important issue is how much and what kinds of data to collect. Are *pervasive* methods, i.e., logging and sampling of events, enabled by the Ubicomp technologies, themselves, the most useful or are ethnographic methods, such as interviewing and videoing, more effective for capturing and analyzing changes in behavior? Or, is a hybrid approach feasible? A further debate is whether to represent *in situ* data as meaningful or significant: are bar charts, vignettes and quotes sufficient or are ANOVAs and regressions needed? Finally, having analyzed the data, how can the findings be fed back into the design process? In particular, how can they be used to improve both the design of the technology and the user experience?

Our research is concerned with explicating the methodological challenges and benefits of using *in situ* studies in the design process. We describe a case study that shows how an in-situ study informed the redesign of a mobile learning device, greatly improving both its situated use and usability. We describe the progression from initial user requirements to prototype design, to in-situ user study and analysis, to reflection and redesign, to a second in-situ evaluation that demonstrated substantial improvements. Section 2 provides the background to the evaluation methods being

used in Ubicomp. Section 3 outlines the initial project goals and the first design iteration of the mobile learning tool. The first in-situ study is then described in Section 4, followed by the findings and analysis in Section 5. Section 6 shows how the user experience and usability problems were categorized and how we used these to iterate further our design. We present the findings from the second in-situ study in Section 7 before concluding with a discussion of the value (and challenges) of in-situ studies during the design process.

2 Background

Usability testing is the conventional approach to evaluating user interfaces that involves collecting data using a combination of methods (i.e., experiments, observation, interviews, questionnaires) in a controlled setting, usually a lab. The primary goal is to determine whether an interface is usable by the intended user population to carry out the tasks for which it was designed [11]. The approach has been extensively and successfully used to evaluate software applications running on PCs and other technologies where participants can be seated in front of them to perform a set of tasks.

Ubicomp applications that are used over a long period of time by people who are moving around and doing other things, however, present a new set of challenges. One approach is to adapt existing HCI methods, such as heuristic evaluation for analyzing ambient displays [21]. Another is to develop new intervention evaluation methods for collecting and sampling data, including cultural probes [12], photo blogging [23] and the experience sampling method [8]. Ethnographies that describe the work people do in their day-to-day activities have also become more popular. The focus has been on explicating the situated nature of the work or other practice with an emphasis on how existing technologies are used by people in places like the home, hospital or church [e.g., 1, 10, 30] with a view to the 'play of possibilities' for designing new Ubicomp-based systems.

A few ethnographically-based, evaluations of prototypes have been situated in physical spaces [6, 7, 26] or by following mobile users around [31, 20]. Based on the findings arising from these studies, various conceptual frameworks have been developed that prescribe or sensitize other researchers to design concerns. [e.g., 3, 5]. While such frameworks can inspire the early phases of Ubicomp development, they offer little guidance on how to iterate a design in order to improve its usability, efficacy and/or enhance the user experience. Alternatively, new conceptual measures have been proposed such as focus, adoption and interpretation [28]. Case studies, such as ours, that explicate the issues, design rationale and choices made in a project, can also elucidate the processes involved [32].

3 The Lilly ARBOR Case Study

Our case study addresses a problem identified as part of an ongoing educational program: how to augment field experiences to better engage students in scientific inquiry processes. A team of environmental scientists had observed that students

performed limited analysis in the field, which was problematic since the program did not have a classroom component. The scientists asked if we could develop a mobile application that would provide the "right kind" of information to improve students' ability "to do more analysis" in the field. This premise was our starting point.

3.1 Overview of Lilly ARBOR Project

The Lilly ARBOR project is concerned with investigating ecological restoration of urban regions while also providing educational opportunities to a variety of students through hands-on learning activities. A one-mile stretch of riverbank in Indiana (US) was restored in 2000, using three of the most common methods for planting trees to restore native forests. The project site was divided into eight plots and over 1400 native trees were initially planted. The site is now evolving into a wildflower meadow and shrub/sapling habitat as the trees grow and other species gradually re-colonize the area.

Twice a year, teams of environmental scientists and students have conducted an assessment of the site, measuring the survival and growth of trees and noting things such as predator damage and the impact of the invasion of other trees and plants. Each team spends the day locating, identifying and measuring the surviving trees for a plot. The learning experience focuses on what is involved in being an environmental scientist: learning about wetland restoration and how to observe, collect, record and analyze data.

Assisted by the team leader, students perform two basic tasks for each tree originally planted at the site: locating and measuring. Students must first identify a particular tree from amongst the self-recruiting species now growing at the site. Once found, students measure the tree with specialized measuring tools. While seemingly straightforward, students need to learn how to hold the instruments and work out which part of the tree to measure, especially if it has multiple branches or has suffered damage. A paper-based chart is used to write down the measurements for each tree and any accompanying comments. It also shows the previous data and comments from the last measurement.

Interviews with the environmental scientists, who lead the student teams in the Lilly ARBOR project, revealed how the paper-based method of recording and looking up data can be laborious and susceptible to errors. In particular, they noted how the lack of space on the paper sheets restricts what information can be written down and revisited, having the effect of limiting exploration of observations and hampering hypothesis testing because previous data is not readily available on site. Instead, students have focused on the task of measuring the tree's dimensions, finding it difficult to reason subsequently about the implications of these with respect to environmental issues.

3.2 Requirements

In further discussions with the environmental scientists, we explored what kinds of contextually-relevant information might encourage students to reason more when conducting the measuring activities. Our aim was to replace the paper-based method

of measuring with an electronic version that would enable the students to switch between observation, data collection and analysis. To this end, our primary design goals were categorized in terms of learning and usability, based on a combination of pedagogical objectives and usability design principles.

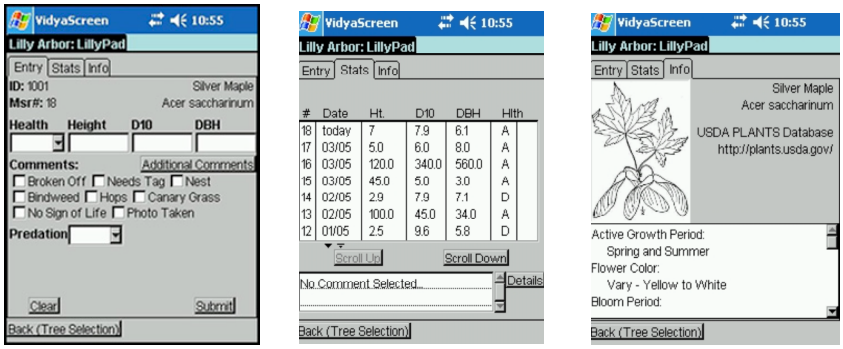
Learning (user experience) Goals. Students should be able to:
- use relevant digital information to understand more about their observations
- share and discuss their observations with other team members
- reflect upon their measuring activities and begin to make inferences about their findings with respect to the planting methods used in the various plots

Usability Goals. The mobile device should allow students to:
- enter measurements and observations into a database (ease of use)
- learn its functionality quickly (learnability)
- use it outdoors while on the move (ease of use)
- discover and locate information (findability)
- read its display in varying environmental conditions (readability)
- show, explain and relay relevant information to others in the team (shareability).

3.3 The Design of LillyPad 1.0

We designed the LillyPad application to provide three core functions: (i) a data entry feature for new measurements and comments, (ii) a historical overview feature showing previously recorded data for each originally-planted tree, and (iii) an information feature showing additional information about the various tree species present at the site. We used a simple and familiar 'tabs' metaphor of interaction, with three tabs representing the functions of data entry ('entry'), historical data ('stats') and additional information ('info'). Clicking on a tab results in a page for that function appearing on the screen (see Fig. 1). The tabs were always visible to enable easy tapping on and switching between. For example, a student could look at the stats page to see previously entered data for a particular tree, followed by tapping on the info tab to see what the leaf for the tree should look like. LillyPad has a page listing all of the trees planted in a given plot and their numbers as an anchor page. Clicking on a tree leads to the data entry page for that tree.

The entry tab page provides a dialog box; data is entered via a combination of checkboxes and a keypad, while comments are entered using a virtual keyboard that pops up at the bottom of the screen. The stats page shows the previous measurements recorded and comments made. This information was designed to help students both locate a tree, and reason about anomalies between the historic data and their current observations. The info tab provides information about the tree species in a small window, together with a thumbnail of professional sketches of the most common parts used to identify a tree (e.g., a leaf) taken from an environmental website (USDA). To see more detail, students could enlarge the sketches to the full screen by tapping on the thumbnails.



(a) (b) (c)

Fig. 1. Screen shots of a) data entry, b) stats and c) info pages for LillyPad 1.0

4 In-Situ Study I

4.1 Methodology

We collected both quantitative and qualitative data to evaluate whether the LillyPad application met our goals:

- logs of page clicks on PDAs throughout the measuring day
- focus group at end of measuring day with all team leaders reflecting on how their team used LillyPad
- commentary by students to roaming researchers throughout the day about their experiences with LillyPad
- vignettes selected from the video material recorded during the day

Having a mix of evaluation methods enabled us to obtain usage pattern data, elicit user feedback (primarily about usability aspects), and observe how LillyPad supported collaborative learning and analysis.

4.2 Procedure

Preliminary user testing of the LillyPad application was carried out by two environmental scientists. Their primary concerns were whether the application was accurate, understandable and easy to navigate. They checked that the database was up-to-date with the appropriate datasets for each plot and tried all functions. We subsequently trained the six scientists who would lead the teams on the measuring day how to use the device. Since technical support would take up to 15 minutes to arrive, they also went through the procedures for what to do when students pressed incorrect buttons or accidentally quit the application. We also designed an outdoor training session for the students, since they would not have the opportunity to become familiar with the application beforehand. Large posters of the most important screenshots were used as visual aids.

On the actual measuring day, eighteen students and eight volunteers from a local corporation that sponsors the program arrived at the restoration site at 8.00 a.m. One of the scientists introduced the restoration project and the three different planting methods used. Six teams were formed, each comprising three students, one or two volunteers and one of the trained scientists. A 10-minute training session was held on how to use the LillyPad application and the PDA (several participants had not used a PDA before). One student per team initially volunteered to be the PDA user. The other students in the team were each given another task and a measuring instrument to use.

Fig. 2. Teams measuring trees in the spring using LillyPad 1.0 and in the fall using LillyPad 2.0

The teams then began to systematically locate and measure the trees in their plot (see Fig. 2). As in previous years, team leaders used any unusual observations, such as if a tree appeared to be missing, as opportunities to probe the students to think about the likely causes. The field day lasted about 6 hours, with a lunch break when the teams had a chance to hear more about the Lilly ARBOR project. Throughout the day, team members switched between using the LillyPad application and the other measuring devices, which was encouraged by the team leader.

Given the physical scope of the project (i.e., a mile long stretch of land), it was impractical for the researchers to observe and record all teams. Instead, we asked a corporate volunteer to video their team's activities with a camcorder we provided. We instructed them to be selective in what they recorded, thereby allowing them to also participate in the group activity. This included videoing measuring the trees using the instruments, the use of and problems with the LillyPad application, and surrounding discussions that ensued. Three researchers roamed the site, staying with one team for an hour or so before moving on to another, while two others remained at base on call should any technical difficulties arise.

4.3 Findings

We analyzed the data in terms of descriptive usage patterns, team leader quotes, summaries of student comments and a detailed analysis of a poignant vignette. These were considered sufficient to assess the learning and usability goals.

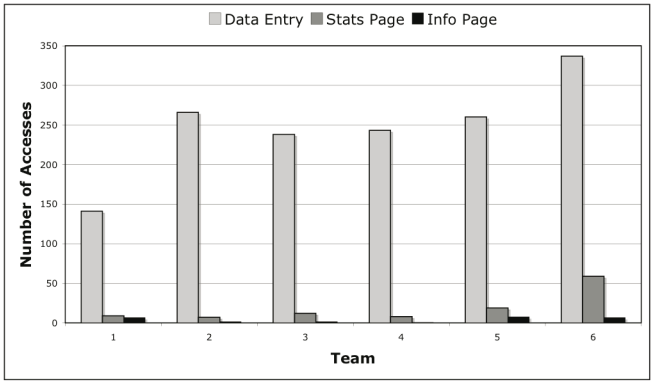


Fig. 3. Mean number of page clicks per team

Usage Patterns: Figure 3 summarizes the page clicks for each team. The number of accesses to the stats and info pages was relatively small, 10-60 for the stats page and less than 10 for the info page. In contrast, the data entry pages were accessed far more, varying between 140 and 330 times per team. This spread reflects, in part, the number of trees surviving in a plot and therefore the number of trees for which data was recorded per plot.

Team Leader Focus Group: All of the team leaders made positive comments about the potential of LillyPad, and said how successful it was for recording data entry. However, they noted that LillyPad was not used very often for other tasks. For example, one team leader pointed out how they *"only used it once but it was very important that one time."* Another pointed out how *"the real advantage was bringing up the stats page so that we could see what a tree was doing multiple times in the past. We found several trees that were missing, and with only the paper then it was missing with no data; but with the device it was very valuable for us to know that this was a beaver-eaten tree covered with reed-canary grass, and that two years ago it was 4cm in diameter."* Another mentioned how it made her change the types of questions she asked, knowing that the students could look up the information on the device that they could not do with the paper-based version.

Student Commentary: Most students learned how to enter data quickly. Several students commented on the difficulty of using the small keyboard to enter data and comments. Some also pointed out how the sketches were not very helpful for identifying trees, and having looked at a couple, they did not bother to access the info pages anymore.

Video Vignettes: In total we collected over 12 hours of video data (2-3 hours per team). The method of selecting certain activities from the total footage that exhibit routines, breakdowns and problems is typical of ethnographic field studies, acting "as a resource, as a set of alerting mechanisms, and as a means of orientation" [13]. One researcher watched all of the videos, marking down and transcribing events where (i) the teams used the PDA to look up information and do any subsequent analysis, and

(ii) there were noticeable breakdowns in communication while using LillyPad. These were viewed with two further researchers who then selected from them a representative set of 10 vignettes to analyze in more detail, showing different teams using the PDA and the problems they encountered.

For (i), the teams worked in an orderly fashion, with different members calling out measurements and comments to the PDA user, who tapped them into the application. We observed the team leaders appropriating the PDA to change their way of engaging students by asking questions that required them to look up information. Rarely did we observe the other team members asking the PDA user for information. It was far more common for the team leaders to ask. The PDA users also rarely showed or read aloud information from the PDA. For (ii) we found that the collaborative process sometimes overwhelmed the PDA user as she translated the multiple measurements called out by the team into numbers, comments and ticks, while simultaneously confirming the entries were correct. During these times, team members had to wait and sometimes repeat their measurement while she completed other parts of the entry task.

While the videos showed how the teams were able to enter data for each of the trees in their plot, the LillyPad application clearly did not meet our learning goals of enabling more analysis to take place whilst in the field measuring. We drilled down on three of the vignettes to explore why this might be the case. Transcribing the minutiae of a poignant moment of an activity, coupled with watching the vignette numerous times, can provide a richer account and interpretation of the interactions within the team, the physical environment and the technologies [15]. It also assists in framing specific recommendations for improving the design. We present one of the transcripts here that reveals the tensions that arose in one team when trying to do both data entry and analysis.

A portion of the vignette is presented in Table 1. The numbers in the text refer to the line in the table. The vignette starts with the team leader (T) noticing a tree that previously had been recorded as dead, re-appearing in the form of a bud (1). Two students (F1, F2) are measuring the height of the tree. A tree appearing to grow after being reported dead is a strange occurrence that warrants reasoning. T is excited and sees this as an opportunity to ask the PDA user (M1) to look up the stats data so they can reason about the tree's disappearance (8). M1 does not heed T's request, but continues to enter data while asking others to confirm what he is entering (3, 10, 13). It appears he is focused on the task and does not 'hear' T. T persists and repeats his request twice (9, 14), yet M1 continues to ignore him. Eventually, T stands up, walks to him, and forcefully gestures at the PDA telling M1 what to do. At this point, M1 does what is asked and brings up the stats page (15). T then reads aloud that the tree has been recorded as dead for the last five years. The other team members marvel and comment on how a tree that has been dead is now alive. M1 continues to be focused on the data entry and does not join in the discussion, only asking how he should record it (20).

It took several attempts by the team leader to access the information that would enable the team to reflect on the unusual sighting. The PDA user clearly focused on

Table 1. Transcript of the team measuring a tree presumed dead but which has grown a new bud

1. T *(team leader crouching next to budding tree holding measuring pole):* "It's come back! That clearly is an Ohio buckeye."
2. F1 *(female student crouching next to him, measuring the height of the tree against the pole)*: "Now are we measuring the flower top or just the stem? I think it's about seven." *T and F1 look over to male student (M1) holding PDA standing 2 feet away.*
3. M1 "Seven point zero?"
4. T: "Yeah. And you can make an estimate for the width. Could be about half."
5. *F1 stands up*. "Yeah, yeah, that was what I was thinking." *F1 crouches down to test her prediction by measuring the diameter of the bud using the calipers.*
6. *F2 (another female student in the team looking on)*: "It is a big flower!"
7. *F1 reads off her measurement:* "Point five zero"
8. T: "We're budding. Rejoice. The tree has resurrected. Let's look at the statistics in there and see how long it has been missing. Is it just one year?"
9. *T waits for a few seconds and then follows up his initial request by being more assertive:* "That will be the middle tab." *M1 still does not reply. T stands up and walks over to M1 and stands in front of him.*
10. *M1 does not look up but asks the others to confirm*. "It's budding you say?"
11. F2: "Yes it's budding"
12. M2 *(a student questions the observations)* "So, we want to figure out when it died?"
13. M1 *(puzzled by M2's comment)* "Once dead, now alive?"
14. *T looks at the PDA screen and points to the data entry accept button:* "Go ahead and accept that. And then look at the stats page." *(Points to the tab on the screen to click on)*
15. *M clicks on stats tab T reads off from stats page:* "Dead, dead, dead, dead, dead, dead, dead, dead. Our every measurement."
16. F2: "Wow, it's been dead?"
17. *M2 reading the screen over M1's shoulder:* "We got"
18. F1: "What a comeback!"
19. M1: "Should I say dead, now alive?" *(returns to task of adding comments)*
20. F1: "Planted and never to be seen for 5 years."

completing the data entry task, ignoring the repeated requests by the team leader. This suggests an inflexibility in our design that needed to be addressed. Specifically:

- data entry is successful but time-consuming
- the PDA user has difficulty multi-tasking when entering data
- the PDA user takes a more passive role during reasoning activities
- the PDA user does not share information from the PDA unless specifically asked.

5 Redesign: LillyPad 2.0

In light of the problems observed with LillyPad 1.0 in the in-situ study, our overarching goal for the redesign was to more fully support analysis *during the measuring activities*. The central objectives were to enable the PDA user to look up relevant data and information *when* it was deemed useful, and to *want* to share and reflect upon this data with the rest of the team. In essence, we wanted the PDA user to

shift from a reactive to a proactive use of the application. We revised our learning and usability goals, accordingly:

- reduce the cognitive demands on the PDA user when entering data by making it less time-consuming and cumbersome
- redesign the stored information to make it more task-relevant and to encourage more active engagement
- include a new set of graphical representations to provide another way of supporting the analysis and reasoning about anomalies
- increase awareness of and reflection on what the other teams are discovering and measuring by enabling communication between teams located in different plots.

5.1 Reduced Cognitive Load

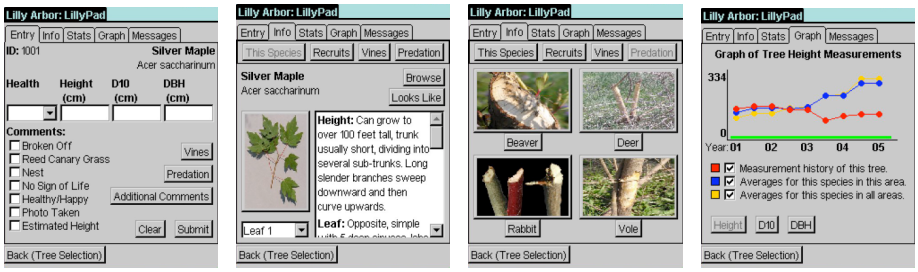
Our first priority was to reduce the data entry burden so that the PDA user can multitask when asked a question or when the team engages in an analysis. We endeavored to improve the interface to make data entry faster, and to make switching between data entry and other tasks easier (See Fig. 4).

Interface Enhancements: We redesigned the data entry page to make it easier and less demanding to fill in. We added white space and enlarged several of the interface widgets to make them easier to select. For example, we introduced a large customized pop-up keypad for easier entry of numerical measurements, reducing the risk of errors. We also included additional checkboxes, thereby reducing the need to type in common comments.

Increasing the size and spacing of the widgets comes at a cost of screen space. The checkboxes could no longer fit on one page, which meant adding sub-pages that appear as pop-up windows. While increasing the navigation path is typically frowned on in mobile application design, the benefits are to make data entry much less cumbersome, including reducing the need for typed comments which our in-situ study found to be particularly problematic in this setting. In addition, the new design should help the PDA users:

- deal with the rapid callouts from the other team members as they could more easily fill in the checkboxes in quick succession
- check that all of the necessary data has been entered in a systematic order
- manage the multiple inputs competently while feeling in control

Two PDAs per Team: We decided to provide half the teams with 1 PDA and the other with 2 PDAs to compare if more analysis would ensue if less work was required by the PDA user. In the 2 PDA condition, one student was assigned the role of 'data entry' and the other as 'information explorer' (i.e. they could view the data, but not enter it). This division of labor allows the data entry person to focus on their role while enabling the other student to look up and share relevant information with the team. We also considered providing each team member with their own PDA but that could have transformed the collaborative activity into individually-based tasks, when



(a) (b) (c) (d)

Fig. 4. Screenshots of LillyPad 2.0: revised (a) data entry, (b) info page, (c) new images of predation, and (d) graphical representation of average tree growth for tree, species and area

our goal was to encourage collaborative learning. Also they would have had to continuously switch between the various measuring activities and holding a PDA, which would have more likely increased cognitive load.

5.2 More Task Relevant Information

We completely rewrote the information pages to support the specific activities involved in tree identification for this particular restoration site. The description of the trees employed a more accessible and enjoyable form of prose. Distinguishing features used for identification appeared first. For example, the text for the Hawthorn begins with "Hawthorns are often affected by crown gall and witches brooms". In addition, we replaced the black and white sketches of the leaves and other identifying features with color photos taken from the Lilly ARBOR site during the fall. Given the next measuring day was scheduled for the fall (where the foliage is quite different from the spring) we wanted to enable them to make comparisons more readily between what they were seeing at the site (see Fig. 3) with what was stored on the Lillypad application (see Fig. 4b-c).

Further design sessions with the environmental scientists resulted in a revised ontology for structuring the information, which included new categories deemed to be more appropriate when identifying, measuring and analyzing. These included the categories of 'looks like', predators, vines and native recruits. Findability was improved by placing photos of the possible vines, trees, or predator damage side by side, so that they could be compared (see Fig. 4c). The rationale was that if a student noticed a tree that had been eaten, covered in a vine, or overtaken by an invasive recruit (e.g., grass), they could easily select a button to obtain relevant information for identifying the predator, vine, or recruit.

5.3 Graphical Representations to Support Analysis

We added a set of graphical representations to visualize the trends and patterns of tree growth over the five years. We thought they would encourage more analysis in situ, since it is easier to make inferences from explicit graphical representations as compared to equivalent numerical data [19]. We wanted the students to have the opportunity to interpret the significance of growth patterns over time in the context of

their ongoing observations and measurements for a particular tree. Three simple line graphs were used to show: (i) the growth of a particular tree over time, (ii) the average growth for that tree species within the current plot, and (iii) the average growth across all of the plots and therefore all of the planting styles (see Fig. 4d).

5.4 Communication Between Teams

We introduced a messaging facility to the LillyPad application to encourage students to communicate their findings and ideas with the other teams and to reflect more globally about the planting method's effect on tree growth. The facility allows students to send short text messages to one another in the different plots at opportune times, such as when they noticed something unusual in their plot (e.g., the oak trees by the river not growing as well as expected). On receiving a message, the PDA users in the other plots could read it out to their team members, triggering the team to reflect upon it with respect to their own measurements (e.g., note if the oaks by the river in their plot were growing less or more). We provided a menu of partially completed messages to make it easy to send messages, such as "our <blank> are doing very well" and "we are seeing a lot of <blank>". We were able to create wi-fi coverage for just over half of the restoration site, using a number of access points and car batteries. Since we anticipated the data entry student to be focused on data entry tasks, we decided to only provide communication to teams with 2 PDAs.

6 In-Situ Study II

During the fall measuring day, a similar number, but different set, of students and volunteers took part. They were divided into teams and trained in the same way as before. However this time, half the teams were given two PDAs and half were given one PDA. The same evaluation methods were used as in the first study. For brevity, we highlight the most interesting results from the logged data, user feedback and video analysis.

Table 2. Average clicks per page types for versions 1 and 2 of the LillyPad application

Page Type	Version 1	Version 2
Data entry	247.5	268
Info	19	48
Stats	4.5	112
Graph	N/A	20

Usage Patterns: The logged data showed a big increase in the usage of pages for the redesigned LillyPad application compared with the first version, as shown in Table 2. To make the comparison fair, the totals were divided by two for the groups with two PDAs. As expected, there was no significant difference for the accesses to the data entry page because the two sets of teams were measuring approximately the same number of trees. However, teams using LillyPad 2.0 accessed the info and stats pages significantly more than teams using LillyPad 1.0 ($t = 4.3$, $P<0.002$ and $t = 2.8$,

P<0.01, respectively). There was very little difference between the teams with one and two PDAs for accessing data entry, stats and info pages. Users in both teams accessed the full range of pages, suggesting that improvements in the design of these pages encouraged greater use. The only significant difference between the one and two PDA teams was the number of graph pages accessed. Significantly more were looked at in the two PDA condition than in the one PDA condition ($t = 6.001$, $P<0.004$), and it was the info explorer who accessed most of them.

Contrary to our expectations, the messaging facility was rarely used. Table 3 shows the entire set of messages sent between the three teams using it, indicating they used it for only a short period. One function was to keep each other informed of progress (in terms of which tree they were on and that it was lunchtime). Another use was to report on unusual sightings. A confusion caused by a typo in a message sent by Area 8 became the topic of conversation for Team 6, where they mention having seen 'catalpzs' among their trees when they meant 'catalpas' (a catalpa is a native recruit, with showy clusters of white flowers, not often found in Indiana). Area 6 misread this and asks them have they seen 'caterpillars' in the trees. Area 8 then reads this as Area 6 having seen caterpillars and asks them on which trees. The video analysis later showed that this misunderstanding sparked a discussion within the team in Area 6 of whether it is possible for caterpillars to be around in the fall. The main reason that teams did not use the messaging facility is that they were too involved in their team's activities. The PDA users did not want to miss out on the discussions and activities that were going on and said that messaging interfered with that. It was considered too distracting; they did not want to be transported to another place, albeit momentarily, as they felt there was enough going on in their own teams.

Team Leader and Student Feedback: The team leaders pointed out that entering data was much easier and the checkboxes quicker to fill in compared with the first version. The students could not think of any problems when entering data or comments about a tree but instead volunteered what additional information could be added (e.g., other images). Some said that the PDA encouraged them to think more about what they were doing. For example, one student mentioned *"It was nice to be able to look up information about the trees, be able to identify it, plus the history, to be able to see if this tree is doing well because a lot of the time you can look at it and say wow that poor little tree has got a lot of competition … So I think it really added to the experience of learning about what it was that we were looking at."* She also commented on the pleasure of interacting with the graphical and numerical data: *"The enjoyment was to look into the history to see what the tree was doing in the last six months, last year."*

Video Vignettes: The videos revealed far more instances of the PDA users sharing information with their team. They took the initiative to contribute to the ongoing activity, reading out information about a particular species, or showing a relevant image which often led to a teammate making a reasoned guess or hypothesis as to why a tree could not be located. This sometimes triggered a more general discussion about what the team was observing in the field and what they were finding out from the LillyPad application (e.g., why a particular species was not growing well close to the river).

Table 3. Text messages sent between the teams with 2 PDAS

10:57:52	Area6	bindweeds are dead
11:36:14	Area7	hi
11:41:12	Area8	we r seeing catalpzs in the trees
11:41:40	Area8	our bindweeds r dead as well
11:46:29	Area6	bindweed dead
11:48:13	Area6	catepillers
11:49:15	Area6	did you mean catepillers
12:00:13	Area8	we have a seedling cottonwood, Lenore is very excited!
12:02:55	Area8	on what are the caterpillers? and what kind?
12:13:41	Area7	is lunch ready
12:21:34	Area7	lunch is ready come get it
13:51:49	Area6	What tree are you on?
13:57:34	Area8	8096

As with the first study, the team leaders tailored their questions in ways that the students could answer with information on the PDA, which sometimes led to more analysis. For example, a team discussed the different rates of growth with respect to the planting method. We saw between 5-10 examples per team of these types of analysis for the one PDA groups, and between 10-20 for the two PDA groups. Both the info explorer and data entry person took part. Illustrative examples of these have been transcribed and analyzed in terms of the interactions and inquiry processes that took place [24].

7 Discussion

This case study has shown how the findings from an in-situ study were used to understand and improve upon the usability and situated user experience of a mobile learning device. The first in-situ study showed the students not using the device other than for data entry and finding this to be time-consuming. Many of the interface changes that were subsequently made to the application led to enhanced usability and encouraged quite a different kind of user experience. The second in-situ study revealed the students enjoying entering data and finding information that in turn encouraged them to engage in more reflective processes. Being able to find pertinent information and share it with others at key moments resulted in discoveries and discussions that were rewarding. Team leaders also noted how the students' interactions with them and each other, together with their shared use of the device, were markedly different from the first version.

While the outcomes of our in situ studies were successful, they were costly in terms of the time and effort involved. Could we have not come up with a much cheaper form of discount usability engineering [22] and achieved the same or even better results by asking a team of experts to predict how it would be used? The answer is, simply, no. Our initial user testing with expert environmental scientists showed them all competently using the LillyPad application and not envisioning any usability

problems. However, placing the device in the palms of students on a cold spring day revealed a whole host of unexpected, context-based usability and user experience problems.

Furthermore, the in-situ setting of our case study revealed how the environment can have a quite different impact on the user experience. In particular, the time of year and the accompanying changes in the foliage affected the way the two versions of LillyPad were used. In the spring the site was barren, making it easy to find trees but hard to identify them as they did not have the typical signs of life, e.g., bright foliage. In the fall, the opposite was true. The site was overgrown, making locating trees more difficult because they were often hidden by grasses, etc., while identifying them easier because of the presence of more identifiable features, e.g., leaves. The cold and clement conditions in the spring and fall, and the time of day also affected the well-being, moods and motivation of everyone. For example, most of the analyses in the second measuring day happened in the morning and very few in late afternoon, when the teams got into a routine and wanted to finish. The effects of and interactions between these situated experience factors made us think quite differently about how to change the design of LillyPad and also our criteria for what counted as successful learning.

Given that in-situ studies are inevitably costly and time-consuming, how do researchers decide upon which methods to use and which of the large amount of potential data they collect to focus on? We used a combination of methods, including logged device data, observations and interviews that enabled a range of data to be collected. A critical part of our analysis was the drilling down on a small number of video vignettes that enabled us to explore concretely the potential and problems experienced by a team when using the LillyPad application as they went about their measuring activities. This provided a 'contextual backdrop' against which to reflect upon the design of the user experience and the mobile device, sensitizing us to how LillyPad *would* (rather than should) be used in practice. It also provided a grounding with which to propose new functions, of which some proved to be successful (e.g. the graphing function) and others not (e.g., the messaging system). Further, this deeper understanding of the situated activities assisted us in explaining why some features were used and others were not.

Finally, it is impossible, and nor is it desirable, to capture everything when in situ. The key is to use various methods that reveal both hoped for and unexpected effects of the context of use. Identifying user experience and usability goals also provides a good framing reference from which to analyze the details of certain events.

Acknowledgments

We thank the Eli Lilly and Company Foundation, the Rotary Club of Indianapolis and the Pervasive Technology Labs, Indiana University for funding the project. Thanks also to to Kara Salazar and Polly Baker at IUPUI and Allen Lee, CJ Fleck, Nick Gentille and Anne Stephenson at IUB for their various contributions. Finally, we thank all the team leaders, students and volunteers who participated in the measuring days.

References

1. Bell, G.: No More SMS from Jesus: Ubicomp, Religion and Techno-spiritual Practices. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 141–158. Springer, Heidelberg (2006)
2. Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown, B., Rowland, D., Benford, S., Hampshire, A., Captra, M.: Interweaving mobile games with everyday life. In: Proc. of CHI, pp. 417–426 (2006)
3. Bellotti, V., Back, M., Edwards, K., Grinter, R., Henderson, A., Lopes, C.: Making sense of sensing systems: five questions for designers and researchers. In: Proc. of CHI, pp. 415–422 (2002)
4. Benford, S., Seager, W., Flintham, M., Anastasi, R., Rowland, D., Humble, J., Stanton, S., Bowers, J., Tandavanitj, N., Adams, M., Farr, J.R., Oldroyd, A., Sutton, J.: The error of our ways: the experience of self-reported position in a location-based game. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, pp. 721–730. Springer, Heidelberg (2004)
5. Benford, S., Schanädelbach, H., Koleva, B., Anastasi, R., Greenhalgh, C., Rodden, T., Green, J., Ghali, A., Pridmore, T., Gaver, B., Boucher, A., Walker, B., Pennington, S., Schmidt, A., Gellersen, H., Steed, A.: Expected, sensed, and desired: A framework for designing sensing-based interaction. Proc. of TOCHI, 12:1, 3–30 (2005)
6. Boucher, A., Gaver, W.: Developing the drift table. Interactions 13(1), 24–27 (2006)
7. Brignull, H., Izadi, S., Fitzpatrick, G., Rogers, Y., Rodden, T.: The introduction of a shared interactive surface into a communal space. In: Proc. of CSCW, pp. 49–58 (2004)
8. Consolvo, S., Walker, M.: Using the experience sampling method to evaluate Ubicomp applications. IEEE Pervasive Computing Mobile and Ubiquitous Systems 2(2), 24–31 (2003)
9. Consolvo, S., Everitt, K., Smith, I., Landay, J.: Design requirements for technologies that motivate physical activity. In: Proc. of CHI, pp. 457–466 (2006)
10. Crabtree, A., Rodden, T.: Domestic routines and design for the home. Computer Supported Cooperative Work: The Journal of Collaborative Computing 13(2), 191–220 (2004)
11. Dumas, J.S., Redish, J.C.: A Practical Guide to Usability Testing. Ablex, Norwood, NJ (1994)
12. Gaver, W., Dunne, T., Pacenti, E.: Cultural probes and the value of uncertainty. Interactions 11(5), 53–56 (2004)
13. Hughes, J.A., Randall, D., Shapiro, D.: Faltering from ethnography to design. In: Proc. of CSCW, pp. 115–122 (1992)
14. Hull, R., Reid, J., Geelhoed, E.: Creating experiences with wearable computing. IEEE Pervasive Computing 1(4), 56–61 (2002)
15. Hutchins, E., Klausen, T.: Distributed Cognition in an Airline Cockpit. In: Middleton, Engeström, Y. (eds.) Communication and Cognition at Work, pp. 15–34. Cambridge University Press, D. Cambridg (1996)
16. Intille, S., Larson, K., Tapia, E., Beaudin, J., Kaushik, P., Nawyn, J., Rockinson, R.: Using a live-in laboratory for ubiquitous computing research. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) PERVASIVE 2006. LNCS, vol. 3968, pp. 349–365. Springer, Heidelberg (2006)
17. Kidd, C., Orr, R., Abowd, G., Atkeson, C., Essa, I., MacIntyre, B., Mynatt, E., Starner, T.: The Aware Home: A Living Laboratory for Ubiquitous Computing Research. In: Streitz, N.A., Hartkopf, V. (eds.) CoBuild 1999. LNCS, vol. 1670, pp. 191–198. Springer, Heidelberg (1999)

18. Kjeldskov, J., Skov, M., Als, B., Høegh, R.: Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In: Brewster, S., Dunlop, M.D. (eds.) Mobile Human-Computer Interaction – MobileHCI 2004. LNCS, vol. 3160, pp. 61–73. Springer, Heidelberg (2004)
19. Larkin, J., Simon, H.: Why a diagram is (sometimes) worth ten thousand words. Cognitive Science 11, 65–99 (1987)
20. Lin, J., Mamykina, L., Lindtner, S., Delajoux, G., Strub, H.: Fish'n'Steps: Encouraging Activity with an Interactive Computer Game. In: Proc. of Ubicomp, pp. 261–278 (2006)
21. Mankoff, J., Dey, A., Hsieh, G., Kientz, J., Lederer, J., Ames, M.: Heuristic evaluation of ambient displays. In: Proc. of CHI, pp. 169–176 (2003)
22. Nielsen, J.: Usability engineering at a discount. In: Salvendy, G., Smith, M.J. (eds.) Human-Computer interaction on Designing and Using Human-Computer Interfaces and Knowledge Based Systems, pp. 394–401 (1989)
23. Olsen, A., Rogers, Y., Sharp, H.: The Snap Method. In: Workshop Proceedings, Designing Methods for New Users, Technologies, and Design Processes, CHI (2007)
24. Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W.R.: Mobile technologies for integrated scientific inquiry. Journal of Learning Sciences (submitted)
25. Rogers, Y., Price, S., Randell, C., Stanton-Fraser, D., Weal, M., Fitzpatrick., G.: Ubi-learning: Integrating outdoor and indoor learning experiences. Comm. of ACM 48(1), 55–59 (2005)
26. Rowan, G., Mynatt, E.: Digital Family Portrait Field Trial: Support for Aging in Place. In: Proc. of CHI, pp. 521–530 (2005)
27. Salovaara, A., Jacucci, G., Oulasvirta, A., Saari, T., Kanerva, P., Kurvinen, E., Tiitta, S.: Collective creation and sense-making of mobile media. In: Proc. of CHI, pp. 1211–1220 (2006)
28. Scholtz, J., Consolvo, S.: Toward a Framework for Evaluating Ubiquitous Computing Applications. Pervasive Computing 3(2), 82–88 (2004)
29. Sharp, H., Rogers, Y., Preece, J.: Interaction Design, 2nd edn. Wiley, Chichester
30. Taylor, A.S., Swan, L.: Artful systems in the home. In: Proc. CHI, pp. 641–650 (2005)
31. Toscos, T., Faber, A., An, S., Gandhi, M.: Chick Clique: persuasive technology to motivate teenage girls to exercise. In: Proc. of CHI, pp. 1873–1878 (2006)
32. Winograd, T.: Bringing Design to Software. Addison Wesley, Reading (1996)
33. Yeh, R., Liao, C., Klemmer, S., Guimbretière, F., Lee, B., Kakaradov, B., Stamberger, J., Paepcke, A.: ButterflyNet: a mobile capture and access system for field biology research. In: Proc. of CHI, pp. 571–580 (2006)