

Issues in Evaluating Ambient Displays in the Wild: Two Case Studies

William R. Hazlewood, Erik Stolterman, Kay Connelly
Indiana University School of Informatics and Computing
Bloomington, IN 47405, USA
{whazlewo, estolter, connelly}@indiana.edu

ABSTRACT

In this paper we discuss the complex task of evaluating ambient displays, concentrating on issues within in-situ deployments. We start by describing how these technologies have been evaluated in lab settings, where the focus has been primarily on issues of usability, and argue strongly for the necessity of in-situ evaluation. We then present two case studies involving in-situ evaluations, and from these derive issues that hindered the researchers from being able to delve more deeply into the overall impact of their implementations. We conclude with our own suggestions on possible alternatives to explore for evaluating ambient displays, which are based on the issues derived from our case studies.

Author Keywords

Ambient Displays, Feedback, Evaluation Studies, Case Studies, Methods.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Design, Human Factors, Measurement, Performance

INTRODUCTION

Over the last decade the advent of ubiquitous computing research has piqued an interest in alternative displays that can provide useful information while blending smoothly into the surrounding environment. These devices are distinguished from more typical informational displays in that they are designed to be minimally attended and perceivable from outside of a person's direct focus of attention, providing a level of pre-attentive processing without being unnecessarily distracting. Such technologies are intended to be embedded in existing environments, often appropriating unused physical and visual aspects of everyday objects to provide an information channel that can be *easily ignored* when there are more important matters requiring attention [19]. Because they can be easily “tuned out,” informational displays such as these may offer a solution to the possible threats of information overload

some have predicted as a side effect of ubiquitous computing environments becoming more common [26]. The challenge we face is how to provide in-depth evaluations on something that is defined as blending with the surrounding world, and meant to be (in some respects) ignored.

While many interesting ambient display technologies have been explored over the last decade, there is a lack of research discussing the difficulty involved in *evaluating the general impact and effectiveness of such technologies*. The lack of research on the evaluation of ambient displays is due in part to the sheer difficulty involved in developing prototypes and implementing user studies for this class of technologies [5]. Since ambient displays are designed to be highly subtle and used indirectly, traditional user interface evaluation methodologies targeting task-oriented users do not provide enough methodological support for the study of how these technologies are experienced in actual contexts of use, or how people's relationship with the technology changes over long periods of time.

By definition, these technologies do not function as intended until they have properly blended into the everyday environment [14], which is difficult to accurately simulate inside a controlled laboratory. Studying these technologies is further complicated in that a sort of paradox arises where intrusively observing the use of an ambient display hinders it from functioning as intended. Direct observation and probing from researchers is a constant reminder that something out-of-the-ordinary is present, and directs the subject's attention unfairly. It is likely that the more aggressively or intrusively one tries to observe people's interactions, the less these results will be an accurate reflection of actual use. Hence, special care must be taken when designing evaluative studies on ambient displays. Since few long-term or in-situ studies have been conducted, it is not clear how to accurately gauge the effectiveness or impact of systems with such atypical styles of use.

In this paper we argue that ambient displays, in all their many forms, constitute a new form of user experience that is radically different from the traditional focused task-oriented user situation. As a consequence we argue for methodological development when it comes to evaluation of these technologies. This paper is based on two case studies, both examples of ambient technology evaluated in real-world settings. We conclude the paper with a

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Copyright 2011 ACM 978-1-4503-0267-8/11/05...\$10.00.

discussion of the insights on evaluation derived from the two cases. We also present some possible directions for further methodological development when it comes to evaluation of ambient technology.

RELATED RESEARCH

The concept of an ambient display has its genesis in Mark Weiser's seminal paper, *Designing Calm Technology*. In this paper Weiser describes a project he called "The Dangling String," which consists of a simple cable hanging in the corner of a room that twitches whenever a packet of information is sent over the local network. If there is little network traffic, the string hangs motionless, but as traffic increases the string wiggles and twitches chaotically. While there are a number of ways to inform people of the local level of network traffic, the advantage of this particular display is that it conveys information in a simple and subtle manner, allowing people in the office to *just know* when the network is under heavy usage without having to think about it. Weiser proposed that this form of representation would allow information to move easily from the periphery of our attention, to the center, and back. So, despite being ubiquitous, such displays could also be easily ignored or disregarded when more pressing concerns required one's attention [24].

Hiroshi Ishii's group at the MIT Media Laboratory further explored the concept of ambient displays by producing several unique implementations, such as ambientROOM [14], the Water Lamp, and Pinwheels [8]. Examples like these allude to several potential advantages from designing technologies that are subtle and make use of our peripheral awareness. However, since the majority of such implementations have not been evaluated [16], and no agreed upon evaluation metrics exists, it is difficult to articulate their overall value.

Evaluations on Ambient Information Devices

While there have been several innovative ambient displays with little or no formal evaluation [e.g., 4, 12, 15, 20, 24], there are some early implementations which have provided important initial evaluations in an attempt to expose how people react to ambient displays, for example LumiTouch [6] and AudioAura [18]. These examples, however, report few details, and focus mainly on informal user feedback. The LumiTouch was given to users who were not emotionally involved with one another, and after using it for a short time comments were collected regarding each individual's experience. In AudioAura nine participants were given a brief introduction, and then asked to complete a set of self-paced tasks, followed by a questionnaire where comments and suggestions were encouraged.

Whereas AudioAura and LumiTouch were intended for designated users, many ambient displays have been designed for public spaces where the information presented can be absorbed by anyone who happens to walk by [e.g., 12, 15, 25]. As with the previous examples, evaluations for these installations typically involve informal interviews

focused heavily on the initial levels of *enjoyment* when interacting with the display but little in the way of long-term effects or general impact.

Laboratory Evaluation Styles

In more elaborate evaluations of ambient displays, methods have focused mainly on the *functionality* and *usability* of the implementation itself. For example [2] describes a device called the "WaterBot" which was developed to encourage water conservation by providing subtle visual and audio cues to those using a water faucet. It was evaluated in a controlled study where ten participants were asked to wash their hands thoroughly in a sink with the WaterBot installed. To gauge the intuitive aspects of the device, participants were not told any details about the device, or its intended function. As they washed their hands repeatedly, the WaterBot cycled through its modes as the researchers observed and recorded the subject's comments. This evaluation provided important insights into how the users made sense of the WaterBot's cues, which could be generalized to show how people might interact with devices they have never before encountered. However, since the study was conducted in a controlled laboratory where participants were requested to complete a task under observation, it is unlikely that the WaterBot ever became embedded into the participant's environment as Weiser suggested. Because of this, we have no real knowledge of the long-term impact this display could have on behavior. It could very well be that during regular use, those using WaterBot would simply become unaware of its presence, or just ignore it intentionally once they were not being observed.

In [7], a system called "eye-q" was developed in which two small LEDs were embedded into a pair of eyeglasses. These LEDs were intended to deliver simple visual cues in the wearer's peripheral visual field without disrupting their vision, or diverting their existing focus of attention. To evaluate this display, a thorough set of experiments were conducted to test the subject's ability to perceive the cues given by the glasses while the subject was focusing on separate primary tasks, which required varying levels of concentration. The study showed that the display was less noticeable when users were under higher workload conditions. This result implies that the information from the display *can* inherently shift to the background when the user needs to focus on something else, but just as with the WaterBot, the eye-q studies were conducted in a controlled laboratory setting, where it is likely that the participants were more sensitive to the display than they might be in a natural setting. It is still unclear whether this display would function as it did in the lab or if it would become one of the many notifications people simply tune out permanently.

An evaluative framework was developed specifically for ambient displays in [16], where the authors successfully modified a version of Neilson's heuristics for evaluating usability, to specifically address the design of ambient

displays. However, while this method is good for locating problems of usability due to design flaws, such analyses tell us only about the *potential* success of an ambient display, not how such displays function once no one is prompting their consideration.

We have come to realize that there are limits on what can be learned about ambient displays in a laboratory setting, particularly implementations that do not address specific tasks or goals. To advance research in ambient displays *new evaluation methods must be explored* that can account for the special requirements necessary for these technologies to function as intended, particularly the blending of these displays into the environment.

Our conjecture is that new methods will require studies focused specifically on *in-situ* deployments conducted over *long periods of time*. This arrangement seems to be the only way to allow a display to become truly *ambient* within a given environment. However, studies like this are difficult and time consuming since they often involve observing low intensity usage styles (e.g., only brief glances at a display) [13]. Also, the ambiguous nature of *use* of these displays demands a rethinking of evaluation. As Hallnäs and Redström state, technologies like these are neither works of art, nor tools with explicit uses, even though they share properties with both [10]. Our goal is to explore more broadly how such studies could be conducted, derive exactly what factors can be measured, and determine what these factors can tell us about the impact ambient displays have in actual contexts of use.

In the following sections, we present two case studies where an ambient display was developed and evaluated in-situ. These case studies are presented in brief, but full accounts can be found in [11] and [21]. These studies were initially not designed as studies of evaluation methodology, but both present evaluation challenges that initiated this paper and our examination of evaluation methodology. The two studies each offer a set of evaluation issues and have provided us with the material necessary for a deeper discussion on evaluation.

CASE STUDY 1: STUDENT FEEDBACK ORB

In the first study, a simple ambient display was developed and deployed in an exploratory long-term, in-situ, study. Considering the research motivations, an initial set of four rubrics was derived for this study's development.

(i) Make use of a simple mechanism

The authors did not want to complicate their study with dense or highly abstract mappings of symbols to meanings, or use anything that would require significant training in order to be understood.

(ii) Provide an information source that was of high interest to the participants

The information channels commonly used for ambient displays typically include sources that can be constructed from publicly available data streams. To create a system

that provides information that is potentially more engaging, the authors chose to create a custom information channel specifically tailored to the participants' interests.

(iii) Be conducted outside of a laboratory setting

The authors wanted to observe a display embedded at a level where participants interact with it naturally, and figured that this would not be accomplishable unless the display was placed in the participants natural setting (i.e., a real home or office).

(iv) Run for an extended time

This rubric was to ensure that the display became embedded into the participant's surroundings, and fully integrated with daily routines. Also, this time frame allowed the display to overcome any initial novelty effects.

To satisfy (i), the authors made use of an existing commercial device known as the "Ambient Stock Orb" from AmbientDevices.com [1]. For (ii), the authors focused on university instructors as the participants in the study, and created an information channel that allowed students to provide feedback regarding their learning experiences. For (iii), the authors simply give an Ambient Orb to each of the participating instructors, and told them to situate the orb wherever they felt it was most appropriate. Finally, for (iv) the system was structured so that it could run automatically, logging information for as long as the study conditions would allow.

The Ambient Display

The final system consisted of an Ambient Orb set to reflect the overall daily confidence levels of students within a particular course, and convey this information to the instructor using a predefined color scale. Additionally, a way for the instructors to view more detailed feedback information was provided in case they were curious as to why their orb had changed one way or the other. Below is a detailed description of the different components of this display.

The Ambient Orb

The Ambient Orb consists of a small sphere made of frosted glass containing an array of multicolored LEDs (see Figure 1, bottom right). These LEDs allow the orb to shift smoothly between thousands of different colors, which are configured by way of an embedded cellular chip communicating with a national pager network in the United States. The result is an information appliance that does not require an existing network connection to function.

The Custom Information Channel

To provide a custom information channel, the system sent automated emails to each participating student, asking them to give a general confidence rating over their understanding of the materials covered in class. This email was sent after each class, and consisted of an HTML formatted message containing a single Likert scale which the students could click (see Figure 1, top). To supplement the data represented on the orb, a separate information page was

developed which the instructor could reference through the web. This page contained a histogram showing the actual scores received in real time, and a small bit of extra statistical information (see Figure 1, bottom left).

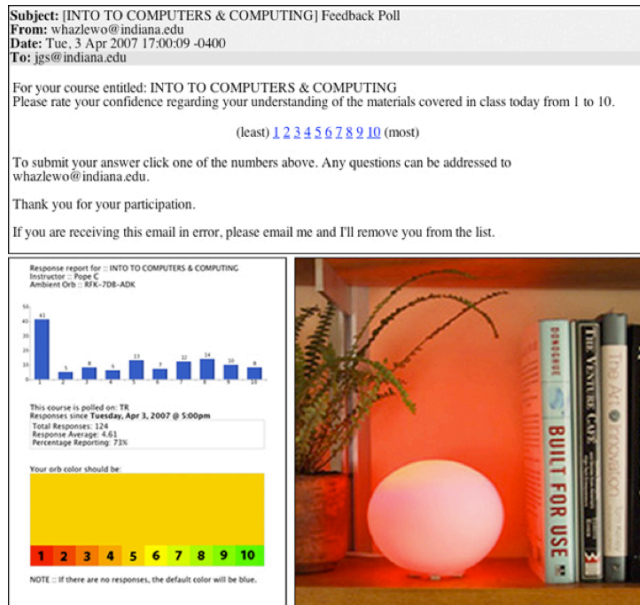


Figure 1: The automated email (top), Informational website (bottom, left), and the Ambient Orb (bottom, right)

Evaluation

Six instructors were gathered from four different departments at the authors' university to participate in this study. Each of the instructors were met with individually, given an orb, and instructed on how it functioned. The instructors were told how to tell if the orb is working correctly and that they should be able to place it anywhere they desired.

Once the orbs were working, and the instructors understood how to use them, the authors began recruiting students from each instructor's class. During this recruitment, students were informed that participants would be receiving an automated email after each class asking them to answer the following question:

"Please rate your confidence regarding your understanding of the materials covered in class today from 1 to 10."

It was explained that to respond they simply had to click on one of the hyperlinks embedded inside the email, that their participation was voluntary, and that the instructor would have no knowledge of who was participating. This was important so that the students were comfortable in giving negative responses.

The instructors were informed that the orb was set to show the average score given by the students on a scale of 1 to 10, with 10 being the best, that the scores were mapped to a color scale ranging from red to yellow to green, and provided the URL for accessing their own supplemental information page. Finally, it was explained that the

instructors were not expected to use the system one way or another and that they should feel free to check the information page anytime they wished.

Abridged Study Results

Except for two instructors (*Instructor 2* and *Instructor 3*) who joined after the initial study had begun, the system ran for 68 days. Since this was an exploratory study, there was no specific phenomenon being measured. The authors were mostly interested in any interesting outcomes that could be used to inform future in-situ studies of ambient displays.

The observation process was made difficult by the fact that the authors constrained themselves from directly observing the participants. The reason for this is that they believed that continually probing the instructors would constantly force the display in the foreground of the participant's attention, keeping it from blending into the surrounding environment and becoming *truly ambient*. The authors chose instead to monitor whatever information was possible *without* disturbing the instructors during the study. However, the pool of non-intrusive monitoring sources was very sparse. The authors had access to the web server hosting the information web page for each instructor's orb, so the frequency which the instructors accessed this resource could be monitored, and they could observe how the students were reporting responses to the biweekly polls. Lastly, they knew structured interviews could be conducted without contaminating the use of the display after the study had concluded. Described below are the observations made along with the interpretations of the authors.

Information Web Page Access

It was surprising how infrequently some of the instructors accessed the informational web page for their orbs. The authors had predicted that the majority of instructors would access the informational site somewhat frequently at the beginning of the study until they became more accustomed to reading the Ambient Orb, but as shown in Table 1, several of the instructors only accessed the informational web page a few times over the entire course of the study. Towards the midpoint of the study an email was sent out reminding the instructors of the informational web page's existence, which may account for the *only* accesses by some of the instructors. Instructors 1, 5, and 6, were the most infrequent visitors to the information web page. During the post-study interviews, the instructors were asked to give their opinion on the value of having the Ambient Orb paired with the informational web page.

Instructor 5 responded:

Instructor 5: "I preferred to look at the website, quite frankly, because it gave more information. The average is kind of a limited score, I preferred the website because of its additional information, and I know how many responses are contributing to the score. Some may not care, but for me, from a research background, teaching, I obviously want to know more information."

However, later in the same response this instructor states:

Instructor 5: "I wouldn't say the orb was a waste, I used that as a barometer on how frequently to go check the web site. I didn't check the website necessarily a whole lot. I checked the orb a lot, and if I'd seen changes I would have gone to the website."

Instructors 1 and 6 both agreed that while they liked the information in the web page more than the orb, they did not feel that they had to check the web page unless prompted to by the orb.

From one point of view, such statements support the value of this display. It was clear that the information was valuable to the instructor, but the Ambient Orb acted more as a filter that informed the instructor that they did not *need* to bother checking the detailed information on the web page. However, during the study a concern was that the lack of access to the informational web page meant that the instructors had simply forgotten about its existence. Despite the initial desire to avoid polling the instructors, the authors decided to send out an email reminding them about the informational web page. The email had a URL embedded in it which they most of them clicked, but even after this reminder there were few accesses to the page. This event is an indication of the difficulty with non-invasive studies. The authors found the lack of access to the web a problem with the study, while it could as well have been an indication of the opposite. That is, that the instructors felt satisfied with the Orb and the ambient information.

To better understand the instructor's own perception of their reliance on the system, they were asked to report how often they made use of the web page. Their responses support what is shown in the logs:

Instructor 1: "I didn't check it often, [the orb] was pretty much always the same color."

Instructor 5: "I'd guess I've checked it about 5 or 6 times. I would have checked it more often had I had more change in the color. That would have prompted me to check again. I had such consistency in color that I didn't go look at the web much. When I did it was just because I was curious about the distribution."

These responses indicate that the instructors were indeed *aware* of the web page's existence, but never felt a compulsion to access it.

Observations on Evaluation Issues

Several issues in regard to the challenges of setting up in-situ long-term evaluations of ambient technologies were observed in this study. Some of the most important findings are:

Limited forms of observation. The non-intrusive observation methods that the authors restricted themselves to made it extremely difficult to collect useful information that would expose anything meaningful about the display.

However, it is the nature of in-situ studies that one has little control over the environment in which a technology is situated. Since we cannot reliably predict what variables will be beneficial in such evaluations, it becomes more important to look for ways to record *anything* that can be logged so that this data is available when researchers realize they need it *post hoc*.

No control of context. While the authors' ambient display functioned exactly as it was designed, the custom information channel did not record any radical events to observe. It would have been beneficial to their analysis to have an occasion where the students reported an extreme lack of confidence over a topic so that the instructor's orb would have dipped into the red. Such an event would have provided something to focus on to during the concluding interviews. Unfortunately, as stated above, when studying displays in-situ, one has little influence over the occurrence of interesting events.

Non-task oriented. The information device discussed here was not intended to address any specific task (e.g. alter the pacing of course topics, or increase the average grades of the students), it only provided information the authors believed could be somehow beneficial to the instructors; information that is just *good to know*. Since the authors were not trying to alter any specific opinion or behavior, it was difficult to know in advance what would potentially be of interest to measure as an effect of the use of the device.

Difficulty measuring awareness without contamination. The study was aimed at examining how the device would influence the instructors' general awareness of their students' confidence. The concluding interviews provided some indications, but while the system was deployed it was difficult to examine the thoughts of the subjects without influencing their actions and personal reflections.

Accounting for overly subtle presentation. While the display functioned technically, for the purposes of this evaluation the method of presentation was far too subtle. Of course the rate of change in an ambient display will influence how near or far it is to the observer's direct focus of attention, but if it is too static, then the display may become nonexistent to the observer. To increase the noticeability of this display the authors could have set it to display rate-of-change rather than the averages, and still have something that contains the specific qualities of an ambient display.

CASE STUDY 2: CLOUDS AND LIGHTS

The second study explores the evaluation issues that arise when developing an ambient display for a shared public space rather than one designed to convey information towards a specific user. To account for a more general audience, displays such as these have to present information content that is *less* personal, and somewhat generalized, such as weather or stock values [9]. For example, [17] describes a public display that conveys the relative

economic strength of the Euro, the Dollar, and the Yen, using the heights of water jets within a stylized fountain. Another display within this theme is “The Source,” which is an ambient display developed for the London Stock Exchange. This display consists of a grid of 162 steel cables running from floor to ceiling within the 32-meter high atrium in the center of the building. This system of cables supports a 9x9x9 matrix of 729 translucent spheres, which can travel up and down the cables, and illuminate in a manner similar to the Ambient Orb. Throughout the day the form and motion of the installation are determined by real-time stock information. At the end of the day the spheres return to their cubed arrangement, and the spheres illuminate to show the stock market’s closing price.

With each of these displays, the information presented is not directed at any particular individual. Instead, the information is simply put in the open, with no specific intention as to how this information is to be appropriated by those walking by. Again, it is simply information that is potentially *good to know*. How then, does one go about evaluating the impact of an informational display such as this? One way would be to place such a display in a location where the information would have an effect on some observable behavior, and watch to see how that behavior changes after deployment.

In the study discussed here, a series of three large-scale ambient displays was deployed in a shared space, with the intent of altering a simple behavior. Similarly to “The Source,” these displays were situated within the atrium of a newly constructed building. Atriums such as these are ideal for ambient displays, as they often act as a channel through which many people have to travel from one location in the building to another. The behavior selected for augmentation was whether people decided to take the stairs or the elevator when moving through the building. This was done by constructing a system to log the usage of these resources, along with a series of three separate displays conveying stair vs. elevator usage in different ways. The construction of this system was no trivial task, requiring over nine months of research and development, and consultation from a wide range of experts.

The Ambient Display(s)

Rather than a single display, three different displays were implemented. These displays were intended to work together as a single system in order to subtly alter the behavior of those within the building. Each of these displays is briefly described below.

The Clouds

This display consists of two separate clusters of spheres, colored orange and gray, which move up and down independently. The orange spheres represent the usage of the elevators while the grey spheres represent the stairs. The relative heights of the clouds change depending on the number of people who use the stairs versus the elevator, repositioning themselves every quarter hour. The higher up

the orange cloud the more people are using the stairs and the higher up the grey cloud the more people are using the elevator (Figure 2, left).

Follow the Lights

This display is designed as an abstract representation intended to be playful and interactive, with the aim of luring and encouraging people to take the stairs. It was implemented using a series of interconnected LEDs embedded into the existing carpet tiles (Figure 2, right). As someone enters the building and walks toward the elevator, the lights begin to twinkle in a pattern that flows toward the stairwell. If the same person continues on to the elevator, the lights begin to slowly pulsate red, as if to imply that the display is becoming annoyed at the person ignoring its suggested path.



Figure 2: The Clouds (left) and Follow-the-Lights (right)

The History

This display consisted of a simple set of dynamic pie charts displayed on a large existing plasma display situated near the entrance of the atrium (Figure 3). The design was deliberately more literal than the other two, and gave those within the building a historical perspective of the stair vs. elevator usage throughout the week.

Logging the amount of stair and elevator usage was done by placing pressure sensitive mats beneath the carpet at the thresholds of each stairwell and elevator within the building. The data from these mats was forwarded to a central machine and aggregated.

Evaluation

In order to see if behavior had changed the authors developed a baseline data set by running the pressure mat sensors for six months prior to the deployment of the ambient displays. Once the baseline data had been collected, the system of displays was installed over a weekend when no one was in the building. The authors wanted to record people’s initial reactions, and observe how they went about making sense of the displays naturally. The authors decided to propagate information about these displays only by word of mouth. All project members were told to explain the displays and the general project goals to

anyone who asked directly, but avoid making any sort of broadcast announcement.



Figure 3: The History

With the system deployed, an 8-week evaluative study was initiated. This study was conducted using a mix of different data collection methods, including: observations and interviews *in situ*; an online survey sent to all building occupants after four weeks; and the data logged by the sensor network. The first two weeks after deployment consisted exclusively of remote observation. This involved project members inconspicuously positioning themselves within the atrium, and taking copious semi-structured notes using a template generated prior to the study. This template noted group size, trajectories of people moving through the space, whether they stopped and discussed the displays, and what they discussed. In weeks 2-4 interviews were conducted where project members approached 25 people randomly as they entered the atrium. These people were asked whether they had noticed the displays, and what exactly they thought the displays meant. This included note taking as before, along with some audio recording. At the midpoint of the study, an email was sent out to the entire population of the building (approx. 200 people), asking them to participate in an online survey. The four-week mark was selected because it was predicted that this was adequate time for most people to have noticed the display, and for the initial novelty factor to wear off. The survey consisted of 13 questions regarding their use of the building, where they were located in the building, and what their impressions/interpretations were of the various displays. Finally, to analyze the actual behavior versus perceived behavior, the responses to this online survey and face-to-face interviews were compared with the log data recorded by the physical sensors.

Abridged Study Results

In terms of people's ability to understand the meaning of these displays, *Follow-the-Lights* was found to be the most intuitive according to the survey and interviews. By week 4 of the study most had figured out that Clouds were supposed to represent some relationship between the stairs and the elevators although few knew the exact mapping

between the display and the information it was providing. During interviews people reported a wide variety of interpretations as to exactly how this mapping worked. Additionally, many reported that they would just prefer to have a simple numerical representation rather than the abstract displays.

In terms of behavioral change, when people were asked whether the displays had caused them to alter their behavior, the vast majority stated that the displays had no effect on them personally. However, when the baseline data was compared with what was being captured by the floor sensors after the displays had been deployed, a statically significant change in the ratio of elevators versus stairs was detected. This implies that the displays were having some sort of effect on people of which they were not even aware.

Observations on Evaluation Issues

As with the previous case, this study provided several insights into the issues that arise when conducting such evaluations. Some of our findings are:

Notions of 'user' and 'use' are ambiguous. When observing/evaluating a public ambient display like this one, it's not clear exactly who the users are. Technically, it is *anyone* who could potentially occupy the space. This could include the people who work in the building, the cleaning crews, the caterers, and any visitors. In addition, it is not clear how to evaluate *use* since we have not established what it means to be *using* the display (e.g., are we using it just by being in the same space, or does it require direct observation and consideration? When exactly are we *not* using it?). This issue makes it difficult to establish any precise expected outcome of the system, and further complicates the decision as to what should be measured. How exactly does one conduct a user study when they do not have a specific user, or a defined task?

To what extent is it acceptable to directly interfere with people in the study? In this study, people in the building were observed and interviewed frequently, which of course influenced the overall public awareness of the system. At the same time, direct input from the public provided rich insights into how people value, and think about, the display. This hints at a possible tradeoff between preserving the natural environment in which a display is employed, and the quality of data collected via the study.

Discrepancies in self-reflection. The opinions and behaviors reported in the interviews did not match what was captured with the sensors logging actual use. While it has become agreed upon that people are not generally good at self-report, this can be seen as a potential opening for evaluations that compare and contrast data from several sources. How to accomplish this (i.e., what sources and how to gather data) is not obvious.

DISCUSSION

The goal of an in-situ study is to observe a technology or practice in a natural context of use. Of course any time we

observe or probe for insights, we are having some sort of effect on the *authenticity* of this natural context. With many technologies this lack of authenticity does not hinder researchers from collecting valuable insights, but in the case of ambient displays, which are intended to blend into the surroundings, it would seem that *any* amount of probing would contaminate the potential findings.

In the case of many public ambient displays, there is no specifically targeted user, no uniquely specified task, and it is not always the case that the designer has any real intent as to how the information being presented is to be used. In some cases, studying the use of a real ambient display *in-situ* may be impossible. For instance, one of the pre-defined information channels for the Ambient Orb is the U.S. National Threat Level, which has only gone red once since its creation in 2002. It is not necessarily the case that this is a bad display, or that people would not be aware of it. There would just not be many ways in which researchers could establish the overall usefulness of such a display until some national catastrophe was to occur. While there are available ways to incorporate some form of gaze/glance detection into the design of a display, there does not seem to be a way to measure how frequently someone inadvertently siphons information from something that enters our field of view.

Implications for research

Based on our findings we are convinced that evaluation “in the wild” of ambient displays poses severe problems. We have discussed some of them above with the purpose of making them as visible as possible. Our assumption is that there will be a continued increase in use of ambient technology in all forms. So, the question then becomes, how can HCI research approach this evaluation challenge? Here we discuss some potential strategies.

The methodological approach

The most obvious and natural direction would be for the HCI research community to engage in a continuous methodological development based on the rich knowledge already existing in the field in regard to evaluation and testing. There exist already a large number of evaluation methods and techniques that might be possible to adapt and redesign. Of course, this development has to be carefully crafted to engage with the unique issues related to ambient technology.

For example, to address the lack of anomalous events observed in both our cases, one could incorporate artificial events into the structure of a study. This would allow researchers to impose a small level of influence over the *in-situ* environment, potentially without corrupting this environment’s natural state. This practice could have ethical concerns however, dependent on the sensitivity of the information being conveyed by the ambient display. For instance, we would want to consider very carefully the ramifications of implying to an instructor that his or her students are giving negative feedback, when they are not.

Such a method would have to be considered on a case-by-case basis.

This is only one example of what might be done, however, it shows that there are potentially great opportunities to creatively develop new methods for specific ambient devices and contexts of use.

The design approach

An approach that is somewhat different would be to augment the way ambient technologies are designed so that the technology and devices themselves include functions and features that make evaluation a *built-in* property. To some extent this is what the simple logging of visits to the informational web site provided in our first case, and the tracking of behavior changes in the second case.

Another possible example of this approach is presented in [22], where the authors incorporated gaze detection into their evaluation of an ambient display. This allowed the authors to—at the very least—know if their display ever actually entered the field of view of those within the space. Such information could be correlated with other measurements to provide a better picture of how the display is functioning. Other things that would aid in these evaluations would include: absolute location tracking (to know how long a person was in the presence of the display), and galvanic skin response (to correlate any physiological responses to changes in the display).

As we learned from our case studies, neither the *user* nor the *use* of such displays is always clear. Hence, it is hard to predict what data will be useful once the analysis begins, or which data could be beneficial to gain deeper insight into other forms of observation (e.g., surveys). With the second case study for example, contrasting the results of an online questionnaire with the automated data logging provided by the pressure mats led to the primary finding (i.e. that the display seems to be having an unconscious effect on behavior). We anticipate that as new technologies for monitoring human behavior and physiological responses become more refined, these *built-in* possibilities will become more available and less intrusive to the user and hence provide more opportunities for automatic capturing of data, leading to richer analysis and interpretation opportunities.

A major advantage of this approach is, of course, that the researchers do not need to directly interact with the subjects in order to gain the extra information. A possible disadvantage is in the possible issues of privacy that could be introduced by this form of monitoring. Such issues may, to some degree, also be handled with careful design of the technology since the purpose of these studies is not necessarily based on the actions of one individual user, allowing us to disregard information on individuals.

The “Darwin” approach

This approach is based on the assumption that generally good technologies prevail over time in a manner analogous

to Darwinian Natural selection. Here, the method would involve building a large number of different ambient displays, situating them in real environments, and letting them function until time reveals which implementations have a propensity for sustained use, and which eventually become discarded.

When using this method however, the concept of *best* must be carefully operationalized when dissecting the features or attributes that make up a particular ambient display. This is because it is not always a specific attribute that determines the success or failure of an implementation. Rather, it is the total collection of attributes that dictate whether or not it will succeed. For instance, many point out that the Betamax video standard provided superior video quality to the more successful VHS format, implying that video quality was not the only attribute that mattered. Initially, the VHS format offered a much larger storage space (3 hours vs. 1 hour for Betamax), and VHS players were much cheaper to produce. Continuing the Darwin analogy, the combination of these factors provided the VHS format a stronger reproductive advantage over Betamax, and hence it was able to survive longer.

Evaluating ambient displays in this fashion would focus more on rigorous *post mortem* analysis rather than monitoring activities, and development of in-depth user studies. It is plausible that analyzing failed or successful designs, perhaps using theoretical frameworks such as Actor-Network Theory, or Distributed Cognition, could expose general characteristics or qualities. These could, in turn, be used to produce new design principles specific to the development of ambient displays.

The Interaction Criticism approach

Due to growing interest in topics such as affective computing, experience design, and intimate and embodied interaction, some are calling for an increased awareness of “the symbolic level of mood and meaning” within the field of Interaction Design [23]. To address this, some suggest that an expressive language of interaction design be explored and developed. A possible solution offered by Bardzell et. al. [3] is to incorporate an *interaction criticism* phase into existing interaction design processes. This phase would resemble the style of critical analysis employed among music, film and literary critics, as well as academic criticism in art and architecture. In these cases, a critic is able to apply interpretive reasoning to some of the more subjective aspects of a given artifact. This ability comes from developing a level of expertise on a given subject (e.g. ambient displays) through lengthy encounters, theories, and in the examination of other works of criticism from peers and colleagues. Such a method would involve the creation of a community of practitioners who routinely evaluate each other’s designs via criticism, and use this criticism to help interpret observations made of these displays in contexts of use.

A justification for this approach is that many of those who are developing ambient displays are either artists themselves, or are borrowing heavily from the fine arts and architecture. For example, the “DataFountain” and “The Source,” were both produced by artists. Hence, it is reasonable that we would also adopt some of the forms of evaluation used within these disciplines.

CONCLUSION

We have argued for the necessity of conducting long-term in-situ studies to better understand how people make use of ambient displays. From the two case studies presented here, we have identified several issues that should be considered when developing similar studies on this kind of technology, and a series of alternative approaches for evaluation that warrant investigation. We believe that the classification of these issues will require further iteration in order to become more specific, and more theoretical work has to be done to refine the approaches. For instance, there is a need to define or redefine already established concepts such as *use*, *user*, *interaction*, *awareness*, *attention*, *presence*, and others in the context of in-situ ambient displays.

We have proposed four potential directions (approaches) that HCI research could develop in order to better handle evaluation of ambient displays. However, we do not foresee any of these approaches to be the only way forward, and we anticipate that all of these approaches will likely be developed in relation to additional theoretical work. Additionally, we believe that the methods proposed in this article warrant further investigation as a possible means of adding richness and depth to existing evaluations, and we would expect future evaluations to incorporate some form of hybridization of these approaches. This work represents only the early steps of a vein of research that will eventually lead to the development of frameworks and methodologies appropriate for understanding how people will live and interact with this particular technology.

ACKNOWLEDGMENTS

We would like to thank all those who aided in the construction of the installations presented within our case studies, as well as the conference reviewers for their insightful comments.

REFERENCES

1. Ambient Devices, "The Ambient Stock Orb." <http://www.ambientdevices.com/cat/orb/orborder.html> (accessed September 20, 2010).
2. Arroyo, E., Bonanni, L., and Selker, T. 2005. Waterbot: exploring feedback and persuasive techniques at the sink. In *Proceedings of CHI '05*. ACM, New York, NY, 631-639.
3. Bardzell, J. and Bardzell, S. 2008. Interaction criticism: a proposal and framework for a new discipline of hci. In *Proceedings of CHI'08*. ACM, New York, NY, 2463-2472.

4. Brewer, J., Williams, A., and Dourish, P. 2005. Nimio: An Ambient Awareness Device, Demonstration at the European Conference on Computer-Supported Cooperative Work (ECSCW). 18--22 September 2005, Paris, France.
5. Carter, S., Mankoff, J. 2004. Challenges for UbiComp evaluation, technical report UCB/CSD-04-1331, computer science division. University of California, Berkeley.
6. Chang, A., Resner, B., Koerner, B., Wang, X., and Ishii, H. 2001. LumiTouch: an emotional communication device. In *Proceedings of CHI '01*. ACM, New York, NY, 313-314
7. Costanza, E., Inverso, S. A., Pavlov, E., Allen, R., and Maes, P. 2006. eye-q: eyeglass peripheral display for subtle intimate notifications. In *Proceedings of MobileHCI '06*, vol. 159. ACM, New York, NY, 211-218
8. Dahley, A., Wisneski, C., and Ishii, H. 1998. Water lamp and pinwheels: ambient projection of digital information into architectural space. In *Proceedings of CHI 98*. ACM, New York, NY, 269-270
9. Greyworld, "The Source (2004)." June 20, 2004. <http://greyworld.org/archives/31> (accessed September 20, 2010).
10. Hallnäs, L. and Redström, J. 2001. Slow Technology – Designing for Reflection. *Personal Ubiquitous Comput.* 5, 3, 201-212.
11. Hazlewood, W. R., Dalton, N., Marshall, P., Rogers, Y., and Hertrich, S. 2010. Bricolage and consultation: addressing new design challenges when building large-scale installations. In *Proceedings of DIS '10*. ACM, New York, NY, 380-389.
12. Heiner, J. M., Hudson, S. E., and Tanaka, K. 1999. The information percolator: ambient information display in a decorative object. In *Proceedings of UIST '99*. ACM, New York, NY, 141-148.
13. Holmquist, L.E.. 2004. Evaluating the comprehension of ambient displays. In *CHI '04 extended abstracts on Human factors in computing systems (CHI '04)*. ACM, New York, NY, USA, 1545-1545.
14. Ishii, H., Wisneski, C., Brave, S., Dahley, A., Gorbet, M., Ullmer, B., and Yarin, P. 1998. ambientROOM: integrating ambient media with architectural space. In *Proceedings of CHI 98*. ACM, New York, NY, 173-174
15. Ishii, H., Ren, S., and Frei, P. 2001. "Pinwheels: visualizing information flow in an architectural space," In *Proceedings of CHI '01 extended abstracts on Human factors in computing systems*, pp. 111-112, 2001.
16. Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., and Ames, M. 2003. Heuristic evaluation of ambient displays. In *Proceedings of CHI '03*. ACM, New York, NY, 169-176.
17. Mensvoort, K. "DATAFOUNTAIN - Money currency rates displayed with an internet enabled water fountain." <http://www.koert.com/work/datafountain/> (accessed Sep 20, 2010).
18. Mynatt, E. D., Back, M., Want, R., Baer, M., and Ellis, J. B. 1998. Designing audio aura. In *Proceedings CHI '98*. ACM Press/Addison-Wesley Publishing Co., New York, NY, 566-573.
19. Olivier, P., Cao, H., Gilroy, S. W. and Jackson, D. G. 2006. Crossmodal Ambient Displays. In *Proceedings of British HCI 2006*, Springer, London, 3-16
20. Redström, J., Skog, T., and Hallnäs, L. 2000. Informative art: using amplified artworks as information displays. In *Proceedings of DARE'00*. ACM, New York, NY, 103-114
21. Rogers, Y., Hazlewood, W. R., Dalton, N., Marshall, P., Pantidi, N. and Hertrich, S. 2010. Ambient Influence: Can Twinkly Lights Lure and Abstract Representations Trigger Behavioral Change? In *Proceedings of Ubicomp '10*, 2010 (in press).
22. Shen, X., Eades, P., Seok-Hee, H., Vand Moere, A. 2007. Intrusive and Non-intrusive Evaluation of Ambient Displays. In *Proc. of 1st Workshop on Ambient Information Systems*. Colocated with Pervasive 2007, Toronto, Canada. vol. 254, May 2007.
23. Smith, G. What is interaction design? In Moggridge, B. *Designing Interactions*. The MIT Press, 2007.
24. Weiser M, Seely Brown J. Designing calm technology. PowerGrid Journal 1.1.1996. Available at: <http://www.powergrid.com/1.01/calmtech.html>
25. White, T. and Small, D. 1998. An interactive poetic garden. In *Proceedings of CHI'98*. ACM, New York, NY, 335-336.
26. Wurman, R. S. 2001. Information Anxiety. Doubleday, New York.