User-Centered Evaluations of Ubicomp Applications

Jean Scholtz, Larry Arnstein, Miryung Kim, Tim Kindberg, Sunny Consolvo

IRS-TR-02-006

May 2002

User-centered Evaluations of Ubicomp Applications

Jean Scholtz[1], Larry Arnstein[2], Miryung Kim[2], Tim Kindberg[3], & Sunny Consolvo[4]

[1]National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899
jean.scholtz@nist.gov
[2]Department of Computer Science & Engineering, University of Washington, Box 352350
Seattle, WA 98195-2350
{larrya, miryung}@cs.washington.edu
[3]Hewlett-Packard Laboratories, 1501 Page Mill Road, MS 1138, Palo Alto, CA, 94304-1126
timothy@hpl.hp.com
[4]Intel Research Seattle, 1100 NE 45th St, 6th Floor; Seattle, WA 98105
sunny@intel-research.net

Abstract. As more evaluations of ubiquitous computing (ubicomp) applications are being undertaken, researchers are finding that the evaluations are more difficult than those for typical desktop computing applications. In this paper, we discuss some aspects of ubicomp that make these evaluations challenging. We present some properties unique to ubiquitous computing applications and suggest metrics to evaluate them. Several case studies of ubicomp evaluations are discussed with respect to these properties.

1 Introduction

Researchers in ubicomp applications are attempting to produce systems that will fill a user need. To understand if this has been achieved, it is necessary to conduct user-centered evaluations. Because the deployment of this technology will certainly impact how users work and play, user-centered evaluations have become a moving target [1]. In this paper we discuss how evaluating ubicomp applications presents research challenges above those associated with traditional desktop-based and mobile computing applications. In particular, evaluating ubicomp requires understanding and measuring how people perform tasks or engage in activities that may only partially involve computing. We discuss lessons learned about evaluation from several case studies. We also outline some research directions that could facilitate future evaluation of ubicomp applications.

2 Ubicomp Application Properties and Implications for Evaluation

We start by proposing some properties that are characteristic of ubiquitous computing applications in general. Kindberg and Fox [8] raised the issues of spontaneous interoperability and integration with the physical world. Based on those, we elaborate the aspects that have implications for evaluation. These aspects are not exhaustive, nor are they completely separate. Identifying characteristic dimensions of ubicomp systems will help provide insights into the types of evaluations that could be conducted, as well as help identify appropriate evaluation techniques and mechanisms. The characteristics entail new types of phenomena to be evaluated; they also raise logistical difficulties for evaluators.

Physical integration: Kindberg and Fox use the term "physical integration" to describe the use of sensors and actuators to link a computing system with physical entities. Those entities include everyday objects not normally associated with electronic functionality, such as supermarket goods, coffee cups or pieces of paper. Sensors in the environment and in handheld devices enable information and services to be associated with users, places, devices and non-electronic objects. Actuators may be used to provide enhanced behaviors to what appear to be everyday objects; for example, at Xerox PARC, the height of the water in a fountain was used to represent the performance of the company's stock.

Physical integration encompasses the use of ubicomp technologies in real-world physical activities such as laboratory experiments or visiting museums. It also addresses interaction design, including the design of the different devices and modalities provided for access to the digital aspects of the system.

Spontaneous interoperability: Traditional desktop systems work as initially configured until new hardware/software is added. Ubicomp systems are continually under change – new users, devices and software enter and leave the system continually; the system cannot be shut down and reconfigured to support a new object entering or leaving the system. Moreover, users expect the system to become available once they enter the room, building or other ubicomp environment.

A spontaneously interoperable system is "calm" in Weiser's [12] sense: it adapts to changing circumstances without distracting the user from her activities. Evaluators can ask about the degree of system transparency for the user as she moves around or as devices re-associate during other types of change.

We now elaborate aspects of physical integration and spontaneous interoperability, which have implications for evaluation.

- Interleaving – One aspect of physical integration is that ubicomp applications are designed to seamlessly interleave with user activities that principally concern artifacts other than computers. For ubicomp applications to be useful, they must not disrupt these other activities. Characteristics such as size, shape, weight and device robustness affect the usability and acceptability of the system. Evaluators need to be concerned with the amount of cognitive resources the user must devote to the ubicomp application. Measures should include the amount of disruption imposed on the user when interacting with the ubicomp application and changes, both positive and negative, in the user's activity. Evaluation efforts looking at interleaving should identify switches between the physical and virtual worlds, including any difficulties involved and the time spent in each world.

- Interaction Design: Ubicomp applications are often not accessed using traditional mouse/keyboard/displays. Because of the interleaving with current activities, speech, gestures and physical manipulations may be more appropriate. Metrics should include: universality of the interaction for the expected user population; how natural the interaction is; how robust interaction recognizers must be to be usable; degree of support for novice and expert users; understandability of the interaction method, including factors such as vocabulary and new electronic affordances of physical objects; and the amount of training needed. In addition, there is a whole range of issues concerning error management and error handling appropriate to the interactions and environment. These factors will certainly affect the usefulness of the applications and must be carefully evaluated. Ubicomp evaluations of interaction design are perhaps the most similar to traditional HCI evaluations, with the exception that novel interactions will occur, necessitating evaluators to find ways to capture those interactions. Evaluations will often take place in collaborative environments, so it becomes necessary to capture and synchronize multiple interactions while attributing them to specific individuals.

- Inferences of context and activity: Some ubicomp systems use the information provided by sensors and actuators to make inferences about what a user is doing, where a user is, and what actions would be helpful to the user. Evaluators need to measure the accuracy of the inferences and the appropriateness of the resultant actions taken by the system. They must also assess how easy it is for the user to undo any unwanted actions.

- Boundaries of Responsibility: It may be difficult for the user to determine what the ubicomp application can do and the division of responsibility between the user and the application. This is particularly true as users move between various ubicomp environments. For example, a user may wonder whether the "smart room" will automatically locate her slides based on the agenda, or whether she has to physically bring her slides with her. Evaluation of boundary confusions should collect instances where something doesn't happen because the user assumed the system would do it, the user and the system both attempt the same task or the user attempts a task that the system could have performed.

- Security and Privacy: Issues of privacy and trust occur when users interact with wireless ubicomp environments, particularly when the environments are unfamiliar to them. Evaluators need to determine if the user understands what records, if any, remain behind after she leaves the environment. If information is saved, the user needs to understand who will have access to that information and how it will be used. Ubicomp researchers need to explore different means of conveying security and privacy information to users, what level

of data transmissions need to be explicitly approved by users, and what levels are implicitly agreed upon by consent to use a given ubicomp application. Evaluations of security and privacy should consist of quantitative data and subjective ratings by system users. Quantitative ratings could identify incidents where users refused to allow or attempted to prohibit the system from obtaining data. Subjective ratings could ask users how comfortable they feel using the system, how and where they believe system data is being used and how certain they are of their answers.

3 Previous Work in Evaluation of Ubicomp Applications

In this section, we present examples of attempts to evaluate one or more of the ubicomp application properties described above. Five of the six applications below were evaluated in the actual setting for which they were designed.

Classroom 2000

The goal of Georgia Tech's Classroom 2000 [2] [3] was to allow students to adopt a note-taking and study strategy most appropriate for their learning style, rather than one that is constrained by the lecture technology. This is accomplished by providing an instrumented classroom that captures and organizes a live lecture in a way that can be easily accessed later. Classroom 2000 was evaluated in two stages: a small-scale formative evaluation was performed on an early prototype in one class with no control group, followed by a formal summative evaluation on a more robust system that involved 60 classes in several institutions.

The formative evaluation focused on the interaction design of the system, rather than on specific educational benefits. The initial prototype consisted of two components: an active whiteboard platform for capturing the instructor's activities and an electronic notebook for capturing student notes. Observation and survey techniques were used to assess how these two platforms were used and perceived by the students. An important outcome of the observation was that the student platforms did not seem to offer significant added value. Furthermore, the surveys indicated that the students found the physical interface to be cumbersome. The observation and survey results served to reinforce a difficult conclusion: that the student platform should be dropped. This is an example of why it is important to apply more than one user study technique, as either technique alone might not have been convincing enough to support such a radical design change.

The summative evaluation performed on Classroom 2000 is the most extensive one that has been reported in the ubicomp literature. The evaluation results are based on 18 months of use of various prototypes in 60 undergraduate and graduate courses at a variety of universities in Atlanta. The evaluation used control groups. The evaluators gathered quantitative usage statistics, obtained qualitative data through interviews and surveys and assessed the impact on learning through comparative studies. This data provided a comprehensive picture of how instrumentation of the classroom altered the live lecture experience, as well as how it altered student behavior outside of class.

Tivoli

Tivoli is a meeting capture and salvage system [9] [10]. Tivoli was evaluated in 60 authentic meetings about intellectual property management at Xerox PARC over a two year period. Data was collected in a variety of ways: a) the meetings were recorded on video, b) artifacts of the meetings were kept, c) the users were interviewed, and d) logs were kept of the users' interaction with the system.

The evaluation of Tivoli raised an interesting trade-off between the utility of the system and the security and privacy that it offered. The audio capture feature of Tivoli helped meetings flow more smoothly because members knew that they could later retrieve important points, but it also had drawbacks. In the beginning of the project, for privacy purposes, the members chose to allow recording of only meeting recaps. But they soon found that it was easier to record the entire meeting except during particularly sensitive discussions. User acceptance of constant audio recording depended on trust that the recorded material would always remain confidential to the members of the meeting. Unfortunately, it was not practical to omit all offensive remarks, making it more difficult for the members to apply their human capacity to "forgive and forget."

Sotto Voce

Sotto Voce [13] [14] is an electronic guidebook designed to improve the experience for visitors to Filoli, a historic house in Woodside, California. The evaluation focused on the trade-offs between the richness of the educational experience and the impact on social interactions by assessing how the guidebook facilitated social interactions. Sotto Voce was tested with 14 visitors in several rooms of the house. Each visit consisted of three phases: a partial tour using a paper guidebook, a partial tour using an electronic guidebook, and an interview. Data was collected in a variety of ways: a) conversations were recorded and transcribed; b) the visitors were videotaped and directly observed by the research escort; c) the interactions with guidebooks were logged; e) interviews were conducted; and f) feedback from visitors was collected by email.

The evaluators applied conversation analytic methods to identify patterns of behavior. As a result of this analysis, it was clear the PDA-based guidebook provided a point of reference for social interactions between visitors to the historical house, which was inhibited only when the audio was delivered through a headset. The evaluators found that the guidebook enhanced the experience by allowing the visitors to process the visual aspects of the exhibit while listening to a running commentary. Though not quantitative, the transcribed, coded dialogs provided compelling evidence of how use of the guidebook was interleaved with social interactions.

Exploratorium

HP Labs' Cooltown group is deploying and evaluating ubicomp applications at the Exploratorium in San Francisco [7], an interactive science museum. The project has experimented with applications such as “Informer,” which informs the user about the exhibit they are visiting via a PDA, in the form of Web pages obtained when the user picks up a URL from an infrared beacon. “Rememberer” enables users to “bookmark” an exhibit as a Web page containing links to information and photographs taken in real time, which they can view after the visit. To register exhibits, users carry an RFID tag which they “swipe” at readers, or a PDA to collect URLs from beacons.

The project evaluated the applications by observation, interviews, follow-up email questionnaires, and logs of Web accesses. Control groups were also observed and interviewed. The Exploratorium raised logistical issues such as the need for robustness in a changing, crowded environment and problems of scale. However, the evaluation verified that interleaving was a significant issue, and that a prototype of Informer interfered with exhibit manipulation and social interactions. By contrast, Rememberer enabled users to concentrate largely on the exhibits and companions, postponing virtual interactions to when the user wanted to reflect on or interact with others about the visit. However, the evaluation of Informer was not able to isolate the “wow” factor: users found the technology so “cool” that it artificially competed with the exhibits for their attention, making it hard to verify its value for the intended purpose. Moreover, some
interleaving effects were difficult to identify, such as when users became confused about which exhibit's Web pages they were seeing.

The applications also involved new interaction models and unfamiliar boundaries of responsibility: picking up Web pages from an infrared beacon, and causing a camera near the exhibit to take photographs (instead of taking pictures with a hand-held camera). These have unusual parameters such as beacon placement and the timing and framing of photographs, which raise logistical issues since they cannot be systematically evaluated except at great cost.

Labscape

Labscape [5] is a smart environment designed to make cell biology laboratory work easier, while enhancing the ability of biologists to communicate and collaborate with each other. The goal is to simplify laboratory work by making information available where and when it is needed, collecting and organizing data where and when it is created, and organizing the data into a formal representation of the laboratory work that colleagues can understand. Similar to the Exploratorium project, Labscape’s evaluation [6] focused on how well the system was integrated into the physical and intellectual activity of the biology laboratory environment.

In order to understand how the system changed work practices, metrics were chosen that could be applied with and without the system in place; this allowed a baseline of the environment to be established. For example, from the data, one can determine how frequently a biologist accessed information in the same one minute period in which they also manipulated a native tool or device. If the frequency of this form of interleaving increased after Labscape was

deployed, then one may conclude that the information is more conveniently located. However, such a conclusion would have to be supported by interviews and field observations.
The tasks that were analyzed in Labscape's evaluation were authentic experiments that spanned several hours of work in a laboratory, which would have occurred independent of the evaluation. Though quite similar, the experiments that were observed varied enough to prevent direct comparison through statistical means. Thus, the quantitative results are used primarily in a descriptive rather than statistical sense. Artificially constraining the tasks to facilitate statistical analysis would have been impractical due to the time commitments of the participants. Moreover, such constraints would have caused the biologists to significantly alter their behavior, as the outcome and details of the experiments would have been of no interest to them.

Rasa

Rasa [11] is a system designed to support situation assessment in military command posts. The goal of the system is to maintain a digital representation of a physical deployment and planning map that is produced by a team of military personnel. Rasa tracks the use of existing physical tools such as post-its, maps, and plastic map overlays, and it interprets multi-modal input from users, including speech, writing, pointing and drawing. The evaluation was conducted in a simulated military setting at the human interface laboratory on the Oregon Graduate Institute campus. Six male subjects from the Oregon Army National Guard used Rasa and traditional physical tools alone to track an artificial ongoing military situation. In an attempt to prove that the system maintains the proper boundary of responsibility under adverse conditions, Rasa was completely disabled midway through the situation with no advanced warning to the participants, leaving them with only traditional physical tools. When the system was re-enabled, the participants were asked to reconcile the state of the virtual and physical representations of the situation. Evaluation in the face of partial or complete failure is an important element for ubicomp applications.
Another goal of Rasa's evaluation was to understand how users dealt with errors due to the use of multi-modal input methods and recognizers. Errors were classified into a) Recognition errors – the system misunderstanding written or speech commands, b) Performance errors – commands that the system could perform, but which were issued incorrectly, c) System errors – software or hardware design flaws, and d) Guaranteed errors – commands issued by the users that the system was incapable of performing. A sequence of failed error recovery attempts was referred to as a *spiral*. Assessing type, number and depth of such spirals is a potentially useful metric for evaluating ubicomp environments that rely on inference about context and activities.

4 Looking Forward

Ubicomp applications involve much more than "computing." Because they are situated in physical environments and change the way people interact with those environments, evaluation methodologies need to reach beyond traditional HCI techniques. Our case studies exemplify the ubicomp properties we identified as requiring a fresh set of evaluation metrics; they also show some of the difficulties that they raise in practice. Ubicomp applications are inherently interdisciplinary and need interdisciplinary evaluations involving cognitive psychologists, social scientists, organizational experts, industrial designers, human factors experts and experts in particular application domains, as well as HCI experts. The logistical aspects of evaluating ubicomp applications in realistic use scenarios need to be simplified to make evaluations less costly and more timely. There is a need to develop evaluation techniques, including data capture and analysis methods, to address ubicomp properties earlier in the design cycle. The ubicomp community needs to determine what existing techniques from HCI and the social sciences in the literature pertain to ubicomp evaluations, and to augment those methodologies as necessary for ubicomp.

References

[1] Abowd, G.D, Mynatt, E, D., and Rodden, T. (2002) The Human Experience, Pervasive Computing, Vol. 1(1), 48-57.
[2] Abowd, G.D, Atkeson, C.G., Brotherton, J.A., Enqvist, T., Gulley, P., and Lemon, J. (1998). Investigating the capture, integration and access problem of ubiquitous computing in an educational setting. CHI 1998. 440-447.

[3] Abowd, G.D. (1999) Classroom 2000: An experiment with the instrumentation of living educational environment. IBM Systems Journal Vol 38. No 4.
[4] Aoki, P.M., Grinter, R.E., Hurst, A.,Szymanski, M. H., Thornton, J. D. and A. Woodruff, A. (2002). *Sotto Voce:* Exploring the Interplay of Conversation and Mobile Audio Spaces. CHI 2002. 431-438.
[5] Arnstein, L. F., and Borriello, G. (2002) Labscape: The Design of a Smart Environment, Intel Research Seattle Technical Report IRS-TR-02-008.
[6] Consolvo, S., Arnstein, L., Franza, B.R., (2002). User Study Techniques in the Design and Evaluation of a Ubicomp Environment. Intel Research Seattle Technical Report IRS-TR-02-012.
[7] Fleck, M., Frid, M., Kindberg, T., O'Brian-Strain, E., Rajani. R., and Spasojevic, M. From Informing to Remembering: Ubiquitous Systems in Interactive Museums. IEEE Pervasive Computing, Vol. 1 (2). Apr-Jun 2002, pp. 11-19.
[8] Kindberg, T. and Fox, A. (2002). System software for Ubiquitous Computing. IEEE Pervasive Computing, Vol. 1 (1), 70-81.
[9] Moran, T., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., van Melle, W., Zellweger, P. (1997). "I'll get that off the audio": A case study of Salvaging Multimedia Meeting Records. CHI 1997. 202-209.
[10] Moran, T.P, Chiu, P., Harrison, S.R., Kurtenbach, G., Minneman, S.L, and van Melle, W. (1996). Evolutionary Engagement in an Ongoing Collaborative Work Process: A Case Study. Computer Supported Cooperative Work 96. 150-159.
[11] McGee, D.R., Cohen, P.R., Wesson, R. M., and Horman, S. (2002). Comparing Paper and Tangible, Multimodal Tools. CHI 2002. 407-414.
[12] Weiser, M. and Brown, J.S. (1996). "Designing Calm Technology", http://nano.xerox.com/hypertext/weiser/acmfuture2endnote.htm
[13] Woodruff, A., Aoki, P.M., Szymanski M.H., and Hurst, A. (2001). Electronic Guidebooks and Visitor Attention: Proc. International Cultural Heritage Informatics Meeting 2001.
[14] Woodruff, A. Szymanski M.H., and Hurst A. (2001). The Conversational Role of Electronic Guidebooks: Proc. International Conference on Ubiquitous Computing, Atlanta, GA. Sep 2001. 187-208.