Evaluating Ubiquitous Systems with Users (Workshop Summary)

Christian Kray[1], Lars Bo Larsen[2], Patrick Olivier[1], Margit Biemans[3], Arthur van Bunningen[4], Mirko Fetter[5], Tim Jay[6], Vassilis-Javed Khan[7], Gerhard Leitner[8], Ingrid Mulder[3], Jörg Müller[9], Thomas Plötz[10], and Irene Lopez de Vallejo[11]

[1] Newcastle University,Newcastle upon Tyne, United Kingdom
{c.kray}@ncl.ac.uk
[2] Aalborg Universitet, Aalborg, Denmark
[3] Telematica Instituut, Enschede, The Netherlands
[4] University of Twente, Enschede, The Netherlands
[5] Bauhaus-University Weimar, Weimar, Germany
[6] University of Bath, Bath, United Kingdom
[7] Eindhoven University of Technology, Eindhoven, The Netherlands
[8] Alpen-Adria University of Klagenfurt, Klagenfurt, Austria
[9] University of Münster, Münster, Germany
[10] Technische Universität Dortmund, Dortmund, Germany
[11] University College London, London, United Kingdom

Abstract. Evaluating ubiquitous systems with users can be a challenge, and the goal of this workshop was to take stock of current issues and novel approaches to address this challenge. In this paper, we report on the discussions we had during several plenary and small-group sessions. We first briefly review those evaluation methods that we identified as being used in ubiquitous computing, and then discuss several issues and research questions that emerged during the discussion. These issues include: data sources used for evaluation, comparing ubiquitous systems, interdisciplinary evaluation, multi-method evaluation, factoring in context and disengaged users.

1 Introduction

A significant number of ubiquitous systems have been built to support human users in performing a variety of tasks. These applications cover a large range of scenarios, including safety-critical medical applications, customized support in the workplace, and leisure-related applications. Users interact with these systems via implicit or explicit means, e. g. by moving about in an environment or using custom-built devices. At the same time, many ambient applications deploy sensors to gather information about the context, in which those interactions take place, and system behavior can change depending on contextual factors. All this adds up to a fairly complex situation, which poses new challenges for evaluation, potentially pushing the boundaries of traditional evaluation methods and opening up opportunities for novel approaches.

The goal of this workshop was to bring together researchers with an interest in this area to discuss the current state of the art in evaluating ubiquitous systems with real users, to identify shortcomings and benefits of traditional evaluation methods and to explore novel approaches. In order to facilitate the exchange of ideas, all but the first session were dedicated either to working in small groups or to plenary discussion. This paper tries to summarize the main issues that were identified during the discussions, and to highlight some areas, where further research is needed.

In the remainder of this paper we will first briefly summarize existing approaches to evaluation (in section 2) before discussing several issues and research questions that arise from evaluating ubiquitous computing (in section 3). While we engaged with a number of issues during the workshop, we would not want to claim to fully cover every possible aspect relating to the evaluation of ubiquitous systems. There are, however, several related events that have explored this area from different angles, which we briefly describe in section 4. In the final section of this paper, we summarize the key outcomes of the workshop and briefly discuss a number of research challenges, which will need to be addressed in the future.

2 Current Approaches to Evaluation

In principle, almost any evaluation technique used in human computer interaction (HCI) can also be applied to evaluate ubiquitous systems. However, a significant portion of ubiquitous technology is meant to weave itself invisibly into the users' life [1], so that both the users' task and their interactions with a system are defined less clearly than in traditional settings. For this reason, it is not always straightforward to apply the techniques, which are widely used in HCI (e. g. methods based on task performance), for ubiquitous systems [2]. The same can be said for some commonly used metrics in HCI, such as task completion time and error rate. Consequently, different measures have been proposed for ubiquitous applications, e. g. Scholz's and Consolvo's set of conceptual measures and associated metrics [3].

Preece et al. [4] identified four main paradigms for evaluating systems with users. *Quick and Dirty* evaluations have the benefit of delivering results fast and cheap and are used by many projects during the requirements analysis phase. *Usability tests* usually take place in a laboratory, where user performance scores such as task completion time or error rates can be easily measured. *Field studies* aim at gathering data in a natural setting. There has been an intense debate about the value of field studies compared to lab-based studies. Some authors consider it necessary to gather data on how ubiquitous systems are used in the real world [2] [5]. Others believe that in many cases the associated effort is not justified by the additional insights gained [6]. Field studies can considerably vary in terms of their duration. While some studies take place in a single day or week, some projects go as far as letting users live with the technology for months or even years to gain insights on the long term effects of technology [7] [8] [9]. *Predictive techniques* use experts, heuristics and user models to evaluate systems without incorporating users.

Evaluation techniques can also be classified according to a number of dimensions. *Formative evaluations* are employed during design iterations to inform design. *Summative evaluations* are used after the design phase has finished to compare the system to other systems or a set of predefined goals. *Introspective techniques* ask for what users think or believe, while *observation techniques* look at the actual behavior of users. *Qualitative techniques* gather data to describe behavior, establish usage scenarios or build categories, while *quantitative techniques* gather data for statistical data analysis. *Short term studies* look at the immediate effects a system has on its users, while *long term studies* aim at identifying effects that only occur after months or even years of usage. In applying any of these techniques, the degree of sophistication or fidelity of the system can vary widely. Sometimes, only user behavior without any prototype is evaluated. Most of the time, an instantiation of a ubiquitous system is part of the evaluation, and it can take the form of a paper based prototype, interface mockups, or (partially) functional prototypes.

A further way to categorize evaluation techniques is according to the way in which users are involved, i. e. whether they are being observed, whether users/experts are being asked directly, whether they are brought into a usability lab, or being modeled using a user model. Observing users can happen directly, with the experimenter directly witnessing the fact, or indirectly, where the experimenter can merely analyze artifacts that were created during the experiment. Direct observation often employs techniques from ethnography. Information can be kept using a notebook and a still camera, using audio recording and a still camera, or using video, for example [4]. Indirect observations can use (photo) diaries, guest books [10], interaction logs [5] or logs from diverse sensors [11]. Cultural probes [12] are small artifacts like still cameras or modeling clay that are given to users. Within a certain time period, users can use these artifact to capture their experiences. Systems that intent to change user behavior can be evaluated by measuring user behavior before and after they used the system. Similarly, the change of the environment around the system can be observed.

Common ways to directly gather user feedback are *questionnaires* and *interviews* [4] [13]. Interviews can be structured (i. e. they follow a rigid predefined procedure), or unstructured. Semi-structured interviews often start with pre-planned questions but then probe the interviewee for more information. *Focus groups* [14] are widely used to let users from different user groups react to each other. *Laddering* [8] is a special interviewing technique to establish users values regarding a system. *Online questionnaires* hold the potential to reach a large number of users, but it is harder to control the sample. *Experience sampling* [15] is a widely used technique to ask the user simple questions many times distributed over a certain time period. The *Day Reconstruction Method* [16] combines features of time-budget measurement and experience sampling. It is used to assess how people spend their time and how they experience the various activities and settings of their lives. *Conjoint analysis* [17] asks users to rank a number of paper based prototypes, where system features are systematically varied. Using this technique, the relative values users attribute to system features can be established. The *repertory grid* technique [18] aims at eliciting so-called personal constructs (e. g. bad-good, playful-expert-like). It can be used to identify a users perceived dimensions regarding a system. With *participatory design,* users are directly involved in the design phase, such that design and evaluation become closely entangled.

Some techniques aim at observing and asking the user at the same time. In a *contextual inquiry* [8], the interviewer takes on the role of an apprentice and the interviewee shows and explains important tasks. With the *think-aloud* technique, the user is asked to state what he currently is thinking while completing a task. Asking experts is also a widely employed evaluation technique. Heuristics [19] guide the expert along defined constructs to evaluate a system. In a usability lab, certain variables of a system can be measured, while context variables can be held constant. Usually, users are asked to solve a clearly defined task, and task completion time and error rate are measured. The real context of system usage can be reconstructed to a certain degree, and context variables can be held constant [20] [21]. If *user models* [22] are used, users are often modeled with respect to completion of a certain task. This can be done, for example, using GOMS [23] or ACT-R [24].

3 Issues and research questions

In the previous section we briefly listed a number of evaluation techniques that have been used in ubiquitous computing. While there certainly are a large number of options available to evaluate a ubiquitous system, it is not necessarily clear which technique is best suited for a particular system or context of use. In addition, it is not obvious whether a technique can be applied straight away or whether it needs to be adapted to accommodate the specific properties of ubiquitous computing (such as context-dependency or potentially invisible interfaces). Ideally, a framework or a set of guidelines would provide help in selecting the most appropriate evaluation methods based on specific properties of a system and the aims and objectives of the evaluation. Working towards this goal, we discussed a number of issues and questions at the workshop (see Figure 1 for an overview), which we report on in the remainder of this section.

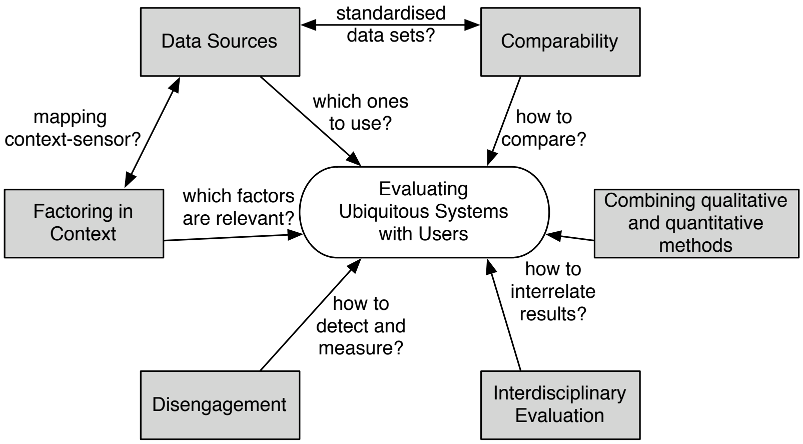
Data Sources
standardised data sets?
Comparability
mapping context-sensor?
which ones to use?
how to compare?
Factoring in Context
which factors are relevant?
Evaluating Ubiquitous Systems with Users
Combining qualitative and quantitative methods
how to interrelate results?
how to detect and measure?
Disengagement
Interdisciplinary Evaluation

Fig. 1. Overview over issues identified during the workshop

3.1 Data sources for evaluation

A key ingredient for realizing ambient intelligence is the acquisition, processing and analysis of data from multiple sensor sources at a large scale. The motivation for this approach is to enable a system to become aware of its context by constantly integrating different sources of information. Compared to more traditional settings (e. g. a desktop use scenario), where oftentimes the users' tasks are very clearly defined, ubiquitous applications can be more difficult to evaluate. This is due to tasks possibly being less well defined, to the tight integration of a system with its environment, and to the complexity of the context in which it is used. For example, whereas a speech recognition system "only" needs to perform the transcription of the recorded utterances, a ubiquitous system may have to analyze the context of a particular situation in order to derive an appropriate system behavior. In the first case, a system can be evaluated by counting correctly classified words, whereas in the second case, such a measure would be insufficient. To illustrate this, consider the example of a smart conference room analyzing context using multiple sensor sources. Using a combination of both microphones and further sensors (such as active tags) the system might infer that currently a very important business meeting is being held. A reasonable system behavior might include that no disturbances occur (e. g. all telephones are muted), all attendees are kept up-to-date with relevant information only (e. g. delivery of urgent emails only) and the creation of a convenient ambience (perfect lighting conditions, constant temperature and humidity).

The more integrated a system is into an environment and the more complex its behavior can be, the more difficult it is to evaluate the system and, thus, the more data sources may have to be taken into account during the evaluation. Although all components of a ubiquitous system can (and have to) be evaluated separately[12] this usually does not give detailed insights into the effectiveness of the overall system. Oftentimes, there is no accepted quality measure which can be evaluated numerically in terms of e. g. recognition rates. In the above mentioned example (the smart meeting room), a good indicator for the effectiveness of the system would be the number of successful deals made *due to* the corresponding meetings being held in the smart conference room.

Obviously, such an effect is rather difficult to measure. Nevertheless, it seems advisable to gather as much data as possible in order to have as much information available for analysis as possible. The benefit of such an approach is that once collected the data can be analyzed in many ways. Furthermore, if data sets are made publicly available, it opens up the opportunity to run different algorithms on them and to compare their performance (see also section 3.2). In addition to the data captured by actual sensors, system logs are another good source of information. They contain a wealth of information regarding the overall behavior of the system and may also help to put the sensor data into perspective (e. g. to detect system errors that might have lead to erratic system behavior).

When combining these primary and secondary sources of information (i. e. sensor data and system logs) with feedback obtained from users (e. g. using one of the methods discussed in the previous section), it is obvious that a potentially very large body of

[12] Examples are the recognition rate of a speech recognizer built into the system, or the robustness of video-based person identification

data has to be analyzed. While this can be beneficial in a number of ways (e. g. by increasing the coverage, or by enabling cross-checking of different sources), it can also entail significant problems. The latter include dealing with contradicting data, the effort required to analyze a very large body of heterogeneous data and the lack of tools to facilitate the task.

Generally speaking, the evaluation of ubiquitous systems oftentimes requires the analysis of multiple data sources. Individual results need to be integrated to include both technical aspects (e. g. in terms of multi-modal sensor fusion) and socio-psychological factors. The key goal of this integration process is to fully evaluate the performance of a ubiquitous system with respect to both overall user satisfaction and the 'technical' performance of the system.

While the large amount of data being gathered can be beneficial in terms of getting a clearer picture of the context and regarding the detailed evaluation of a system, there are also some drawbacks and further implications. The data being gathered may potentially include sensitive data with respect to the security and privacy of people or places. It is hence the responsibility of the researchers collecting the information to ensure that the data is handled appropriately. Furthermore, there is the danger that the volume of data is so larger that extracting meaningful information can become very time-consuming.

3.2 Comparability

The comparability of evaluations is very important, yet problematic. The purpose of evaluations is to inform the development of future applications and systems. Later systems should be better than those they replace. That term 'better' causes problems because it implies measurability. The evaluation of two systems should enable us to say that one system is better than another, according to some set of goals.

At present, there are a number of obstacles to this. One significant problem, as can be seen from preceding sections of this paper, is the number of methods available to researchers working on ubiquitous computing. While within this paper we advocate the use of multiple methods in evaluation, it is clear that this can make it difficult to compare results of the evaluations of systems.

A second issue concerns the clarity of goals of given systems. Having the ability to compare a system or the outcome of an evaluation requires the definition of user goals that are to be reached or exceeded. However, goal definition in ubiquitous computing is inherently problematic for many ubiquitous applications due the the complications of contextual factors. However, the basic requirements on goal setting can be borrowed from usability engineering literature, for example [25]. The Usability Engineering Lifecycle differentiates between quantitative and qualitative goals. Quantitative goals usually have clearly measurable properties, such as the time to access information. By contrast qualitative goals relate to less tangible requirements such as the facilitation of high quality collaboration between users, which can best evaluated through qualitative methods such as interviews and focus groups.

Often the unambiguous definition of goals is not possible, because of the richness of the environment in which the systems are used. This is particularly true of ubiquitous computing application when the a value proposition cannot be formulated in a meaningful manner. For example, in ambient assisted living applications, the cumulative benefit of the application is the independent living of the occupants, but this is often hard to decompose into specific achievement of goals in a new smart home system. Such a system might typically observing the everyday activities of the occupants e.g. activities relating to dressing, washing and medication taking, prompting and calling for care as appropriate. However, the individual achievement of goals does not add up, in a simple manner, to the achievement of the goal of independent living. Achievement of the broader goal is itself determined by the disruption that such a system is deemed to make to a user's daily routine, the reliability and predictability of the system, and ultimately the user's (and their carers') confidence in it.

The problem of comparing two systems in ubiquitous computing can be significantly more difficult than in other domains. This is partially due the unpredictability of contextual factors and its impact on repeatability, which is a prerequisite for rigorous investigation and the comparison of systems with one another. For example, while it may be quite easy to measure whether buying a book using one website is faster than using another website, but comparing two ubiquitous systems providing adaptive navigation support depending on the context may be more difficult. The main areas for consideration at this time then, is to consider carefully how evaluations are to be compared across systems for the benefit of continued development of ubiquitous computing and to develop a shared conception of the goals that are to be achieved through this development.

3.3 Benefit of interdisciplinary evaluation

Interdisciplinarity is an important issue in the evaluation of ubiquitous systems. An unusually wide variety of skills and knowledge are required for this kind of research. We need an understanding of design, of software and electrical engineering, of what is possible to achieve in ubiquitous computing – along with a thorough understanding of research methods, including a knowledge of how to choose appropriate methods for a given situation. Of course, interdisciplinarity can bring its own complications.

It can be argued that thoroughly understanding evaluating ubiquitous computing in context requires cross-domain research. The increasing complexity that comes with ubiquitous technologies along with the dynamic characteristic of these technologies even furnishes more proof. Bridges across disciplines are necessary. Theoretical insights, methods, best practices, and experiences from several disciplines, such as constructivist theories, behavior sciences, educational science, computer science, electronics, ethnography, discourse, and sociology, should be combined, and should feed back to the current ubiquitous computing research. These bridges are necessary, not only between theory and practice or between social scientists and technological scientists, in order to evaluate ubiquitous systems with users. The traditional disciplinary borders must be crossed to integrate different standards and approaches as well as different evaluation methods, to enable a holistic understanding of the impact and qualities of a ubiquitous system.

3.4 Combining quantitative and qualitative methods

Currently the ubiquitous computing community is mainly driven by the rapid advance of technological solutions. The mechanics of these technological solutions (network connectivity, mobile devices, sensors, programming interfaces etc.) are plentiful, accessible and inexpensive. This enables the community to easily experiment by creating prototypes. However, it is often the case that the evaluation goals of such prototypes are unclear.

There is always an anticipated value proposition of the prototype. However it might not always be possible to subsequently evaluate it in a rigorous, scientific way. We would argue that this is somehow expected since in many cases it is either difficult to deploy a prototype with many users or the purpose of the prototype is exploratory. Such prototypes are still in their early development. That fact makes it challenging for researchers to evaluate their benefits and costs since there is little knowledge about the way users would interact with such systems. Thus, a thorough understanding of the available methods to evaluate such systems is needed.

To better understand the use of ubiquitous systems and to rigorously evaluate their proposed benefits and costs, we argue that a combination of qualitative and quantitative methods is needed. Since these systems are still in development qualitative methods should be first deployed to evaluate the way people interact and fit into their lives such systems. Then, when a clearer idea of the context and the proposed benefit is established, quantitative methods are needed to rigorously evaluate that proposed benefit. In this way their potential can be generalized.

For example, an ethnographic study, diary study or interview can be used to assess the way users interact with a system. Such qualitative methods would allow the user to openly discuss about all the aspects of the interaction. In this way researchers decrease the possibility of having overlooked either a benefit or a cost that their system might bring. Having clearly established what to evaluate in the system, the use of questionnaires or a log of behavioral measures would give the means to the researcher to meticulously demonstrate the effect of the system.

3.5 Factoring in the context

As ubiquitous technologies become more and more personal, they increasingly stay with one person at a time and are consequently used in various contexts. One way to deal with these challenges is the Living Lab concept. Living Labs move research out of laboratories into real-life contexts to stimulate innovation. The Living Lab concept is acknowledged in Europe as an open innovation instrument, and refers to a network of infrastructure and technologies as well as a network of people; it seems appropriate to study questions related to the design and evaluation of ubiquitous technologies that improve and enrich everyday life. The Living Lab approach represents a research methodology for sensing, prototyping, validating and refining complex solutions in multiple and evolving real-life contexts. The user experience focus involves areas of user interface design and ergonomics as well as user acceptance, extending to user co-design process, finally leading to service or product creation.

It might be clear that the Living Lab concept opens a wealth of possibilities to exploit the evaluation of ubiquitous technologies in context with and by real users. However, as indicated by Mulder and Kort [26]:

> there are no agreed upon generic methods for logging yet. Only system events, but these are detailed and not always complete. Often logging is implemented into the ICT product or service during implementation. This implementation is not always straightforward or even possible, when you do not have access to the source code. Many of the automated tools alone do not deliver the desired insight, they need to be combined with common methods such as interviews and focus groups which either provide input for the automated measurements (which things should be captured and asked for during experience sampling) or provide additional information after the automated measurements (clarifications of specific experience sampling data, behaviors or contexts in which it appeared).

Moreover, there is still a need for research in methodological guidelines and tool requirements for data-analysis. In particular, analysis techniques for correlating objective behavior and subjective user experience data into relevant design context parameters.

3.6 Disengagement

The fact that pervasive and ubiquitous systems are often designed for public consumption can add some interesting issues in terms of their evaluation. As well as evaluating the usability of the system, it is necessary to understand how the system engages or fails to engage users.

Taking an example of interactive or intelligent public displays, there will be a subset of users who have engaged with the system to some extent and a subset who have not. It is clearly important, in evaluating such systems, to work towards an understanding of both of these behaviors. With regard to informing the development of future systems, it is arguably more important to understand the reasons for disengagement than it is to understand patterns of use by engaged users.

Those who engage with a system can be assumed to a certain extent to have a certain range of expectations of that system, whereas it is much more difficult to predict the expectations of those who have failed to engage. Similarly, the motivations of the engaged users, with respect to the use of a system, are relatively easy to predict, compared to the motivations of the disengaged.

There are many potential reasons for disengagement with any public system. There may be issues of investment of time and effort, coupled with difficulties to assess benefit. There may be issues of feeling that a system is likely to be too complicated or otherwise for a exclusive group - ’not for me’. There may be issues of embarrassment of self-consciousness. There are likely to be many and varied reasons amongst a population for disengagement - it is vital that these are considered an integral part of evaluation. The main goal of evaluation of such a system must be the future development of more effective, more usable systems. To this end, we need to gain an understanding of the barriers to use amongst a population of users. We need to do this through evaluation with potential users who have all the means necessary to engage with a system but choose not to.

In the case of systems designed for a workplace or an educational institution, there is a 'captive' audience or user-group. Where there is a defined user-group, access to the disengaged is possible through random or stratified sampling of the population. In these cases, the user-group has a certain level of commitment to, and investment in, the evaluation. For public systems, however, there is not necessarily a well-defined user group. Also, the (potential) user group has no commitment to, or investment in, the evaluation of such a system. They may indeed have the opposite motivation - having chosen not to engage with a system, they may want to actively avoid being questioned about it.

There are two main issues, then, to keep in mind. The first is that an understanding of disengagement is a vital part of the evaluation of public systems. The second is that some thought must be given to the development of suitable methods for gaining this understanding.

4 Related Events

This workshop and its results have to be seen in the context of a series of event aiming in the same direction. One of the early events focusing on the evaluation of ubiquitous computing, the *Workshop on Evaluation Methodologies for Ubiquitous Computing* took part during the 1st UbiComp conference in 2001 organized by Jean Scholtz et. al. [27] resulting in a first sketch of a framework for evaluation, outlining four relevant dimensions: universality, utility, usability, and ubiquity. Along these dimensions the workshop participants exemplarily identified new metrics and challenges as well as needed tools and methodologies specific to the evaluation of ubiquitous computing and so delivered a sound starting point for the definition of an overall framework.

Since then, a number of further workshops on this topic were held. In the following we want to briefly describe the two events that took place immediately before this workshop, namely the *In-Situ* workshop [28] and the *1st International Workshop on Ubiquitous Systems Evaluation (USE '07)* [29] both held in September 2007.

The In-Situ workshop focussed on tools and methodologies for evaluating user behavior and user experience particular aiming at pervasive and mobile systems. The presented approaches predominately concentrated on methods and tools for an evaluation in the wild and can be divided into two categories. The first category consists of approaches where classic methodologies and tools like Thinking-Aloud, Interviews etc. were applied - some times in new combinations - to pervasive systems in field experiments. The approaches in the second category all proposed methods in which the ubiquitous computing technology itself was used to evaluate systems, for example by analyzing and capturing the data of the various sensors in a mobile phone.

Only one week later USE '07 also brought together researchers from various fields of the Ubiquitous Computing domain, fostering the idea of creating an overall framework for the user-centered evaluation of ubiquitous systems by identifying specific techniques. During the workshop a set of challenges, needs and requirements were identified that are special to the evaluation of ubiquitous systems, e. g. the limitations regarding the reproducibility of experiments in dynamic environments, privacy issues and the question on how to compare personalized evaluations. In addition, the workshop participants identified the need for published data-sets, that make research results more comparable in the community and for extendible high-reaching benchmarks. The combination of the benefits of in-situ, virtual, and immersive were found useful as well as the evaluation of personal experiences during natural interaction with the system and deriving implicit feedback from that unlike tests that require the user to follow a script and asking for explicit feedback.

All these events advanced the state of the art with the goal to establish an overall framework for user-centered evaluation of Ubiquitous Computing systems by framing relevant dimensions, identifying specific tools and techniques, and formulating requirements and needs. Furthermore, this series of events reflects both the relevance of this topic to the community and the continued effort to tackle the problem by delivering a common set of tool, techniques and methods.

5 Conclusion and Outlook

During the discussion at the workshop, it quickly became clear that evaluating ubiquitous systems with users poses some new challenges while at the same time opening up opportunities for research. Due to the nature of ubiquitous computing – e. g. the impact of contextual factors, the tight integration into everyday life and interfaces that may be invisible – some evaluation methods do not work well or need to be adapted. We have identified several challenges relating to this issue, including the need to incorporate/control the context of use, comparing different systems that serve similar purposes and the question how to cope with disengaged users. While these are important areas that should be tackled in the future, a key problem in evaluating ubiquitous systems with users is the lack of clear guidelines for the selection of evaluation methods tailored to ubiquitous computing.

At the workshop, we discussed an interesting idea related to this problem. Since many ubiquitous systems already capture a lot of sensor data and thus information about the context, a promising way to optimize the efficacy of evaluation would be to automate the selection of particular evaluation methods based on the context. For example, a system could pick different sets of questions depending on the current context, or it could choose a method such as contextual enquiry in one case and a post-hoc interview in another case.

References

1. Weiser, M.: The computer of the 21st century. Scientific American (1991) 94–100
2. Abowd, G.D., Mynatt, E.D.: Charting past, present, and future research in ubiquitous computing. ACM Trans. Comput.-Hum. Interact. **7** (2000) 29–58
3. Scholz, J., Consolvo, S.: Toward a framework for evaluating ubiquitous computing applications. IEEE Pervasive Computing **3** (2004) 82–88
4. Preece, J., Rogers, Y., Sharp, H.: Interaction Design. Wiley (2002)

5. Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W., Kurtz, A., Hall, R., Hursey, J., Toscos, T.: Why its worth the hassle: The value of in-situ studies when designing ubicomp. (2007) 336–353
6. Kjeldskov, J., Skov, M.B., Als, B.S., Hegh, R.T.: Is it worth the hassle? exploring the added value of evaluating the usability of context-aware mobile systems in the field. (2004) 61–73
7. Abowd, G.D., Atkeson, C.G., Bobick, A.F., Essa, I.A., Macintyre, B., Mynatt, E.D., Starner, T.E.: Living laboratories: the future computing environments group at the georgia institute of technology. In: CHI '00: CHI '00 extended abstracts on Human factors in computing systems, New York, NY, USA, ACM (2000) 215–216
8. Mueller, J., Paczkowski, O., Krueger, A.: Situated public news and reminder displays. (2007) 248–265
9. Cheverst, K., Dix, A., Fitton, D., Rouncefield, M., Graham, C.: Exploring awareness related messaging through two situated-display-based systems. Human-Computer Interaction (**22**) 173–220
10. Taylor, N., Cheverst, K., Fitton, D., Race, N., Rouncefield, M., Graham, C.: Probing communities: Study of a village photo display. In: Proc. OzCHI 2007. (2007)
11. Mueller, J., Krueger, A.: Learning topologies of situated public displays by observing implicit user interactions. In: Proceedings of HCI International 2007. (2007)
12. Hutchinson, H., Mackay, W., Westerlund, B., Bederson, B., Druin, A., Plaisant, C., Lafon, B.M., Conversy, S., Evans, H., Hansen, H., Roussel, N., Eiderbck, B., Lindquist, S., Sundblad, Y.: Technology probes: inspiring design for and with families (2003)
13. Breakwell, G., Fyfe-Schaw, C., Hammond, S.: Research Methods in Psychology. Sage Publications Ltd (1995)
14. Marshall, C., Rossman, G.B.: Designing Qualitative Research. Sage Publications, Inc (2006)
15. Consolvo, S., Walker, M.: Using the experience sampling method to evaluate ubicomp applications. IEEE Pervasive Computing **2** (2003) 24–31
16. Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., Stone, A.A.: A survey method for characterizing daily life experience: The day reconstruction method. Science **306** (2004) 1776–1780
17. Gustafsson, A., Herrmann, A., Huber, F.: Conjoint Measurement: Methods and Applications. Springer (2001)
18. Hassenzahl, M., Wessler, R.: Capturing design space from a user perspective: The repertory grid technique revisited. International Journal of Human-Computer Interaction **12** (2001) 441–459
19. Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Lederer, S., Ames, M.: Heuristic evaluation of ambient displays. In: CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM Press (2003) 169–176
20. Singh, P., Ha, H.N., Kuang, Z., Olivier, P., Kray, C., Blythe, P., James, P.: Immersive video as a rapid prototyping and evaluation tool for mobile and ambient applications. In: MobileHCI '06: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services, New York, NY, USA, ACM Press (2006) 264–264
21. Schellenbach, M.: Test methodologies for pedestrian navigation aids in old age. In: CHI'06 Extended Abstracts on Human Factors in Computing Systems. (2006) 1783–1786
22. Jameson, A., Krüger, A.: Special double issue on "user modeling in ubiquitous computing". User Modeling and User-Adapted Interaction (2005)
23. John, B.E., Kieras, D.E.: The GOMS family of user interface analysis techniques: comparison and contrast. ACM Trans. Comput.-Hum. Interact. **3** (1996) 320–351
24. Anderson, J.R., Matessa, M., Lebiere, C.: Act-r: A theory of higher level cognition and its relation to visual attention. Human-Computer Interaction **12** (1997) 439–462
25. Mayhew, D.J.: The Usability Engineering Lifecycle: A Practitioner's Guide to User Interface Design. Morgan Kaufmann Publishers (1999)

26. Mulder, I., Kort, J.: Mixed emotions, mixed methods: the role of emergent technologies to study user experience in context (in press)
27. Scholtz, J., Richter, H.: Report from ubicomp 2001 workshop: evaluation methodologies for ubiquitous computing. SIGCHI Bull.: suppl. interactions **2002** (2002) 9–9
28. Mulder, I., ter Hofte, H., Kort, J., Vermeeren, A.: In situ workshop at mobile hci 2007. http://insitu2007.freeband.nl/ (last accessed on 22 November 2007)
29. Neely, S., Stevenson, G., Terzis, S.: Ubiquitous systems evaluation workshop (use '07). http://www.useworkshop.org (last accessed on 22 November 2007)