INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION, *22*(1 & 2), 103–118
Copyright © 2007, Lawrence Erlbaum Associates, Inc.

Conducting In Situ Evaluations for and With Ubiquitous Computing Technologies

Sunny Consolvo
Intel Research Seattle, WA
Information School, University of Washington

Beverly Harrison
Ian Smith
Mike Y. Chen
Intel Research Seattle, WA

Katherine Everitt
Jon Froehlich
Department of Computer Science & Engineering, University of Washington

James A. Landay
Department of Computer Science & Engineering, University of Washington
Intel Research Seattle, WA

To evaluate ubiquitous computing technologies, which may be embedded in the environment, embedded in objects, worn, or carried by the user throughout everyday life, it is essential to use methods that accommodate the often unpredictable, real-world environments in which the technologies are used. This article discusses how we have adapted and applied traditional methods from psychology and human–computer interaction, such as Wizard of Oz and Experience Sampling, to be more amenable to the in situ evaluations of ubiquitous computing applications, particularly in the early stages of design. The way that ubiquitous computing technologies can facilitate the in situ collection of self-report data is also discussed. Although the focus is on ubiquitous computing applications and tools for their assessment, it is believed that the in situ evaluation tools that are proposed will be generally useful for field trials of other tech-

Thanks to the many researchers, staff, and participants who have contributed to each of our projects, and the continued support of Intel management and the University of Washington. We especially thank Gregory Abowd, Morgan Ames, Ken Anderson, Gaetano Borriello, Micah Brodsky, Eithon Cadag, Cherie Fenner, Ken Fishkin, Batya Friedman, Anthony LaMarca, Lenny Lim, Scott Mainwaring, David McDonald, Fred Potter, Polly Powledge, Christine Riley, Peter Roessler, Brett Shelton, Ken Smith, Jeff Towle, and Miriam Walker.

Correspondence should be sent to Sunny Consolvo, Intel Research Seattle, 1100 NE 45th Street, 6th Floor, Seattle, WA 98105. E-mail: sunny.consolvo@intel.com

nology, applications, or formative studies that are concerned with collecting data in situ.

1. INTRODUCTION

At Intel Research Seattle and the University of Washington, we apply user-centered design principles to the development of ubiquitous computing (ubicomp) technologies. As Moran and Dourish (2001) pointed out, "[ubicomp technologies] move the site and style of interaction beyond the desktop and into the larger real world where we live and act" (p. 88), thus the techniques appropriate for the design and evaluation of traditional computing are often insufficient for ubicomp (Scholtz & Consolvo, 2004). Ubicomp technologies may be embedded in the environment, embedded in objects, worn, or carried by the user throughout everyday life. Interactions may be implicit or even sensed, unlike the explicit and intentional interactions typical of user interfaces in desktop environments. Ubicomp technologies may be used in unpredictable situations, changing contexts, and highly mobile applications. They impact the user of the technology (i.e., the *direct stakeholder*), as well as the other people who are affected by it, but may not directly interact with or know about it (i.e., the *indirect stakeholders*; Friedman & Kahn, 2003). Traditional lab-based evaluation techniques tend to be insufficient as they omit too many of these critical variables. Instead, our approach adapts techniques from fields such as psychology and human–computer interaction (HCI) to conduct in situ evaluations of ubicomp technologies where user actions can be critically situated within more realistic and crucial contexts of usage (Suchman, 1987).

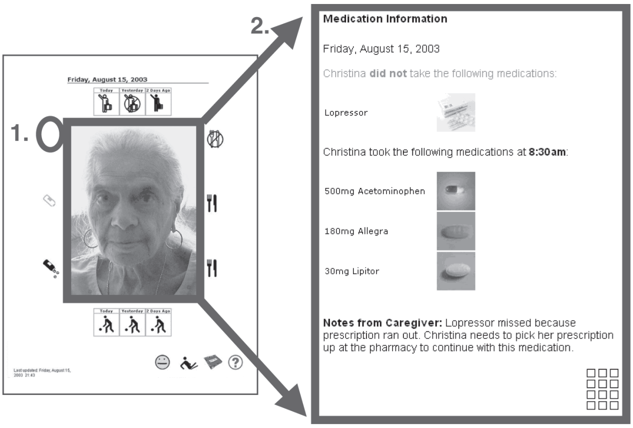
Our evaluation approach typically combines pre- and poststudy interviews with various in situ data collection methods. In in this article we highlight three dramatically different ubicomp technology platforms to illustrate three very different strategies for collecting in situ data. The first study investigated a Wizard of Oz (WOz) prototype to mimic sensors deployed in home settings for supporting eldercare. This study combined participant interviews and questionnaires with daily in situ data collected by an evaluator by phone. The second study investigated a location-aware application using global positioning system mobile phone data to understand how user preferences related to choices in destinations. This study combined interview data with in situ self-reports from participants collected via instant messages triggered based on the participants' arrival at a destination being sensed. The third study investigated daily physical activity habits and whether sharing activity-related data with a small group of friends might influence attaining physical activity goals. This study combined interviews and questionnaires with in situ user-initiated logging of pedometer data that were entered, displayed, and shared using a mobile phone-based application. In each instance, the interview and questionnaire data alone would not have provided an accurate picture of technology use and issues. In addition, the very nature of the applications themselves meant laboratory studies would not be sufficient (i.e., to concept test whether sensors could provide useful data about an elder's well-being while living independently, the elder needed to be in his or her own environment carrying out normal daily routines; to test location tracking and different destinations, participants had to travel to these; to test ongoing physical activities and the effects of sharing daily information, participants needed to be able to update activity-related data as their day unfolded). Thus in each case, the application itself demanded a representative or realistic field setting to make sense, and the evaluation needed to happen concurrent with the in situ use of these applications. We discuss the challenges that this in situ data collection represents for each of these cases.

2. THREE APPROACHES TO IN SITU DATA COLLECTION AND EVALUATION

2.1. In Situ Data Collection Using the WOz Technique

As is typical in traditional application or device design, we often wish to concept test our ubicomp application and system ideas before making significant investments of effort, time, and cost in their development. To do this, we need to build prototypes early in the design process that give users a realistic feel for the envisioned technology while retaining the complex usage contexts provided by in situ evaluation. An approach we use to get early feedback on technologies with which participants can interact is to adapt the WOz technique from HCI. In WOz evaluations, key aspects of "the system" are simulated by evaluators, however, the prototype appears fully (or at least partially) functional to the participant (Dahlback, Jonsson, & Ahrenberg, 1993). These simulations are typically done in lab settings where the participant's computer is actually watched and controlled by a hidden but (usually) proximate evaluator who mimics the responses anticipated from a working system. We have adapted the WOz technique to retain the simulation properties of early stage (or nonworking) ubicomp technologies while conducting evaluations in the field settings and contexts where such technologies are expected to be used. This provides essential usage context and pragmatic environmental constraints that might otherwise go unnoticed in more controlled settings.

Applying WOz to the CareNet Display. The CareNet Display is a digital picture frame (inspired by the Digital Family Portrait; Mynatt, Rowan, Craighill, & Jacobs, 2001) that augments a photo of an elder with information about his or her day. This information is needed by the friends and family who live near the elder and provide his or her day-to-day care. Icons surround the elder's photo and subtly change to represent status about events (e.g., a red morning medication icon indicates a problem or something unexpected like forgetting to take a pill, whereas a black icon suggests that the event occurred as expected). To get details about an event (e.g., what medications and doses the elder took and when), the user touches the icon and the photo of the elder is replaced with more information (Figure 1). For the in situ evaluation, the CareNet Display conveyed information about the elder's medications, meals, activities, outings, mood, and falls, and a shared calendar (see Consolvo, Roessler, & Shelton, 2004, for details).



Medication Information

Friday, August 15, 2003

Christina **did not** take the following medications:

Lopressor

Christina took the following medications at **8:30am**:

500mg Acetominophen

180mg Allegra

30mg Lipitor

Notes from Caregiver: Lopressor missed because prescription ran out. Christina needs to pick her prescription up at the pharmacy to continue with this medication.

FIGURE 1 Interacting with the CareNet Display. At left is the CareNet Display's main screen as it appeared in the in situ Woz evaluation. By clicking on the morning medication icon (1), the photo of the elder is replaced by a detail screen (2) showing what happened with the elder's morning medications.

To accurately assess the use and value of this technology for an elder and his or her support network, it was critical to situate it within the environment of its intended users. Thus, to gather early feedback about the CareNet Display concept and design, we deployed a "working" prototype in the homes of two to three family members per elder (four elders total) for 3 weeks each ($N = 13$). The prototype consisted of a touch-screen tablet PC housed in a custom-built wooden picture frame (Figure 2) and, to the participating family members, appeared to be fully functional. However, the sensors that we envisioned using to collect data about the elder's activities and events were not ready to be used in a multiweek deployment in an elder's home. Therefore, we simulated the sensor data collection by phoning the elder and/or primary caregiver several times per day each day of the 3-week deployments. We then updated the displays manually via a secure Web connection. We used General Packet Radio Service (GPRS) to provide always-on wireless

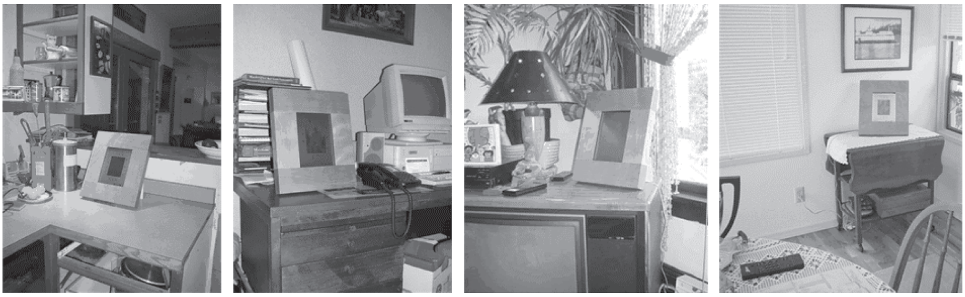


FIGURE 2 The CareNet Display prototype as deployed in the homes of participants during the in situ WOz evaluation. Participants kept the display in various "public" places within their homes including (from left to right) the kitchen, home office, TV room, and dining room.

Internet access so that updates could be sent to the displays at any time without the participants explicitly taking any action.

As the CareNet Display concept extended the idea of the Digital Family Portrait, so did its in situ WOz-style evaluation. In the original evaluation of the Digital Family Portrait, one family participated in what Mynatt et al (2001) refer to as a "field trial." The field trial lasted for 9 days and consisted of each participant having a laptop where the Digital Family Portrait was represented as a Web page. The researchers spoke to each participant once per day to collect the information to update the portrait, then prompted the participant to look at his or her version of the portrait after the daily phone call with the researcher.[1]

Lessons learned specific to in situ, evaluator-initiated data collection. A key advantage of WOz-style evaluations is that important design and system requirements can be discovered before much development effort has been put into building the underlying system. In situ WOz-style evaluations contribute even better requirements, as participants get a realistic feel for what it would be like to actually use the technology as part of their everyday lives. In the case of the CareNet Display evaluation, we learned that one of the seven types of information that we displayed (i.e., "falls") was less effective than another type of information that participants would prefer to have on their displays (i.e., "household needs," such as a lightbulb in the elder's bathroom needs to be replaced). To choose the initial seven types of information that were displayed, we conducted interviews and focus groups with various stakeholders involved in eldercare (Roessler, Consolvo, & Shelton, 2004a, 2004b). In these initial studies, falls ranked as the most important type of information about which familial caregivers wanted to know, whereas household needs ranked eighth. Prior to the in situ evaluation, participants agreed that falls should be on the display. However after living with the display for a few weeks, they felt that, although fall information was still critical, household needs would be better suited to a technology like the CareNet Display. This provided an important requirement to the sensor and system developers as well as the CareNet Display designers that we would not have learned at such an early stage in the development process had we limited ourselves to in-lab studies.

However, the important issue of the overall amount of time and the rigid schedule required of the researchers to effectively simulate the CareNet Display became clear from this evaluation. A researcher had to call the elder or primary caregiver up to six times per day from the early morning until late evening every day for four deployments, each of which lasted for 3 weeks. This schedule included weekends, nights, and holidays. The researcher required an Internet connection to send the updates to the displays after collecting the data by phone. Sometimes she or he had to go into a café with wireless Internet access and use his or her mobile phone to call

[1] Other researchers have also used WOz-style evaluations in situ. Examples are gathering requirements for organization-wide groupware systems (White & Lutters, 2003), simulating location information (Dearman, Hawkey, & Inkpen, 2005), and capturing images with a mobile field guide (Davies, Cheverst, Dix, & Hesse, 2005).

the elder at a time that was convenient for the elder (but inconvenient for the researcher). It also meant that the researcher needed his or her laptop, all necessary phone numbers, and the interview script much of the time. Furthermore, although we tried to rotate this responsibility between three researchers, the elders had established a rapport with one particular researcher and preferred talking with him; therefore the responsibility basically fell to one individual.

The scope of the study (number of participants, method of data collection, duration of deployments) made this research fairly labor intensive, exacerbated by conducting the evaluation in the field. It required careful planning, preparation, coordination, and effort. However, critical results would have been missed in more controlled laboratory settings where the system was not part of people's daily lives in their own homes.

2.2. In Situ Data Collection for Self-Report Techniques

Self-reported data (e.g., users responding to questions or creating a diary) help provide details about users' context, intentions, and actions that a log of system activity cannot capture. These data are particularly useful and more accurate when collected during or shortly after key moments of interest while still fresh in the user's mind. In situ self-report procedures have a distinct methodological advantage over ex situ inquiries as they

> do not require retrieval or reconstruction of data from memory but rather [involve] access to and accurate reporting of information available to conscious awareness. As a result [self-report data collected in the actual situation minimizes] the biases [associated with standard self-report procedures]. (Barrett & Barrett, 2001, p. 176)

In the past, in situ self-reports were often done via paper and pencil; that is, participants would carry around and fill out notebooks, typically formatted with predefined questions (Barrett & Barrett, 2001). Participants would complete these questions at predetermined times during the day (interval-contingent sampling), in response to events of interest (event-contingent sampling), or at random intervals triggered by an electronic device like a beeper or a phone call (signal-contingent sampling; Wheeler & Reis, 1991). This technique is often referred to as the Experience Sampling Method (ESM). These questionnaires may be administered over the phone, on paper, or most recently, using mobile devices (e.g., Barrett & Barrett, 2001; Intille et al., 2003; Consolvo & Walker, 2003). An advantage of conducting ESM using mobile devices such as personal digital assistants (PDAs) and mobile phones is that the evaluators may precisely control the amount of time a participant has to respond to a prompt before the prompt is recorded as missed, thus ensuring that recall bias has been minimized. In previous ESM studies that used beepers to trigger prompts, evaluators had much less control over such compliance. Barrett and Barrett offered a nice discussion of several other methodological advantages afforded by computerized ESM.

We recently adapted ESM to leverage mobile devices and technology sensed events (e.g., a location sensor indicates that a person has arrived at a particular destination, an audio sensor indicates that a person has completed a conversation, a motion sensor indicates that a certain object has been picked up). Once an event of interest has been sensed, an appropriate in situ questionnaire can be delivered. This sensor-based sampling technique, pioneered by Massachusetts Institute of Technology's Context-Aware Experience Sampling toolkit, is called context-triggered sampling (Intille et al., 2003). We extended it for use on mobile phones.

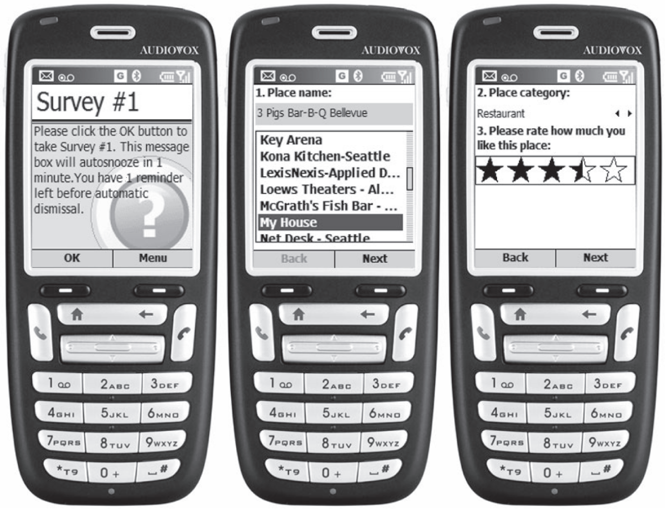
Combining ubicomp sensing technologies, ESM, and mobile device platforms, we can perform more sophisticated in situ evaluations. We discuss a specific investigation, "voting with your feet," in which we employed mobile phone-based ESM with signal-contingent and context-triggered sampling.

Applying self-reports to a study of "voting with your feet." Building on previous work on location systems (Hightower, Consolvo, LaMarca, Smith, & Hughes, 2005; LaMarca et al., 2005), we hypothesized that a person's visits to and from places in the physical world was an implicit form of expressing preference for that place. Visiting a place, in this sense, is analogous to voting—hence, "voting with your feet." To test our hypothesis, we conducted a 4-week in situ study with 16 participants using signal-contingent and context-triggered sampling on the SmartPhone (Froehlich, Chen, Smith, & Potter, 2006). The software used for this evaluation was a .NET-based experience sampling toolkit that we developed called *My experience* (*Me*), which runs on SmartPhones, PocketPCs, TabletPCs, and desktop machines running Microsoft® Windows. The *Me* toolkit has wireless Web synchronization, sensor plug-in architecture, multimedia capture, and a highly flexible XML-based input for survey construction. It supports interval-contingent, event-contingent, signal-contingent, and context-triggered sampling as well as all of the functionality available in an earlier PDA-based open source tool we built called iESP[2] (Consolvo & Walker, 2003).

In the study, questionnaires were triggered based on the movements of participants. When our phone-based location sensors determined that a participant had been at a place for about 10 min, a questionnaire was triggered asking about that place (Figure 3). We used the phone's wireless connection to monitor participation and detect technical issues with the mobile phones, which allowed us to react quickly to problems as they occurred.

An inherent problem with current mobile devices for conducting experience sampling is that text entry for open-ended questions is difficult and time consuming. To maximize user response, multiple-choice questions, true–false questions, text auto-completion, and numerical ratings (e.g., Likert scales) are much more efficient. However, this efficiency is at a cost of losing qualitative data. Thus, in this study, we augmented the mobile questionnaires with Web-based diaries to gather short, qualitative accounts. The Web diary data were partially structured by auto-

[2] The *My Experience (Me) Toolkit* and *iESP* are available from http://seattleweb.intel-research.net/projects/esm/



Survey #1

Please click the OK button to take Survey #1. This message box will autosnooze in 1 minute.You have 1 reminder left before automatic dismissal.

1. Place name:

3 Pigs Bar-B-Q Bellevue

Key Arena
Kona Kitchen-Seattle
LexisNexis-Applied D...
Loews Theaters - Al...
McGrath's Fish Bar - ...
My House
Net Desk - Seattle

2. Place category:

Restaurant

3. Please rate how much you like this place:

FIGURE 3 ESM questions on the SmartPhone Audiovox SMT 5600. Shown are examples of the survey prompt screen (left), text entry auto-completion (middle), and rating scale (right) questions used in the "voting with your feet" study.

matically displaying the time-stamped location activity data supplied from the field, which were sent to and synchronized with our Web server once an hour. When participants logged in, they would see a chronologically ordered journal view of their place visit activity. The diary system would then ask specific qualitative questions about one or two randomly selected public place visits. In particular, we asked the reasons for their visit, feedback about their rating, and if they would recommend the place to others. Despite the obvious additional burden that this technique placed on participants, 5.8 Web diary sessions per week were completed per participant. This exceeded our expectations, as the study consent form requested only four sessions per week.

Using context-triggered sampling, we were able to collect 3,458 in situ questionnaires from the 16 participants about their travel behaviors, place visit activities, and the accuracy of our mobility inference algorithms (i.e., we could sense when the phone was stationary vs. mobile based on Global System for Mobile Communication signals; Smith, Chen, Varshavsky, Sohn, & Tang, 2005). As we found during interviews, remembering the places that people go from day to day is a challenging memory-recall task. Most participants had difficulty remembering where they had been in the past few days much less during the course of a 4-week study. By collecting data in situ, we gained a much more accurate picture than we would have been able to collect from retrospective surveys or intermittent interviews. Instead, we were able to determine it empirically during the course of the study. In addition, we were able to assess the accuracy of our mobility inference algorithm by relating in situ questionnaire responses to the sensor data streams. In this way, the questionnaires allowed us to study human behavior and validate aspects of our ubicomp technology.

Lessons learned specific to context-triggered, self-report data collection. There are a number of advantages of adapting in situ self-report/ESM techniques for mobile devices and sensor technologies. For example, participant compliance can be assessed more accurately (Barrett & Barrett, 2001) as it is difficult for participants to modify data or diary entries post hoc and the data can indicate exactly when a self-report was completed, how long it took, and if any responses were changed before completion.

Wireless connectivity provides real-time information about participation and allows the researcher to access responses as they are being supplied, thus enabling the preparation of targeted ex situ inquiries (e.g., individualized interviews). With real-time data transmission, preliminary data analysis can begin almost immediately and researchers can identify and troubleshoot problems as they occur during the study. Researchers are also able to remotely send new questions to the devices or trigger questionnaires dynamically.

Computerized self-reports can also be more sophisticated in their presentation of questions than, for example, beeper studies that use paper diaries. For example, questionnaires delivered by mobile devices may contain the following:

- Conditional questions (e.g., ask Question C only if Question B's response was "yes" and Question A's response was "no")
- Probabilistic questions (e.g., ask Question A 30% of the time and Question B 100% of the time)
- Multimedia questions (e.g., video-based questions)
- Specific question frequencies (e.g., only ask Question E once per unique answer to Question D)
- Question–answer order randomization (i.e., the presentation of answer data is randomized from questionnaire to questionnaire to reduce response bias; Barrett & Barrett, 2001)

In addition, auxiliary inputs like cameras or microphones can be used to augment responses and provide a richer understanding of the participants' experiences (Consolvo & Walker, 2003).

Finally, context-triggered sampling (e.g., Intille et al., 2003) where sensors automatically detect events of interest or infer context and then trigger a questionnaire can be used. This technique has several advantages: (a) questionnaires occur only during events of interest, which reduces participant disruption when compared with interval- or signal-contingent sampling; (b) context data can be continuously saved, allowing the researcher to validate participant responses and potentially uncover unanticipated behavioral patterns; and (c) the effectiveness and accuracy of new technologies can be tested.

However, despite the targeted sampling we were able to achieve with the context-triggers, participants did not always react positively to the questionnaire prompts. During interviews, participants responded that they typically did not mind answering a questionnaire unless it came at a particularly inconvenient time (e.g., while they were driving, engaged in conversation, or eating). In addition, occasionally, the mobile phones themselves would crash. Fortunately we were able to detect the technical issues early because of the wireless connectivity (e.g., if our servers did not receive a signal from a study mobile phone for a few hours, we would call or e-mail the participant and help him or her reboot the study phone). Participants also mentioned during the interviews that they began to expect a questionnaire when they arrived at a new place. If no questionnaire was administered, they wanted the option of triggering one manually (at the time of this study, the *Me* toolkit did not support manual triggers, although it does now).

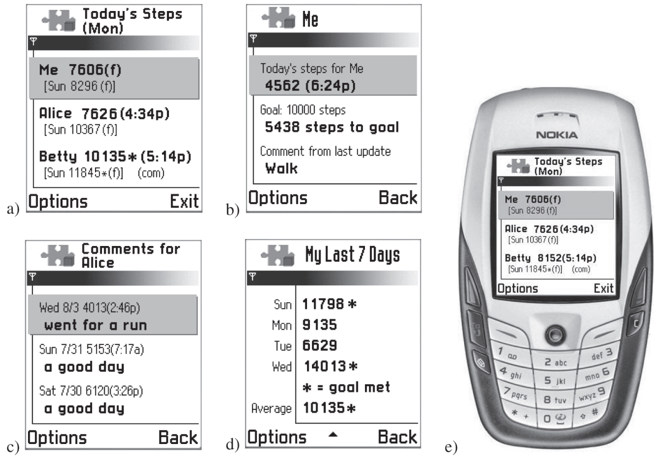
2.3. In Situ Data Collection for a Mobile Application Prototype

Thus far we have described an in situ evaluation method where data were collected manually by an evaluator directly contacting participants to obtain data (evaluator initiated and evaluator entered) and a second method where in situ data collection was initiated by context triggers (location sensing in particular), and the participants' entered self-reported data. We now discuss an in situ evaluation of a ubicomp application where the data collection was participant initiated (rather than evaluator or context determined) and participant entered.

Applying in situ data collection to a study of social influence on physical activity. As part of our UbiFit project, we are investigating how technology can help encourage people use a mobile phone and pedometer to provide personal awreness of physical activity. In our first application (Consolvo et al., 2006), we use a mobile phone application to provide personal awareness of activity level, progress toward a daily goal, and mediate physical-activity-related social interaction among a small group of friends (or "fitness buddies"). The user carries a mobile phone and wears a pedometer throughout the day. The user manually enters her or his step count into the mobile phone application at any time throughout the day, at which time she or he can add a comment (e.g., "went for a run," "slow day—paper deadline") and send the count (and comment) to any or all of her or his fitness buddies. Users can also send messages to buddies independent of entering a step count (e.g., "Great job!," "Want to go for a walk?"). In addition, the mobile phone application provides information regarding how users are performing toward their daily step count goal, calculates their average daily step count, and provides access to final counts for the past 7 days. If a user's buddies have shared their step counts, the user can also see information for their buddies' performance (Figure 4).

We conducted an in situ evaluation where three groups consisting of four to five women per group used the mobile phone application for 3 weeks throughout their everyday lives ($N = 13$). Each group of women was from a preexisting social network. Prior to the in situ evaluation, we piloted the mobile phone application on several members of the research team and nonresearch staff for more than 6 weeks.

The mobile phone application for the evaluation was developed in Python on the Nokia Series 6600 mobile phone and was accompanied by the Omron HJ-112 pedometer. When a participant entered her step count, an e-mail was sent to our central server, which harvested the e-mail and recorded the content in an XML da-



a)

Today's Steps (Mon)

Me 7606(f)
[Sun 8296 (f)]

Alice 7626(4:34p)
[Sun 10367(f)]

Betty 10135*(5:14p)
[Sun 11845*(f)] (com)

Options Exit

b)

Me

Today's steps for Me
4562 (6:24p)

Goal: 10000 steps
5438 steps to goal

Comment from last update
Walk

Options Back

c)

Comments for Alice

Wed 8/3 4013(2:46p)
went for a run

Sun 7/31 5153(7:17a)
a good day

Sat 7/30 6120(3:26p)
a good day

Options Back

d)

My Last 7 Days

Sun	11798*
Mon	9135
Tue	6629
Wed	14013*
	* = goal met
Average	10135*

Options ▲ Back

e)

Today's Steps (Mon)

Me 7606(f)
[Sun 8296 (f)]

Alice 7626(4:34p)
[Sun 10367(f)]

Betty 8152(5:14p)
[Sun 11845*(f)] (com)

Options Exit

FIGURE 4 Mobile phone application screen shots and the Nokia 6600. (a) main screen, (b) detail screen, (c) recent comments, (d) trending information, and (e) the main screen on the Nokia 6600 mobile phone. The * in (a) and (d) signifies that the participant met her daily step count goal.

tabase, and then the server sent out any data to be shared to fitness buddies. The phones kept a record of data entered by the participant, allowing us to retrieve any unreceived data at the end of the study (e.g., if the participant sent an update when her phone was out of range of phone service). We provided a study mobile phone, Subscriber Identity Module (SIM) card, and prepaid phone service to each participant (i.e., the study phone was carried in addition to each participant's personal mobile phone).

The goal for this study was not to see if this particular technology had successfully affected a sustained increase in physical activity but rather to understand important design requirements for physical activity and exercise related applications that we plan to develop in the future. In particular, one future project we are planning in this area involves using sensor technology that will detect and differentiate various types of physical activities as they occur and can automatically communicate these data via Bluetooth to a mobile phone. Participants can augment these data by answering ESM-style questions on the mobile phone. However, eliciting these physical activity monitoring requirements was only possible through an in situ evaluation of the mobile phone application's use in everyday life. It was important to see how this technology was used in the context of people's physical activity throughout the day.

Lessons learned specific to in situ, user initiated data collection. Some important results of this study involved the pedometer as an (in)adequate measure of physical activity and how continual awareness affected behavior. For example, despite the wide-spread use in medical studies of pedometers as a measure of physical activity (Tudor-Locke, 2002), many participants expressed frustration with the pedometers misrepresenting their overall level of physical activity. It was particularly important to participants that the pedometer accurately convey their overall level of physical activity because their performance relative to a goal was being logged in the mobile phone application for personal and shared use (i.e., if the participant's daily step count was 3,000, her friends would not necessarily know if she had not been very active that day, or if she had an active day that involved an activity that the pedometer did not track well, such as cycling; Consolvo et al., 2006). Several participants also reported that knowing how they were doing with respect to their daily goal throughout the day in many cases prompted an unplanned walk, particularly if they were sharing progress with their buddies. These results are helping us design an application and physical activity detection device to better support our target users and prepare for a longer-term study to investigate affects on sustained behavior change.

2.4. General Lessons Learned From In Situ Data Collection

In addition to the lessons learned that are specific to the in situ data collection techniques previously mentioned, we have also learned lessons that apply more generally to collecting data in situ.

For example, we learned critical considerations that impacted the design of the CareNet Display and had implications for its underlying infrastructure that were not due to the simulated, WOz-style nature of the evaluation per se but rather the in situ nature of the evaluation that allowed participants to experience the technology in their everyday lives. For instance, we found that elders' initial impressions about how they thought the CareNet Display should work changed after their experiences in the deployments. The elders originally claimed that they would be very open about sharing the types of information the display conveyed to the local family and friends who provided their care. However, postdeployment, elders were more conservative about with whom they would share such information. For instance, one elder reflected that she would not be comfortable sharing information regularly with her grandson (who was an alcoholic), though in the initial interview, she claimed to be comfortable sharing such information with him. Day-to-day use of the technology gave her a clearer idea of the implications of what it would be like to actually use it. In another case, elders who shared information about their medication management through the CareNet Display discovered that this perceptibly altered the nature and dynamic of conversations with family members. Because this information was now conveyed automatically to family, family members no longer felt it necessary to "nag" the elder about whether the elder had taken his or her pills that day. Instead, conversations were spent on more social and less "monitoring-like" topics, making it feel less stressful and more enjoyable to all parties. This outcome could not have been predicted or observed without in situ use. These evolutions in technology use are a direct consequence of the often subtle differences between understanding and projecting how a technology will be used and actually using it day to day with real people in their real roles with their real social relationships.

We also learned about using mobile phones to collect in situ data. For example, although current mobile phones often last several days under normal use before needing to be recharged, the types of nonstandard, mobile phone applications that we have built often need to be recharged each night, as they typically require more processing power, interaction, and/or display use. The requirement of recharging nightly was particularly problematic with participants who were not in the habit of charging their phones every night. It takes time for participants to establish this as a routine practice, and thus for shorter term studies (even several weeks in duration) this happened inconsistently.

Another consideration is response time. It may take several seconds to send a message with the phone, access a database, or do sophisticated calculations. In this time, the participant may become frustrated, bored, or think he broke the application and try various buttons to "fix" it—a behavior that may crash fragile, early-stage applications. This delayed response time may have a negative effect on the participant's perception of the application (Shneiderman, 1984). In addition, the delivery times of Short Message Service messages can vary depending on phone network traffic, potentially creating difficulties for applications that involve social mediation.

An additional consideration is cost—both monetary and resource. Simply put, phones that are able to run nonstandard software are often expensive, and they require significant setup and maintenance time from a skilled technical researcher. There are also general difficulties that researchers face with the in situ deployment of any device: Participants may lose or damage the device, mistakenly (or intentionally) reset it, or quit the study application and be unable to restart it; we have experienced all of these, though thankfully not often.

When running studies that use mobile phones, it is important to determine whether to use the participant's personal SIM card in the study phone or provide a study phone with a study-provided SIM card to carry in addition to the participant's personal phone. People rely on their mobile phones; prototype applications can potentially crash the phone and prevent the participant from receiving or making calls (which may make their phone unreliable during the study). If using the participants' SIM cards, the researchers may need to reimburse participants for phone service charges incurred as a result of participating in the study (and the participants may not know of some charges until after receiving their bill). In addition, information that participants rely on from their personal phones (phone numbers, addresses, calendar entries, familiar ring tones, etc.) often must be replicated on the study phone—a potentially time-consuming and error-prone task that raises privacy concerns. However, a study-specific phone requires participants to carry two phones—their personal phone and the study phone; thus participants may be less likely to carry the second phone with them or use the application with the same frequency as they would if it were on their primary phone.

Which approach to take thus depends on several factors, including the type of application, robustness of the application, expected frequency of interaction, type (and size) of phone used, and the remuneration scheme. Unfortunately, although we have used both tactics in our studies, we have not found the answer to which is "right," as both tactics have important advantages and disadvantages. This is a key consideration that we continue to investigate.

Even "simulated" designs may assume certain technologies are at hand and available for data collection and transmission (e.g., Internet access, phone lines), which may not be the case in field settings. Homes may not have needed services or technologies (e.g., Internet access, mobile phone coverage), and even if they do, there are ethical and financial implications of using participants' personal connections or services for the researchers' or study's use. Use of these services can be an invasion of privacy; incur hidden costs (e.g., use of mobile phone minutes); and unacceptably tie up services, making them unavailable (e.g., phone is busy, Internet connection in use). In the CareNet Display evaluation, we used wireless GPRS to exchange data because it was more feasible than installing temporary extra services such as DSL or additional phone lines. Thinking through these issues is a critical but nonobvious part of the study design and planning process, particularly when such services are required to simulate aspects of an only partially working system.

Aesthetics also play a key role (which should be no surprise to designers) yet emerge very early when doing in situ concept tests. Likewise, a number of ergonomic or industrial design issues are identified almost immediately, many of which would not be apparent in lab-based simulations. For instance, in our study of physical activity over the course of the day, participants in a pretest at some point used a washroom. This uncovered a problem with the device attachment/clip design whereby pedometers could dislodge and fall into the toilet. We subsequently used pedometers that had a clip and a safety strap to provide a secondary attachment mechanism should the clip slip off. We also had to investigate how the placement of the devices affected the reliability of the data. For instance, we discovered that pedometers produce less reliable data when attached to pants that have unstructured waistbands such as drawstring ties (these garment styles are commonly worn in summer or when when performing physical activities). These practical design considerations impact both the device design and the instructions and training guidelines for participants.

Finally there are a number of important points relating to security and privacy. (These are not topics we were able to cover in this article given the focus, though they are crucial and we do not wish to underemphasize them.) We wish to raise two particular points as they relate to in situ data collection in the examples just presented. First, any time researchers visit the home of someone they do not know, they must take precautions to make the participants and the researchers feel safe and comfortable. We try to have pairs of male and female researchers (vs. only male, only female) during such visits, particularly if the participant is a woman who may be home alone. In addition, any allergies or fears the researchers may have of domestic pets need to be considered prior to a home visit. Finally, it is important that participants have a clear idea of what data are being collected, when data are being collected, and who will have access to the data. This is generally viewed as a key part of conducting ethical research in any situation, however, it is of particular importance when dealing with technology that is embedded into the fabric of people's daily lives and relationships.

3. CONCLUSIONS

End user applications developed for or with ubicomp technologies pose particularly challenging demands for in situ use and hence for in situ evaluation. Traditional evaluation methods that are well suited for more predictable domains (e.g., evaluating graphical user interfaces, explicit user interactions, or desktop applications) need to be augmented for these unpredictable and often mobile real-world contexts where user interactions may be implicit or sensed as well as explicit. In this article, we discuss how we have augmented interviewing and questionnaires with three different in situ data collection methods to be more amenable to the real-world demands of ubicomp applications. Based on our experiences with a variety of applications and user studies, some of which were outlined here, we have summarized benefits and challenges we have observed in using these in situ methods. Although our focus has been on ubicomp applications and tools for their assessment, we believe that the in situ evaluation tools we propose will be generally useful for field trials of other technology, applications, or formative studies that are concerned with collecting data in situ. We are collaborating with other companies and universities to apply our in situ tools to their field studies to this end. We continue to work on evolving the tools and methods proposed here, and we are adapting other techniques to apply to in situ evaluation in the future.

REFERENCES

Barrett, L. F., & Barrett, D. J. (2001). An introduction to computerized experience sampling in psychology. *Social Science Computer Review, 19,* 175–185.

Consolvo, S., Everitt, K., Smith, I., & Landay, J. A. (2006). Design requirements for technologies that encourage physical activity. *Proceedings of the Conference on Human Factors and Computing Systems: CHI 2006*, 457–466.

Consolvo, S., Roessler, P., & Shelton, B. E. (2004). The CareNet Display: Lessons learned from an in home evaluation of an ambient display. *Ubiquitous Computing: 6th International Conference on Ubiquitous Computing*, UBICOMP 2004, pp. 1–17.

Consolvo, S., & Walker, M. (2003). Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing Magazine: The Human Experience, 2*(2), 24–31.

Dahlback, N., Jonsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—Why and how. *Proceedings of the International Workshop on Intelligent User Interfaces: IUI 1993*, 193–200.

Davies, N., Cheverst, K., Dix, A., & Hesse, A. (2005). Guiding and navigating: Understanding the role of image recognition in mobile tour guides. *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services: Mobile HCI '05,* 191–198.

Dearman, D., Hawkey, K., & Inkpen, K. M. (2005). Effect of location-awareness on rendezvous behavior. *CHI '05 Extended Abstracts on Human Factors and Computing Systems,* 1929–1932.

Friedman, B., & Kahn, Jr., P. H. (2003). Human values, ethics, and design. In J. A. Jacko & A. Sears (Eds.), *The human-computer interaction handbook: fundamentals, evolving technologies, and emerging applications* (pp. 1177–1201). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Froehlich. J., Chen, M., Smith, I., & Potter, F. (2006). Voting with your feet: An investigative study of the relationship between place visit behavior and preference. *Ubiquitous Computing: 8th International Conference on Ubiquitous Computing, UBICOMP 2006*, 333–350.

Hightower, J., Consolvo, S., LaMarca, A., Smith. I., & Hughes, J. (2005). Learning and recognizing the places we go. *Ubiquitous Computing: 7th International Conference on Ubiquitous Computing, UBICOMP 2005,* 159–176.

Intille, S. S., Rondoni, J., Kukla, C., Ancona, I., & Bao, L. (2003). Context-aware experience sampling tool. *CHI '03 Extended Abstracts on Human Factors in Computing Systems,* 972–973.

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I. E., Scott, J., et al. (2005). Place lab: Device positioning using radio beacons in the wild. *Pervasive Computing: 3rd International Conference, PERVASIVE 2005*, 116–133.

Moran, T., & Dourish, P. (2001). Introduction to this special issue on context-aware computing. *Human–Computer Interaction, 16,* 87–97.

Mynatt, E. D., Rowan, J., Craighill, S., & Jacobs, A. (2001). Digital family portraits: Supporting peace of mind for extended family members. *Proceedings of the Conference on Human Factors and Computing Systems: CHI 2001,* 333–340.

Roessler, P., Consolvo, S., & Shelton, B. (2004a). Phase #1 of Computer-Supported Coordinated Care Project (IRS-TR-04-005). Retrieved May 10, 2006, from http://www.intel-research.net/Publications/Seattle/022320041335_230.pdf

Roessler, P., Consolvo, S., & Shelton, B. (2004b). Phase #2 of Computer-Supported Coordinated Care Project (IRS-TR-04-006). Retrieved May 10, 2006, from http://www.intel-research.net/Publications/Seattle/030520041041_233.pdf

Scholtz, J., & Consolvo, S. (2004). Toward a framework for evaluating ubiquitous computing applications. *IEEE Pervasive Computing Magazine, 3,* 82–88.

Shneiderman, B. (1984). Response time and display rate in human performance with computers. *Computing Surveys, 16,* 265–285.

Smith, I., Chen, M., Varshavsky, A., Sohn, T., & Tang, K. (2005). Algorithms for detecting motion of a GSM mobile phone. *ECSCW '05 Workshop on Location Awareness and Community,* Paris, France.

Suchman, L. (1987). *Plans and situated action: The problem of human-machine communication*. New York: Cambridge University Press.

Tudor-Locke, C. (2002). Taking steps toward increased physical activity: Using pedometers to measure and motivate. *President's Council on Physical Fitness and Sports: Research Digest* (Series 3, No. 17).

Wheeler, L., & Reis, H. T. (1991). Self-recording of events in everyday life: Origins, types, and uses. *Journal of Personality, 59,* 339–354.

White, K. F., & Lutters, W. G. (2003). Student research projects: Behind the curtain: Lessons learned from a Wizard of Oz field experiment. *ACM SIGGROUP Bulletin, 24,* 129–135.