User Study Techniques in the Design and Evaluation of a Ubicomp Environment

Sunny Consolvo[1], Larry Arnstein[2], and B. Robert Franza[3]

[1] Intel Research Seattle
sunny@intel-research.net
[2] Department of Computer Science & Engineering, University of Washington
larrya@cs.washington.edu
[3] Cell Systems Initiative, Department of Bioengineering, University of Washington
bfranza@u.washington.edu

Abstract. To be successful, ubicomp applications must be designed with their environment and users in mind and evaluated to confirm that they do not disrupt the users' natural workflow. Well-established techniques for understanding users and their environment exist, but are not specifically designed to assess how well the computing and physical task environments blend. We present strengths and weaknesses of several qualitative and quantitative user study techniques for ubicomp. We applied these techniques to the design and evaluation of a ubicomp application for cell biology laboratories (Labscape). We describe how these techniques helped identify design considerations that were crucial for Labscape's adoption and demonstrate their ability to measure how effectively applications blend into an environment.

1 Introduction

Weiser described a vision of ubiquitous computing where computers are so well integrated into the task environment that they vanish into their surroundings [21]. For ubicomp applications to "vanish," they need to be designed with their environment and users in mind and evaluated to confirm that they augment, not disrupt, the users' natural workflow. Yet as Abowd et al. [3] have pointed out, little research has been published on user study techniques that best address the challenges of ubicomp. We present an approach to the iterative design and evaluation of ubicomp environments that establishes a baseline assessment of the environment against which subsequent enhancements and modifications can be evaluated.

Our approach includes a combination of existing qualitative and quantitative user study techniques, some of which have been borrowed from other disciplines. The contribution of this paper is to present the strengths and weaknesses of several user study techniques based on our experiences in applying them to a ubicomp application for cell biology laboratories. Our long-term goal is to establish a principled approach for the design and evaluation of ubiquitous computing environments.

We begin with a discussion of the current state of the art in the evaluation of ubicomp environments. In Section 3, we continue with a brief survey of relevant user

G. Borriello and L.E. Holmquist (Eds.): UbiComp 2002, LNCS 2498, pp. 73-90, 2002.

study techniques to set the stage for our main contributions. In Section 4, we introduce Labscape, a smart environment that serves as the example application to which we applied the techniques. In Section 5, we discuss results from the user study techniques that were applied to the design of Labscape. In Section 6, we present initial results obtained from Labscape's evaluation. We end with a discussion, future work, and conclusions.

2 Related Research

Evaluations have been conducted for a variety of ubicomp applications. We focus on evaluations of capture and guide systems, as they are most similar to Labscape.

Classroom 2000 [1] is an instrumented classroom that captures live lectures in a form that can be accessed later. An iterative design process was used, involving representative users in an authentic setting. Qualitative data was collected from surveys, quantitative data from usage logs, and a comparative study was done to assess the impact on student performance. The main differences between their evaluation and the technique we discuss in Section 6 are that they used a control group rather than a baseline, and none of their quantitative data came from observations.

Tivoli is a meeting capture and salvage system [12]. It was evaluated in an authentic setting. Data was collected in a variety of ways: the meetings were captured on video, meeting artifacts were kept, users were interviewed, and logs were made of the users' interactions with the system. The main differences between their evaluation and our technique are that they did not use a control group or establish a baseline, and their quantitative data came from surveys and usage logs, not observations.

The GUIDE project is a context-aware electronic tour guide that was deployed in the city of Lancaster [7]. Similar to our work, they performed interviews and observations to influence their design. Their evaluation consisted of an expert walkthrough and a field trial. The field trial took place in an authentic setting with representative users. The main differences between their field trial and our technique are that most of their data were qualitative, relying on direct observation and interviews; their quantitative results came from system usage logs, not observations.

Similar to the GUIDE project is E-graffiti, a context-aware electronic guide for the Cornell campus [6]. E-graffiti was evaluated with representative users in an authentic setting. Participants were asked to perform a combination of real and contrived tasks. Data were collected in the form of system usage logs and questionnaires that users completed after using the system. Observation was not used in the evaluation, nor was a control group used or baseline established.

3 Survey of Relevant User Study Techniques

The discussion in the remainder of this paper assumes a general understanding of user study techniques. We include this section to introduce readers to the techniques most

relevant to our work. Much of what is discussed below can be found in detail elsewhere [9, 15, 16, 17, 20].

These techniques are appropriate for different stages in the development of an application, from initial concepts, to design and evaluation of a working application. Based on previous experience, we are in favor of using multiple techniques and advocate that the participants be representative of the target user population.

3.1 Contextual Field Research

Contextual field research (CFR) is a technique for gathering qualitative data by observing and interacting with users as they go about their normal activities. It is typically used to discover how users think and act rather than to test preformulated hypotheses. Data is collected by a combination of note taking, video, audio, and photographs. Some benefits of CFR are that it is conducted in the user's environment rather than the laboratory, users perform their normal activities rather than contrived tasks, and because no application needs to be in place to conduct CFR, it may be used to help guide the application's requirements and design.

However, CFR has disadvantages. Users may alter their behavior when they know they are being observed. It can be more expensive than other qualitative techniques. The cost of CFR can be difficult to gauge before it begins, as the evaluator may not know what he will learn, how much data he will have to collect, or how long the observations will take. He may also not know how long the data will take to analyze. The evaluator cannot guarantee that the sessions he observes are typical for the users.

Despite the disadvantages, the quality of data from CFR is often better than that from other techniques; evaluators do not have to rely on the user to remember everything about his work and environment, nor must evaluators worry about inventing appropriate tasks for the user to attempt.

3.2 Intensive Interviewing

Intensive interviewing is a technique for gathering qualitative data by asking users open-ended questions about their work, background, and ideas. Unlike more structured interviewing techniques, question order and content may vary from user to user. As with other interviewing techniques, evaluators must ask questions in such a way as to not influence users' responses. Several hours are often spent with each user over a series of one to two hour sessions; the total time spent with each user is typically between six and fifteen hours. Similar to CFR, data is captured by a combination of note taking, video, and audio. Some benefits of intensive interviewing are that evaluators learn about the user's work in the user's own words, and it is relatively inexpensive compared to observational techniques. Intensive interviewing also helps evaluators establish a rapport with the user, which can be particularly useful when the evaluators intend to use additional user study techniques. Because intensive interviewing does not need to be performed in the user's environment, the evaluators do not have to disrupt that environment, and scheduling may be easier.

Because everyday actions can become automatic, a significant disadvantage of intensive interviewing is that users will often fail to mention important aspects of what they do [23]. Similar to CFR, evaluators do not know how much time they will need to spend with each user. In theory, the interview process stops when the evaluator is not learning much new information. In practice, the interview process often stops before that point is reached, due to resource and time constraints. Another disadvantage is that audio transcription is time consuming. As mentioned above, intensive interviewing does not need to be conducted in the field. Though that has its benefits, it also has disadvantages. When interviewing is conducted in the field, being in the user's environment may serve to jog his memory; for example, he may be more likely to explain how he uses things in his environment. Outside of his environment, he may neglect to mention that information.

Intensive interviewing can be a good technique to use when combined with observational techniques. It provides valuable information, but is not comprehensive enough to be used on its own.

3.3 Usability Testing

Usability testing is a technique for gathering empirical data by observing users as they perform tasks with the application that is being evaluated. There are several variations of usability testing; we discuss informal, qualitative studies involving between five and fifteen users per study. Usability testing may be conducted in the field, but it is more commonly conducted in a usability laboratory where equipment for recording and observing the sessions is available. The goal of usability testing is to create an application that is easy to use and provides appropriate functionality for its users. This is usually done in an iterative process of testing followed by improvement. Usability testing is inexpensive compared to other observational techniques, and results can be generated quickly. If testing is conducted in a usability laboratory, an additional benefit not shared by the other techniques we discuss is that members of the development team can observe the testing as it takes place.

A significant disadvantage of usability testing is that the testing situation is artificial: even if testing takes place in the field, both the tasks and situations are contrived. Even if the application tests well in the study, there is no guarantee that the application will be a success in practice. Another disadvantage is that, as with CFR, users may alter their behavior because they know they are being observed.

Usability testing can be a good technique for some domains. However, as we discuss in Section 6, the disadvantages outweigh the benefits for ubicomp.

3.4 Lag Sequential Analysis

Lag Sequential Analysis (LSA) is a technique for gathering quantitative data by observing users as they perform their normal activities. It is traditionally used in the field of developmental psychology to study the behavior of person to person interac-

tion by measuring the number of times certain behaviors precede or follow a selected behavior; the behaviors are defined by the study evaluators. Data can be captured live with paper and pencil or coded from video. LSA shares two benefits of CFR: it is conducted in the user's environment, and it is conducted while the user performs his normal activities. With LSA, evaluators can generate statistics that capture aspects of observed behavior such as frequency and conditional probabilities of events. If video is used to capture the data, it can be re-coded for different information as evaluation needs change, and it can be used for qualitative observational purposes.

A significant disadvantage of LSA is cost; coding video for LSA is time consuming. When using more than one coder, the reliability of the different coders must be calculated (e.g., by using Cohen's Kappa—a statistic used to assess inter-rater reliability [8]). As with CFR, evaluators cannot guarantee that the activities they observe are typical for the users. As with CFR and usability testing, users may alter their behavior because they know they are being observed.

LSA is an expensive technique that can generate quantitative and statistical data. As we discuss in Section 6, it can be a good technique for ubicomp environments.

4 Labscape

Labscape is a ubiquitous computing application that helps biologists in their laboratory environment. Labscape has two objectives. First, it seeks to make critical information available to biologists when and where they need it to minimize distractions and errors. Second, it allows biologists to easily capture and organize data that is generated in the process of conducting an experiment, in a structured format that is searchable and sharable. In this section, we summarize the physical environment of the biology laboratory and the information needs of the biologists [4]. This material provides the background for the discussion in Sections 5 and 6.

4.1 A Cell Biology Research Laboratory

Our primary collaborator on Labscape is the Cell Systems Initiative (CSI), part of the Bioengineering Department at the University of Washington. Five biologists share the immunology laboratory at CSI—three are full-time researchers and two are students. The laboratory consists of one main room, two auxiliary rooms, and some equipment in the hallway (see Fig. 1). While performing work in the laboratory, the biologists frequently move between various stations, as the stations are highly task-specific. The biologists primarily work in the main laboratory, but occasionally use the other areas. Though the researchers each have a small station in the main laboratory that is considered their personal space, the majority of the laboratory and equipment is shared; the students do not have any personal space.

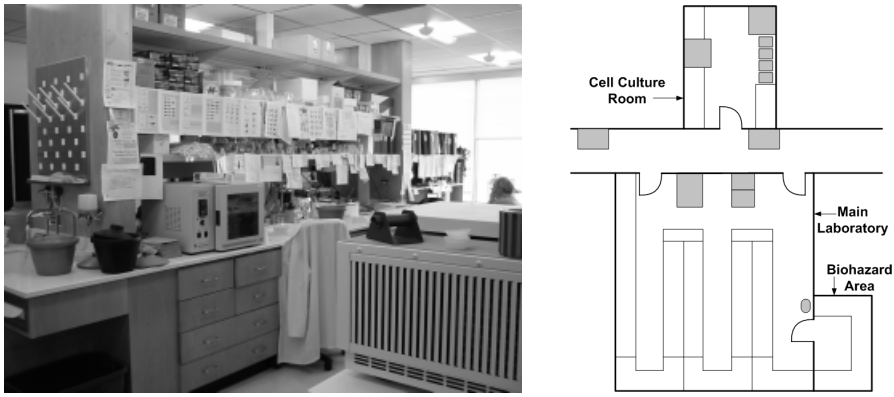

Fig. 1. Photo and layout of the laboratory at CSI. Most of the biologists' work is performed in the main laboratory, though they also use the other areas

Although the laboratory is cluttered with tools and equipment, cleanliness is paramount. Surfaces are kept clean, as contamination is an important concern. In fact, issues with contamination may affect how laboratory notebooks and other notes are used in the laboratory. Motors in the equipment create a constant background noise. A variety of reference documents—print-outs, hand-written notes, copies of pages from books—hang from walls and shelves throughout the main laboratory. Because the biologists perform a variety of procedures in the laboratory, the equipment is not laid out in any particular order; the laboratory does not function like an assembly line.

Biologists perform their work while moving around in their environment. The tools and equipment they use were built by a variety of manufacturers. Biologists are seldom at one station for very long. Pencil and paper is the primary form of information support available to them in the shared areas of the laboratory.

4.2 Information Needs

Biologists need to plan, execute, and document their laboratory work. In planning, records of previous procedures may be consulted to avoid introducing unintended variability into the experiment and to review previous results that may influence their plans. During the procedure's execution, biologists may need to access their plans, track progress, and record observations and data. Finally, biologists must formally document their work for future reference and legal compliance.

Biologists meet their information needs in a variety of ways; the most prevalent is through the use of pencil and paper. In addition, commercial laboratory information management systems and electronic laboratory notebooks can be used to organize and access data produced by laboratory experiments [11, 13]. Such systems have penetrated highly repetitive clinical and production laboratories, especially those having stringent legal record-keeping requirements. However, these tools are rarely found in research-oriented laboratories that require flexibility and rely on voluntary use of

information technology. New computing tasks that do not contribute to the biologists' abilities to perform good experiments are quickly abandoned.

Labscape is a ubiquitous laboratory assistant that satisfies these information needs without distracting biologists from their work: it presents needed information in the context of the experiment, it records experiment data and observations as the work is performed, and it provides ubiquitous access to the experiment record [5]. As we develop a better understanding of the biologists' needs and how technology might help, we can further enhance the environment to improve their ability to focus on the biology rather than on the information support system.

Biology research is a goal-oriented activity that allows for iterative assessment of performance on similar tasks before and after the deployment of new technologies. As a result, Labscape is an excellent test case for user study techniques in the iterative design and evaluation of ubicomp applications.

5 User Study Techniques Applied to the Design of Labscape

In this section, we discuss how two user study techniques helped us design Labscape: intensive interviewing and contextual field research (CFR). To design Labscape, we needed to gain a general understanding of the biologists' work and environment. We also needed answers to some specific questions. In particular, we were interested in learning whether computing should be distributed throughout the environment, carried by the user, or a combination of the two. In addition, we wanted to know where, how, and why biologists accessed and recorded information during experiments.

We started with intensive interviewing, as it is a relatively fast way to obtain a lot of information. Because we also intended to use CFR, intensive interviewing allowed us to establish a rapport with the biologists, learn the rules of the biology laboratory, and get an idea of what we would observe. The interviews were conducted at CSI and at Intel Research Seattle. For the interviews conducted at Intel Research Seattle, a floor plan of the biology laboratory and dozens of photos of the laboratory, tools, and equipment were available for reference purposes. Before we finished the interviews, we started CFR.

Notes were taken to capture data for both studies; in addition, audio recordings were made of the interviews, and still photographs were taken during the CFR. Most of the results discussed below came from a combination of the two techniques.

5.1 Results That Influenced Labscape's Interaction Model and Form Factor

The intensive interviewing and CFR helped us learn many things that impacted the design of Labscape's interaction model and form factor.

Upon entering the laboratory, we noticed that the benches were cluttered with tools and equipment, leaving the biologists little room to do their work. This suggested that anything we added to the environment could not occupy much space.

Contributing to the clutter were temporary waste bins and paper posted on the shelves above the benches. We learned that the waste bins were located throughout the laboratory to minimize the number of movements the biologists have to make, as they often dispose of things mid-task. We also learned that the information the biologists need to reference is often not where it is needed. This is largely due to the fact that only information everyone needs can be placed in the shared space; information needed by only one biologist has to be kept in their personal space. These observations suggested that Labscape's design should reduce, or at least not increase, the amount of required movement, and that this goal could be achieved in part by providing information where it is needed.

The biologists frequently wear latex gloves while performing experiments in the laboratory. Though the gloves protect the wearer, they can also spread contaminants. For this reason, biologists must remove their gloves when handling objects that might also be handled by others not wearing gloves. For example, we observed that the biologists removed their gloves while using an imaging workstation that also serves as a general-purpose computer. Often, biologists would only remove one glove to use the keyboard and mouse, as removing gloves in the middle of an experiment can be a nuisance. This suggested that we would need an interaction model that could be used by both gloved and bare hands, without creating contamination problems. Had we only interviewed the biologists and not conducted CFR, we would not have learned about the nuisance factor of removing gloves.

While performing tasks, the biologists remained very focused. They told us that this was largely because they often work from memory for reasons of convenience and to avoid contamination; distractions could cause them to make mistakes. This suggested that Labscape would have to be conveniently located and not attention-demanding. It also suggested that we might be able to remove some of the cognitive load from the biologists if they could rely on Labscape instead of memory.

Perhaps the most important conclusion from this round of user studies was the decision to put the computing in the environment, rather than on the biologist. Though space was a limitation, we came to this conclusion for several reasons. The biologists frequently move around the laboratory carrying objects with one or both hands. Therefore, we could not require them to use a handheld device. Although interviews told us that lab coats were normally worn, through CFR we learned that this is not necessarily the case—it depends on the samples and reagents that are being handled. Based on interviewing alone, we might have decided on a design that would require the biologists to carry a computing device in the pocket of their lab coat; however, thanks to CFR, we learned we could not assume the use of a lab coat.

Another factor that contributed to the decision to put the computing in the environment was the "wearable computer" the biologists already have—a digital timer with a clip. When a biologist is waiting for a specified period of time to pass (for example, samples might need to be incubated for 75 minutes), he sets the timer and changes tasks. In theory, he is supposed to clip the timer to his clothing so that no matter where he is when the timer goes off, he can hear it. Though we consistently saw the

biologists set their timers, more often than not, they left the timer at the station rather than clipping it to their clothing. It was clear that we could not rely on the biologists to wear something for Labscape.

5.2 Results That Influenced Labscape's Functionality

Intensive interviewing and CFR also gave us results that contributed to Labscape's functionality. We noticed that when a biologist was expecting to hear a beep from a piece of equipment, he would respond to it quickly. If he was not expecting a beep, he ignored it. This suggested that we would have to be careful about how we handled alerts. If we chose to alert the biologist by using a beep, we would probably either have to have him consciously set the alert so that he would listen for it, or we would have to personalize the sound.

We also noticed that the biologists do a lot of multi-tasking. This is often because they are waiting for a piece of equipment; for example, they may have to wait for an incubator to heat to the correct temperature. To use their time efficiently, they will often start work on something else during these waiting periods. They often leave their new task to check on the status of the equipment. This suggested that Labscape would have to be able to switch between tasks and experiments easily. It also suggested that we might be able to reduce movement and distraction in the laboratory if we could either give biologists the status of the equipment in which they are interested, or alert them when it is ready.

5.3 Summary

The combination of intensive interviewing and CFR helped us design an application appropriate for the biologists based on how they work (see Table 1 for a list of observations and their design ramifications). If we had not conducted user studies in the design stage, we probably would have built something that the biologists would have rejected. None of this would have been learned had we not involved representative users. Much of this would not have been learned if we had restricted ourselves to interviews and not observed users working in their environment on their usual tasks.

Consistent with many of the observations discussed above, the current implementation of Labscape relies on shared touch tablet computers that are distributed throughout the environment. These devices are used to display a flow graph representation of procedural plans and records when and where they are needed. The flow graph representation also provides a structure for capturing and organizing data that is produced during laboratory work. Details on Labscape's implementation, including design and functionality, can be found in other publications [4, 5]. In the next section, we describe the evaluation technique that we used to assess laboratory work before and after the installation of this system.

Table 1. Table of observations from user studies and what they implied in the design of Labscape

Observations	Ramifications for Labscape…	Found by…
• clutter	• limited space for additions to environment: e.g., no CRT monitors	• int. interviewing & CFR
• waste bins on benches • info on shelves & walls	• cannot increase # of movements • reduce movements by supplying info where needed	• int. interviewing & CFR
• wearing latex gloves • removing glove to use shared computer	• must be capable of being manipulated with gloved and bare hands • consider contamination issues	• int. interviewing & CFR
• not easily distracted	• must be conveniently located • cannot be attention-demanding • reduce cognitive load by guiding the biologists through the experiment	• int. interviewing & CFR
• move around frequently • hands often full when moving • seldom wear lab coats • seldom wear clip-on timers	• put computing in environment, not on user: e.g., don't require PDAs	• int. interviewing & CFR
• ignore unexpected beeps	• handle alerts carefully	• CFR
• multi-tasking	• switch between tasks & experiments • reduce movements by providing equipment status	• CFR

6 Evaluating Labscape

In this section we discuss the quantitative evaluation of a ubicomp environment with respect to aspects we feel are at the core of Weiser's vision: how physical activity relates to the use and creation of information. To do this, we chose metrics such as number of movements in the laboratory, how information is recorded, and the interleaving of Labscape use with physical work. Results of these metrics are discussed in Section 6.4. We also explain why traditional usability testing is not the best solution and how we used lag sequential analysis (LSA) for Labscape's initial evaluation.

For this evaluation, we chose a technique that would allow us to gather a lot of information from a small number of users, rather than a little information from a large number. The application of LSA we discuss was conducted at the CSI laboratory; two biologists participated in the study. In all sessions, the biologists performed their normal activities—no contrived tasks were used. Approximately 18 hours of video comprised of ten biology experiments was recorded and coded using LSA; five of the experiments were conducted as the biologists normally worked; five experiments were conducted while the biologists used Labscape.

6.1 Traditional Usability Testing Isn't the Solution

As discussed in Section 3, traditional usability testing involves observing users as they perform contrived tasks on the application being evaluated. Regarding evaluating ubicomp environments, Abowd et al. [2, 3] have pointed out that "it is not at all clear how to apply task-centric evaluation techniques to informal everyday situations," and that controlled studies in usability laboratories cannot lead to deep, empirical evaluation results: what is needed is real use in an authentic setting. We agree; in addition, even when the user's tasks are well understood, traditional usability testing is not the best solution for evaluating ubicomp applications.

Thanks to intensive interviewing and CFR, we had a good understanding of the biologists' tasks. We also knew from experience that although usability studies based on contrived tasks can provide easy to analyze data and expose problems with the application, they frequently fail to expose more serious problems that might occur in situations of authentic use. These unexposed problems could lead to the failure of the application. Contrived situations and artificial environments are not good enough for evaluating ubicomp: the applications are closely tied to physical movement and must work under a wide variety of conditions.

6.2 Application of Lag Sequential Analysis

Weiser suggested that to evaluate ubicomp, we need to work with people in disciplines such as psychology and anthropology [22]. We consulted a developmental psychologist who helped us choose LSA. The decision to use LSA was based on the type of metrics we wanted to collect, our desire to analyze the data for sequential correlations between observed events, and our need to balance the quality of data with the extent of the coding effort.

In lag-based data collection, an observation period is broken into a sequence of sampling intervals called "lags." For time-sampling methods, each lag represents a fixed period of time; for event-sampling methods, each lag represents the duration of an event. When an event of interest occurs in a lag, that event gets a "yes" for the lag. Event duration and the number of occurrences of an event in a lag may be noted depending on available coding resources and analysis needs. Sampling and analysis variations for lag-based data are discussed in detail by Sackett and Osofsky [15, 17].

To start LSA, evaluators must choose the events of interest. For our initial evaluation of Labscape, we identified 23 event types in categories such as information reference and recording, movements, body positions, and physical work with samples and laboratory tools. We chose events that would apply equally well before and after the introduction of new technology to the environment to give us a way of comparing subsequent iterations to a baseline. We used one-minute lags to keep the coding effort manageable while still providing detailed data. For events that tended to be short in duration and high in frequency, such as movements, we counted the number of occurrences in each lag, rather than reducing them to a binary value. The extra coding effort provided a significant increase in the usefulness of our data, as it helped to create a

more realistic picture of work in the biology laboratory. We did not record duration of any events, as we did not think the additional data would be worth the coding effort. Thus, our data does not currently distinguish between one long event and many short events of certain types, such as the events in the physical work category. If these ambiguities create problems in further data analysis, we can re-code the video.

By including physical location in the laboratory as one of our event categories, we were able to visualize laboratory activity using a map. To do this, we used a floor plan to show us details such as if the biologist was using a specific piece of equipment or if they were getting something from a shelf, cabinet, or refrigerator. We coded the maps to correspond to lags; for example, if the biologist made five movements in lag 27, the map would be marked with 27a, 27b,…27e, noting the biologist's location and sequence of movements. The combination of the event lags and accompanying maps gave us an accurate representation of work in the biology laboratory.

Fig. 2 below shows location and path data for one biologist during 60 minutes of a typical session. Fig. 3 contains some raw lag data collected from the same 60 minutes shown in Fig. 2. The data shows how many times the biologist accessed information, which of the major tasks of the experiment he was performing during each lag, if he was at a laboratory bench, desk, or elsewhere, whether he was sitting or standing, and whether he was empty-handed or carrying something when he moved.

Our data (as shown in Figs. 2 and 3) confirms some of the observations we made during CFR—the biologists move around frequently, at least one of their hands is full for more than half of their movements, and they multi-task. We discuss additional results in Section 6.4.

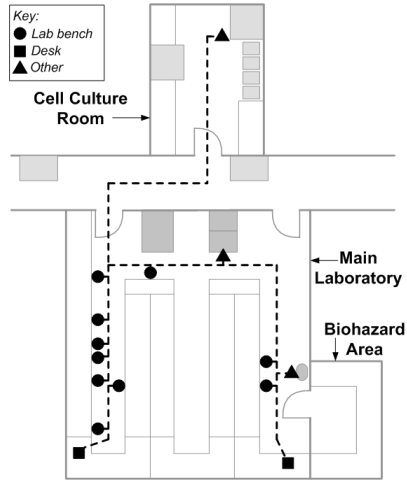

Fig. 2. Map showing a biologist's location and paths for 60 minutes of a typical session. This map does not show the complexity of movement throughout the laboratory; during this 60 minutes, the biologist changed location 76 times

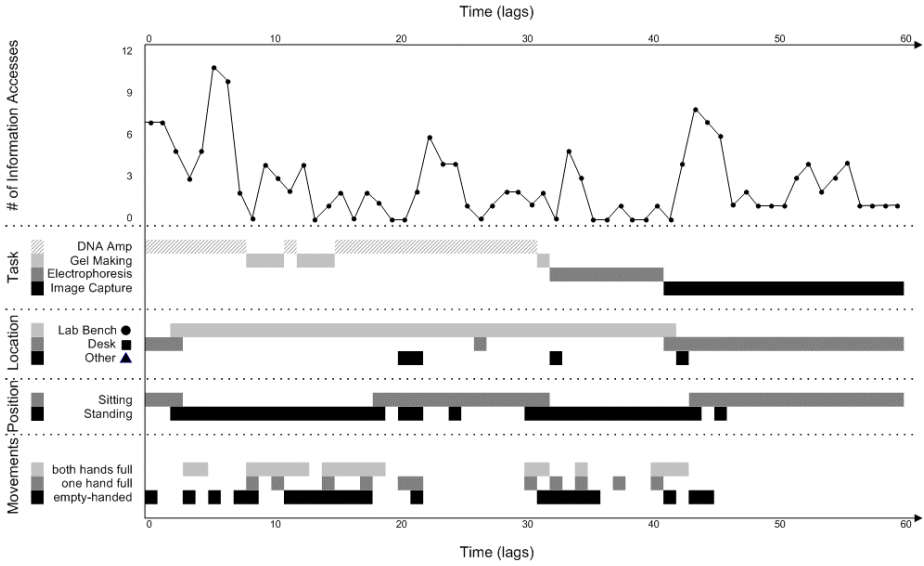

Fig. 3. Sample of the raw lag data collected for the same 60 minutes shown in Fig. 2. This chart shows the number of times the biologist accessed information per lag, tasks he performed, his location, position, and movements, including whether or not his hands were full

6.3 Advantages and Disadvantages of LSA for Ubicomp

Perhaps the biggest advantage of LSA is that it provides a way to measure the effect of ubicomp in an authentic setting. It can be used to establish a baseline of the environment before the ubicomp application is introduced, and against which future iterations can be compared. LSA can be performed with as few as one evaluator, though we recommend two to four; too many evaluators greatly increases overhead, as training of the video recording and coding personnel becomes an issue. Another advantage is for the users—because they are being observed during real use, they are able to do their regular work while the evaluators concurrently evaluate the application.

We captured data for the same users conducting the same type of experiments before and after the introduction of Labscape. However, a disadvantage of our application of LSA is that variables in each experiment made the data more difficult to analyze. The variables included changes in the number of samples used for each experiment, the time of day the experiment was being conducted, other events of the day, and the biologists' upcoming agendas. Based on their agendas, we saw different types of multi-tasking. Differences in the number of samples changed the duration of several tasks, which also created a shift in when and what multi-tasking occurred. The time of day the experiment was conducted and other events of the day sometimes changed the availability of the equipment. These variables made it difficult to perform direct comparisons of the data based on time. For example, we could not compare lags 20-60 of each experiment and get useful aggregate data, as those lags did not

necessarily represent the same type of work. However, obtaining more tractable data would mean asking the biologist to do something contrived: we would lose the benefit of real use. Instead, we have begun to work with statisticians to understand more ways in which we can analyze our data. One thing we can do is aggregate the data by task and not time; examples of this appear in Section 6.4.

Another disadvantage is that LSA is expensive. The most significant contributor to the cost was the video coding. Given the number of events of interest and the dynamic nature of the laboratory, we felt video coding was the right solution for us. Our initial attempts at live coding were not successful: none of the coders' lag sheets matched, the map was missing a huge amount of data, and having multiple coders in the laboratory got in the biologist's way.

Though video coding significantly adds to the cost, it makes coding easier. The advantages include the ability to review what happened, use of a remote control for pause, rewind, fast-forward, and slow motion, and being able to train coders without having to bother users. Tapes can also be used for other purposes; for example, anyone new to the Labscape project can review the tapes without bothering the biologists.

6.4 Results of Lag Sequential Analysis Adaptation

Given that our primary goal was "first, do no harm," we wanted to ensure that Labscape was not changing or adding significant tasks to the biologists' natural workflows. Because we knew that the biologists were concerned with minimizing their number of movements, we needed to know if Labscape was causing the number to increase. Our data shows a slight reduction in the average number of movements when the biologists used Labscape (see Fig. 4).

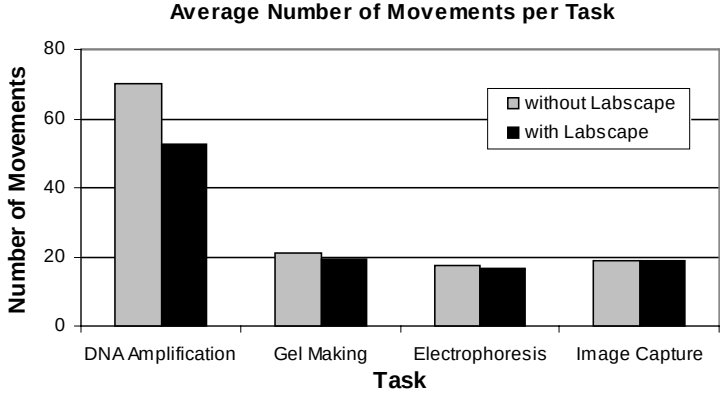

Fig. 4. Average number of movements the biologists made per task over the 10 recorded experiments. This enabled us to confirm that Labscape did not add to the number of movements

As mentioned in Section 5, biologists often rely on memory to keep track of where they are in an experiment. We hoped that Labscape could help reduce some of the

cognitive load by providing information where it was needed and capturing information as it happened. Our belief was that if we saw an increase in the number of times the biologists recorded information using Labscape, they would be relying less on memory. Fig. 5 below shows that with Labscape, the biologists voluntarily recorded information more frequently during each task, while Fig. 4 shows that their number of movements remained the same or decreased slightly.

We believe that in addition to reducing cognitive load by recording information as it happens, the biologists will also have better records of their experiments. These benefits were confirmed during the post evaluation interviews with the biologists. They verified that the ability to easily record the progress of a procedure allowed them to more comfortably switch tasks by reducing their need to rely on memory. They added that the record Labscape created was at least as thorough as any of the records they created before using Labscape.

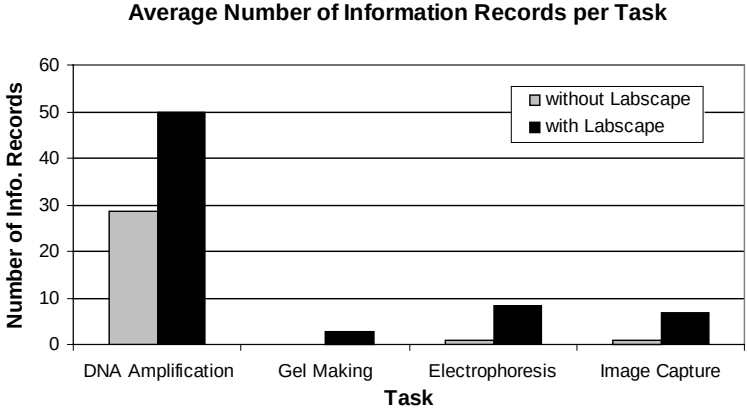

Fig. 5. Average number of times the biologists recorded information per task over the 10 recorded experiments. This enabled us to confirm that information was recorded more frequently with Labscape

Because Labscape's goals include maintaining the natural workflow of the environment and keeping biologists focused on their work, we needed to see how fluidly Labscape fit into their tasks. One metric of fluidity is to see how interleaved voluntary use of Labscape is with physical work. The more lags that contain both Labscape use and physical work, the more likely it is that Labscape is being smoothly integrated into the task environment. The results of this metric are shown in Table 2.

Table 2. Percentage of lags by task over the five recorded experiments conducted with Labscape involving use of Labscape and physical work

% of lags with interleaving	Task
57%	DNA Amplification
39%	Gel Making
42%	Electrophoresis
35%	Image Capture lags

Fig. 6 shows an example of this interleaving over 60 minutes of a typical session with Labscape. Physical work is represented by use of biological materials, tools, and equipment. Based on the results in Figs. 4, 5, and 6, we believe that Labscape has been successfully integrated into the biologists' natural workflows.

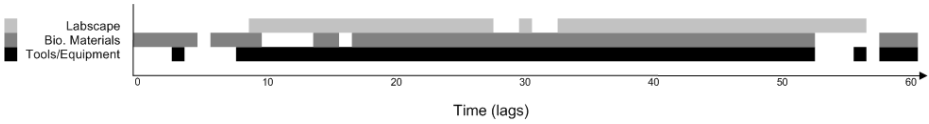

Fig. 6. Use per lag of Labscape, Biological Materials, and Tools/Equipment for 60 minutes of a typical session with Labscape

6.5 Summary of Evaluation

By using LSA for Labscape's initial evaluation, we were able to determine that we met our goal of "first, do no harm." We were also able to establish a baseline of work in the biology laboratory prior to the introduction of Labscape. Our results confirmed that Labscape appears to be successfully blending into the biologists' natural workflows. More analysis of the data is yet to be done.

As expected, the LSA study was expensive. Five people coded the 18 hours of video. Approximately 40.5 hours were spent training the coders and an additional 85.5 hours were spent coding the video.

7 Discussion / Future Work

We have just begun to analyze the data from our initial LSA study of Labscape; more analysis is forthcoming. We are encouraged by the type of data that we have collected; we hope the ubicomp community will join the effort of exploring LSA and share their results with us. If the community finds LSA useful, tools could be developed or adapted to help reduce the cost of analysis, e.g., by automatically collecting and analyzing some of the behavior and location data. Sanderson and Hilbert [10, 18] present surveys of tools and technologies for sequential analysis that have been applied in usability studies on automatically collected user interface event data.

We have recently started to work with another cell biology research laboratory and will be conducting interviews and CFR there; we are trying to learn how typical the behaviors and environment we learned about at CSI are of the general population of cell biology researchers. Labscape will soon be installed at that laboratory.

To help guide a graphic design student at the University of Washington with an upcoming redesign of Labscape's UI, a survey was recently sent to biologists from laboratories across the country; with that survey, we are hoping to validate our assumptions about Labscape's potential users. We also want to learn about their work habits and the types of ubicomp technology with which they are already familiar, e.g., use of

cell phones, digital cameras, PDAs, etc. In addition, we recently completed a heuristic evaluation of the Labscape user interface.

As we learn more about biologists' needs, we will continue to make incremental refinements to Labscape and evaluate them.

8 Conclusions

We have taken a step in the direction of establishing a principled approach for evaluating ubicomp applications. Our approach strives to obtain a combination of qualitative and quantitative data from real use in an authentic setting. Establishment of a baseline of the environment prior to the introduction of any ubicomp technology is paramount to this approach. The baseline provides critical data against which to compare the application's effect on the environment.

We have also shown the importance of applying well-established user study techniques to the design of ubicomp applications, while involving representative users and observing them in their natural environment.

Acknowledgements. The authors wish to thank Jeff Towle, Eithon Cadag, Jeong Kim, and Lenny Lim for the hours they spent filming, coding, and compiling data. We would also like to thank Richard Beckwith and Ken Anderson for their guidance with lag sequential analysis and Andrew Black, Gaetano Borriello, Neil Fanger, Ken Fishkin, Steve Gribble, Chia-yang Hung, Anthony LaMarca, Matt Lease, Michael Look, Bill Schilit, Jing Su, and Qinghong Zhou for their help.

References

1. Abowd, G.D.: "Classroom 2000: An experiment with the instrumentation of a living educational environment." *IBM Systems Journal*, Vol. 38 (1999)
2. Abowd, G.D., Mynatt, E.D.: "Charting Past, Present, and Future Research in Ubiquitous Computing." *ACM Transactions on Computer-Human Interaction*, Vol. 7. (Mar 2000) 29-58
3. Abowd, G.D., Mynatt, E.D., Rodden, T.: "The Human Experience." *IEEE Pervasive Computing*, Vol. 1 (Jan-Mar 2002) 48-57
4. Arnstein, L.F., Borriello, G., Consolvo, S., Franza, B.R., Hung, C., Su, J., Zhou, Q.H.: "Labscape: Design of a Smart Environment for the Cell Biology Laboratory." To appear in *IEEE Pervasive Computing*
5. Arnstein, L.F., Grimm, R., Hung, C., Kang, J.H., LaMarca, A., Sigurdsson, S., Su, J., Borriello, G.: "Systems Support for Ubiquitous Computing: A Case Study of two Implementations of Labscape." *Proceedings of the International Conference on Pervasive Computing*, Zurich, Springer Verlag (2002)

6. Burrell, J., Gay, G.: "E-graffiti: Evaluating Real-world Use of a Context-aware System." *Interacting with Computers*, Article 1228 (2002)
7. Cheverst, K., Davies, N., Mitchell, K., Friday, A.: "Experiences of Developing and Deploying a Context-Aware Tourist Guide: The GUIDE Project." *Proceedings of the 6th Annual Conference on Mobile Computing and Networking*, Boston (2000)
8. Cohen, J. "A coefficient of agreement for nominal scales." *Educational and Psychological Measurement*, Vol. 20 (1960) 37-46
9. Hackos, J.T., Redish, J.C.: *User and Task Analysis for Interface Design*. John Wiley & Sons, Inc., New York Chichester Weinheim Brisbane Singapore Toronto (1998)
10. Hilbert, D.M., Redmiles, D.F.: "Extracting Usability Information from User Interface Events." *ACM Computing Surveys (CSUR)*, Vol. 32, Issue 4. (Dec 2000)
11. Lysakowski: "Comparing Paper and Electronic Laboratory Notebooks, Parts I and II." *Scientific Computing and Automation Magazine* (March 1997 & May 1997)
12. Moran, T.P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., van Melle, W., Zellweger, P.: "'I'll Get That Off the Audio': A Case Study of Salvaging Multimedia Meeting Records." *CHI Conference Proceedings* (1997)
13. Myers, J.D., Fox-Dobbs, C., Laird, J., Le, D., Reich, D., Curtz, T.: "Electronic Laboratory Notebooks for Collaborative Research." *Proceedings of IEEE WET ICE*, Stanford, CA (1996)
14. Norman, D.A.: *The Design of Everyday Things*. Currency and Doubleday, New York London Toronto Sydney Auckland (1988) 12-17
15. Osofsky, J.D. (ed.): *Handbook of Infant Development*. John Wiley & Sons, Inc., New York Chichester Brisbane Toronto (1979)
16. Rubin, J.: *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons, Inc., New York Chichester Brisbane Toronto Singapore (1994)
17. Sackett, G.P. (ed.): *Observing Behavior, Vol. II: Data Collection and Analysis Methods*. University Park Press, Baltimore London Tokyo (1978)
18. Sanderson, P.M., Scott, J.J.P., Johnston, T., Mainzer, J., Watanabe, L.M., James, J.M.: "MacSHAPA and the Enterprise of Exploratory Sequential Data Analysis (ESDA)." *International Journal of Human-Computer Studies*, Vol. 41 (1994)
19. Scholtz, J., Herman, M., Laskowski, S., Smailagic, A., Siewiorek, D.: "Workshop on Evaluation Methodologies for Ubiquitous Computing." http://zing.ncsl.nist.gov/ubicomp01/, *UbiComp '01* (October 2001)
20. Schutt, R.K.: *Investigating the Social World*. 3rd ed. Pine Forge Press, California (2001)
21. Weiser, M.: "The Computer for the 21st Century." *Scientific American* (Sept 1991) 94-104
22. Weiser, M.: "Some Computer Science Issues in Ubiquitous Computing." *Communications of the ACM*, Vol. 36, Issue 7 (July 1993) 75-84
23. Wood, L.E.: "Semi-Structured Interviewing for User-Centered Design." *ACM Interactions*, Vol. 4, Issue 2 (Mar & Apr 1997) 48-61