A Framework for Evaluation of Ubicomp Applications

Mary Theofanos
National Institute of Standards and Technology
100 Bureau Drive, MS 8940
Gaithersburg, Md. 20899-8940
mary.theofanos@nist.gov

Jean Scholtz
National Institute of Standards and Technology
100 Bureau Drive, MS 8940
Gaithersburg, Md. 20899-8940
jean.scholtz@nist.gov

ABSTRACT
There is a clear need for evaluation methodologies that are suited to ubiquitous computing applications. Moreover, these methodologies must be able to evaluate the greater emphasis on values, emotion, privacy, trust and other social aspects that ubicomp applications exhibit. In this paper we examine the utility of our previously proposed user evaluation framework to specifically evaluate the social aspects of ubicomp applications. We test the criteria in our methodology by examining the utility and applicability of the framework to an existing commercial ubiquitous application. We conclude that the proposed framework does contain appropriate metrics to assess whether good design principles were achieved as well as identifying the social aspects and social implications of the application.

Author Keywords
Ubiquitous computing, usability evaluations, framework, metrics.

ACM Classification Keywords
H. Information Systems; H.5. Information interfaces and presentation (e.g., HCI) (I.7); H.5.2 User interfaces (D.2.2, H.1.2, I.3.6).

INTRODUCTION
Ubiquitous computing applications are diverse in nature and are very challenging to evaluate. Computing systems can only achieve Weiser's [1] vision of seamless cohabitation of the world by humans and computers if they are truly woven into everyday work and play. Such systems must go beyond the typical usability achieved by current computer systems. Today's systems are designed and evaluated based on the concept of tasks. That is, can individuals and groups use these systems to accomplish goals efficiently, effectively, and with reasonable satisfaction on the part of the users. But these systems are limited by constraints and assumptions about the user in addition to placing many requirements on the user. In order to achieve systems that are seamlessly integrated into everyday life, we need to understand how to design and evaluate these systems. This involves looking beyond the typical aspects of usability for tasks and considering the human experience. The creation of a common approach to evaluation will require a good deal of experimentation and measurement by the research community. A framework providing a consistent terminology and an initial set of metrics can provide an environment for researchers to share and learn from each other's evaluations. The benefits from such an approach will be validated metrics, effective discount evaluation techniques, and design guidelines which can then be put to use to improve the human computer interaction of these systems.

BACKGROUND
Typical usability evaluations of software today focus primarily on the three usability metrics (efficiency, effectiveness, and user satisfaction) [2] of the application. The majority of applications are also single user, desktop applications (excluding collaborative applications) and we tend to use laboratory evaluations where the context of the real world is not considered. Ubiquitous computing applications need evaluations and hence, metrics, that extend beyond the typical efficiency, effectiveness, and user satisfaction. These guidelines, techniques and metrics have proven very helpful in evaluating traditional desktop computing applications, but they are not sufficient for ubiquitous applications that place more of an emphasis on values, emotion, privacy, trust and other social aspects of computing.

The literature already contains examples of ubiquitous applications that have been less than successful or failed because developers have ignored the social implications. Consider the Boeing application with networked goggles that put diagrams or animations right in front of workers' eyes, eliminating the need to glance at handhelds or laptops while performing difficult tasks. Employees, didn't like wearing the goggles in a location where other co-workers could see them, therefore, the application was not used

despite the fact that it helped users produce excellent wire bundles. Salvador, Barile & Sherry [3] examined transactions in retail settings comparing a mock grocery store using typical check-out procedures (UPC codes) to a ubicomp system which senses items in a user's basket as well as a user's credit card. Using sensors, the system was able to automatically ring up and pay for the items simultaneously. The five participants in the study were uncomfortable using the system because they could not see and verify the transactions as they were occurring. This study raises several issues about trust, accountability, and invisibility that affect adoption and acceptance.

Guidelines, techniques and metrics to evaluate ubiquitous applications that place more of an emphasis on values, emotion, privacy, trust and other social aspects of computing must be developed. This can be facilitated by the establishment of a framework for evaluation of ubiquitous computing applications.

A FRAMEWORK FOR USER EVALUATIONS

Scholtz and Consolvo [4] have developed a framework for evaluating ubiquitous computing applications based on evaluations from traditional desktop computing, personal experience with ubiquitous computing evaluations, and literature reviews. This framework identifies a set of user evaluation areas with associated metrics and measures. Measures are defined as observable values. Associating meaning to those values by applying human judgment results in metrics. Table 1 identifies the nine user evaluation areas (UEA), the associated metrics and measures.

Traditional usability evaluations focus on users, but this framework also emphasizes stakeholders. As defined by Friedman et al. [5], direct stakeholders interact with the application and/or its output in a direct way, while indirect stakeholders are affected by the application in some meaningful way although not directly. Consider a cell phone, the direct stakeholder (DS) is the person who uses the cell phone and makes and receives calls from it. The indirect stakeholders of the cell phone include people who receive calls from the DS, people who call the DS, people with the DS when using the cell phone, people around the DS but not with the DS. Thus to use the framework, evaluators must identify both the direct stakeholders and the indirect stakeholders of the ubiquitous applications. Currently the framework focuses on metrics and measures for direct stakeholders. Metrics and measures for indirect stakeholders must also be addressed.

Many of the areas of evaluation take into consideration impacts outside of the application use itself. The framework was developed so that different applications could use the same vocabulary and hence, learn from each other. An issue in developing evaluations that extend beyond the application is developing the methods and metrics needed. It is essential that we try to develop formative evaluations. Instrumenting an entire parking garage is expensive. It will be disheartening to find that few people make use of the information due to the inaccuracy of the sensors.

APPLYING THE FRAMEWORK

Our objective is to apply the framework to a number of diverse ubicomp applications in order to assess the utility of the framework and the appropriateness of the evaluation areas. We recently completed a case study of a handheld restaurant order entry system that relies on handwriting recognition, mimicking the little green order pad that wirelessly transmits orders to the kitchen [6]. In this application the developers were very aware of the servers' needs both from a technology and social acceptance perspective. UEA's considered included:

- Attention: design strategies were used to provide a clear context for users who know exactly where to look and write on the screen at all times.
- Interaction: in addition to collecting data on *efficiency*, *effectiveness*, and *user satisfaction*, *distraction* is a critical factor for the servers. The primary task is to focus on and serve the customers and the technology cannot interfere. *Scalability* can also be a concern in larger restaurants –how many waiters can be supported at once.
- Impact and Side Effects: the restaurants that introduced the system have experienced measurable economic advantages in productivity, performance as well as quality. The restaurant's profits were increased due to fewer mistakes on the part of the servers. Servers were also able to sell more drinks because they were on the restaurant floor more of the time. Patrons ordered more desserts because the service was faster. The servers' tips were increased as service came more quickly. Moreover, fewer waiters are needed as individual waiters can service more tables.
- Adoption: a downside economically is that waiters have to learn the system and this takes some initiative. This also gives an advantage to people who are more technically competent. Additional *costs* include the cost of initial setup and the issue of maintaining the menus including daily specials.
- Conceptual Model: touch screen systems are popular in the restaurant industry today however; the handheld user interface is based on a different conceptual model.

UEA	Metric	Conceptual Measures
Attention	Focus	Number of times a user needs to change focus due to technology; number of different displays/actions a user needs to accomplish, or to check progress, of an interaction; number of events not noticed in an acceptable time
	Overhead	Percent of time a user spends switching foci; workload imposed on the user due to changing focus
Adoption	Rate	New users/unit of time; adoption rationale; technology usage statistics;
	Value	Change in productivity; perceived cost/benefit; continuity for user; amount of user sacrifice
	Cost	User willingness to purchase technology; typical time spent setting up and maintaining the technology
	Availability	Number of actual users from each target user group; technology supply source; categories of users in post-deployment
	Flexibility	Number of tasks user can accomplish that are not originally envisioned; user ability to modify as improvements and features are added
Trust	Privacy	Type of information user has to divulge to obtain value from application; availability of the user's information to other users of the system or third party
	Awareness	Ease of coordination with others in multi-users application; number of collisions with activities of others; user understanding about how recorded data is used; user understanding inferences that can be drawn about him or her by the application
	Control	Ability of users to manage how and by whom their data is used; types of recourse available to user in the event that the data is misused
Conceptual Models	Predictability of application behavior	Degree of match between user model and behavior of application
	Awareness of application capabilities	Degree of match between user's model and actual functionality of the application; degree of match between user's understanding of his or her responsibilities, system responsibilities, and the actual situation; degree to which user understands the application's boundary
	Vocabulary awareness	Degree of match between user's model and the syntax used by the application
Interaction	Effectiveness	Percentage of task completion
	Efficiency	Time to complete a task
	User Satisfaction	User rating of performing the task
	Distraction	Time taken from the primary task; degradation of performance of

		primary task; level of user frustration
	Interaction transparency	Effectiveness comparisons on different sets of I/O devices
	Scalability	Effectiveness of interactions with large numbers of users
	Collaborative interaction	Number of conflicts; percentage of conflicts resolved by the application; user feelings about conflicts and how they are resolved; user ability to recover from conflicts
Invisibility	Intelligibility	User's understanding of the system explanation
	Control	Effectiveness of interaction provided for user control of system initiative
	Accuracy	Match between the system's contextual model and the actual situation; appropriateness of action; match between the system action and the action the user would have requested
	Customization	Time to explicitly enter personalization information; time for the system to learn and adapt to the user's preferences
Impact and Side Effects	Utility	Changes in productivity or performance; changes in output quality
	Behavior changes	Type, frequency, and duration; willingness to modify behavior or tasks to use application; comfort ratings of wearable system components
	Social acceptance	Requirements placed on user outside of social norms; aesthetic ratings of system components
	Environment change	Type, frequency, and duration; user's willingness to modify his or her environment to accommodate system
Appeal	Fun	Enjoyment level when using the application; level of anticipation prior to using the application; sense of loss when the application is unavailable
	Aesthetics	Ratings of application look and feel
	Status	Pride in using and owning the application; peer pressure felt to use or own the application
Application Robustness	Robustness	Percentage of transient faults that were invisible to user
	Performance speed	Measures of time from user interaction to feedback for user
	Volatility	Measures of interruptions based on dynamic set of users, hardware, or software

Table 1: User Evaluation Areas (UEAs) for Ubiquitous Computing Applications

- Application Robustness: wireless coverage of the system was reviewed. Will the transmitters reach? What happens if the handheld reboots? Hardware issues were also examined including battery life and effective backlighting of the screen for evening shifts.
- Appeal: in this application it is difficult to separate social acceptance from appeal but the *aesthetics* should be considered. Does the device fit in different types of establishments? The servers indicated that the device was a conversation piece for many customers.

For this application the framework was a good fit and was appropriate in specifying the important evaluation areas.

A second example illustrates the social implications of impact and side effects. An airport near the authors uses sensors in the parking lot to provide the number of available spaces overall, the number per floor, and the number per aisle on each floor. One author has actually done a check of the indicated number in several aisles versus the actual number available. Given the error, she has modified her behavior to only choose aisles that have a large number of spaces available (if time is short) – a side effect that the developers probably did not consider.

CONCLUSION

These examples identify some of the areas that must be evaluated in order to design ubicomp systems that serve the public. We must move from application specific evaluations only to looking at social and economic issues.

These initial studies indicate that the framework does contain appropriate metrics to assess if good design principles were achieved and if the design will produce the desired user experience. Additional ubicomp applications should be studied to provide more feedback on the proposed framework in order to refine the framework and address its strengths and weaknesses and determine its utility in evaluating ubicomp applications and their social implications. The framework is a first step providing a structure so that key areas of evaluation are not overlooked and in identifying validated metrics and design guidelines which can then be put to use to improve the human computer interaction of ubiquitous systems.

But many questions must still be addressed. Does this initial framework capture the factors that influence the social aspects of ubiquitous applications? Is it complete, what's missing, is it useful? Are there any user evaluation areas that are missing? Does it provide metrics and measures that can differentiate systems? Can the framework be used to predict which systems will be useful and accepted by users? What are the interactions/correlations between the different user evaluation areas? Which evaluation areas are appropriate for which categories of ubiquitous computing applications? Are different evaluation areas applicable or have more weight depending on the category of ubiquitous computing applications? Can it, for instance, identify that invisibility of certain ubicomp applications deters adoption and acceptance and therefore it needs to be countered with visibility and accountability as seen by Salvador et al.?

REFERENCES

1. Weiser, M. "The Computer of the 21st Century," IEEE Pervasive Computing (1:1), January-March 2002, pp 19-25.
2. ISO 9241-11: 1998, Ergonomic requirements for office work with visual display terminals (VDTs) – Guidance on usability.
3. Salvador, T., Barile, S., & Sherry, J. (2004). Ubiquitous computing design principles: Supporting human-human and human-computer transactions. *Proceedings of the Conference on Human Factors in Computing Systems*, 1497-1500.
4. Scholtz, J., & Consolvo, S. (2004). Toward a framework for evaluating ubiquitous computing applications. *Pervasive Computing*, April-June, 82-88CHI.
5. Friedman, B., Kahn, Jr., P.H., & Borning, A. (2001). *Value Sensitive Design: Theory and Methods*, tech. report 02-12-01, University of Washington, Dec. 2001
6. Theofanos, M., & Scholtz, J. (2005). A Diner's Guide to Evaluating a Framework for Ubiquitous Computing Applications. To be published in *Proceedings of HCI. Hypertext 2001*, ACM Press (2001), 9-18.