# Analysis of the Evaluation of Application-Led Research in Pervasive Computing

Cormac Driver*, Eamonn Linehan† and Siobhán Clarke

Distributed Systems Group, Trinity College Dublin
{FirstName.LastName}@cs.tcd.ie

**Abstract.** Pervasive computing researchers typically conduct their research through the development of prototype applications. Such research is motivated by a well-defined problem and evaluated by assessing the impact of deployed solutions. Accordingly, the evaluation phase assumes a critically important role in this process. Failure to sufficiently evaluate an application can have wide ranging negative effects. Differences between the pervasive computing and standard desktop paradigms preclude the use of established evaluation techniques without significant modification. In this paper we present a survey of the state of the art in pervasive application evaluation. We discuss the prominent challenges in conducting a pervasive computing evaluation and assess how the surveyed application evaluations have been affected by these challenges. We make recommendations for researchers conducting hypothesis-led pervasive technology evaluations.

## 1 Introduction

A great deal of pervasive computing research is application-led, characterised by the development and evaluation of pervasive applications and systems. Researchers investigating well-defined problems use this approach to quickly deploy candidate solutions to test their theories by observing how users interact with applications and how applications interact with the environment. The evaluation phase assumes a critically important role in this process, determining how much can be learned about the worth of a deployed solution. Insufficient evaluation can have negative consequences ranging from leaving researchers unsure as to what users actually think about key aspects of their application, to causing researchers to follow research threads down paths that are eventually proven, by subsequent adequate evaluation, to be unrewarding.

We are motivated by difficulties experienced while evaluating a pervasive computing application built during the Hermes project [9], which is investigating the development of a software framework for pervasive applications. If the worth of a pervasive application cannot be accurately determined through evaluation, it follows that no conclusions can be drawn regarding the desirable features of a software framework. Difficulties in pervasive application evaluation are being experienced by the wider pervasive computing community, which has begun to address the issue of ap-

* Supported by the Irish Research Council for Science Engineering & Technology
† Supported by Intel Corporation

plication evaluation (see related work in section 2). However, it is not simply a matter of raising awareness about the need for evaluation. The community needs guidance in how to go about conducting meaningful evaluations of pervasive computing applications.

While desktop application evaluation techniques are well established, the same support does not yet exist for researchers working in the field of pervasive computing. It has been previously asserted that "The scaling dimensions that characterize ubicomp systems - device, space, people, or time - make it impossible to use traditional, contained usability laboratories." [1].

In this paper we present a survey of the state of the art in pervasive application evaluations. We explore the challenges that make pervasive computing evaluations more complicated than standard desktop application evaluation and investigate how these challenges have affected published evaluations. We also make recommendations for addressing the challenges identified. It is our hope that by identifying the areas in which previous evaluations have underperformed we will contribute to ensuring more comprehensive evaluations in the future. We also seek to motivate the need for significant research in the area of pervasive computing evaluation.

The remainder of this paper is organised as follows: Section 2 discusses related work. Section 3 presents data from our survey of the state of the art in pervasive application evaluation. Section 4 discusses challenges in conducting an evaluation of a pervasive computing application and illustrates, using survey results and our own experience, how these challenges have affected application evaluations. These discussions are followed by recommendations for addressing the challenges. Section 5 contains a summary.

## 2 Related Work

Scholtz and Consolvo proposed a framework for evaluation of ubiquitous applications [18], which has a set of "Ubicomp Evaluation Areas (UEA)" in which such applications can be evaluated. These areas describe requirements common to pervasive computing applications. Each UEA contains one or more metrics or measures that are intended to characterise how well the application performs in that evaluation area. The UEAs are overlapping in some cases but are comprehensive and provide a good reference for researchers carrying out their own evaluations. The authors do not discuss how the metrics can be compared across evaluations of different applications and no insight is provided into how data may be collected to populate the metrics. Our work highlights the practical challenges of the applying evaluation techniques that will enable results to be compared across evaluations.

Ranganathan et al. have also proposed metrics for the evaluation of different aspects of pervasive computing applications [16]. Their research goal was to form a benchmark by which pervasive applications could be compared. The metrics are grouped in three categories: System Metrics, Configurability & Programmability and Human Usability. An important contribution is an attempt to address the problem of ambiguity in the meaning of metrics, their units of measurement and their suitability. Our work builds on this by offering advice on metric adoption and on how evalua-

tions can be conducted to facilitate comparison with evaluations conducted by other researchers.

Consolvo et al. [7] have assessed the strengths and weaknesses of several qualitative and quantitative user study techniques for ubiquitous computing applications. The techniques were applied during an evaluation of the disruption to the user's natural workflow that is caused by the deployment of a ubiquitous computing application. The focus of this work is not on the development of metrics but rather on the application of information gathering techniques e.g., intensive interviewing and lag sequential analysis. Similar to our work, the authors acknowledge the importance of evaluation and the inappropriateness of traditional desktop-based evaluation techniques. The authors also share the ambition of gathering evaluation data from real use in authentic settings. This work differs to ours in that they discuss the merits of specific data gathering approaches as opposed to discussing higher-level challenges.

Sharp and Rehman have published a summary [19] of the 2005 UbiApp Workshop which was held at Pervasive 2005. The workshop concerned application-led pervasive computing research which is defined as the "design, implementation, deployment and evaluation" of pervasive applications. The published report relates closely to the topic of this paper in that the workshop featured significant discourse on the problems surrounding application evaluation. The report contains several key criticisms of the approaches to pervasive application evaluation currently being used. These criticisms are expanded upon here and are amongst the common challenges in pervasive computing application evaluations that we wish to highlight.

González et al. [12] have developed a ubicomp research methodology. Their paper describes the development of a ubicomp medical application and the development methodology they used. The methodology evolved from their experience developing applications in the area of pervasive computing. The authors acknowledge the fact that "the evaluation of ubiquitous computing environments in particular, has become an issue of considerable attention" and attribute this to the fact that field studies with actual users "require mature technology and often considerable investment in infrastructure which makes them impractical". The contribution this work makes is in stressing the importance of requirements gathering and in demonstrating how effective design can lead to successful deployment and a simpler evaluation.

## 3 Survey of Published Pervasive Application Evaluations

This section presents the results of a survey of 29 research papers, each of which discusses the evaluation of a pervasive computing application. The aim of this work is to assess the standard of pervasive application evaluations and by doing so identify the key challenges in evaluating prototype applications. All the application-led papers from the proceedings of two leading conferences in the field, UbiComp 2004 and Pervasive 2005 were considered in the survey. We also included a number of the most widely cited projects in the field.

### 3.1 Average Number of User Study Subjects

The number of subjects surveyed ranged from 4 to 700 with the median being 21.

### 3.2 Subject Demographics

We determined whether an attempt was made to choose subjects that were representative of the target audience for the application. For example, if an application designed to assist nurses in a hospital was evaluated using computer scientists as the subjects then we marked that evaluation as not having chosen subjects representative of the target audience. 50% of papers choose a subject group that was either representative of the target population or was based on a desire to have representative demographics within the group.

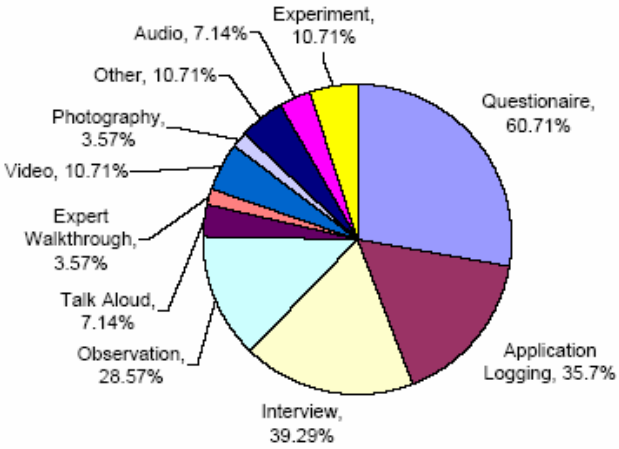


**Figure 1. Data collection techniques used in pervasive application evaluation**

### 3.3 Formative Evaluations

Formative evaluations are carried out to inform the application design phase. In papers where a pre-implementation evaluation of the design was conducted or an iterative evaluation approach was taken the paper was marked as having conducted a formative evaluation. 43% of the papers discussed conducting a formative evaluation.

### 3.4 Data Collection Techniques

On average, each paper used between two and three different data gathering techniques. Some used up to 5 methods but 31% of projects used just a single method. An

'other methods' category was included to cover methods such as presentations by users and lag sequential analysis which were used once each. This category also caters for the paper that explicitly stated 'other data' as a source of evaluation data. As can be seen in Figure 1, the most popular methods of gathering data for application evaluations are questionnaires, interviews, application logging and observation which were used in 60.71%, 39.29%, 35.71%, and 28.57% of the papers respectively. The remaining methods each appeared in between 3% and 10% of the surveyed papers.

### 3.5 Contrived Studies

We investigated how many of the papers had described a contrived study. We consider an evaluation contrived if it places subjects in a non-natural usage environment e.g., they know they are being observed. We found that 54% of papers describe an evaluation in which the usage environment was unrealistic. In contrast to these contrived evaluations, 36% of evaluations involved real world deployment. To be considered deployed we required an application to be used repeatedly without supervision in a natural usage environment for a significant period of time (one month or more).

### 3.6 Presentation Format

The majority of papers relied almost entirely on discussion (89%) to present their results. Statistical metrics appeared in 25% of papers and of these the number of subjects from which statistics were drawn was quite low, the lowest being 5 subjects. Other studies drew statistical conclusions from 11 and 15 subjects. Data was also presented as tables (17.86%), as charts/graphs (14.29%) and as raw data (3.57%).

### 3.7 Evaluation Objectives

In order to determine what was being evaluated we collected data on stated evaluation goals. Where the authors did not state the goals of their evaluation, the results of that evaluation were used to ascertain which aspects of the application were evaluated. Where the authors stated multiple goals for their evaluations the same project was recorded under each of the appropriate headings. These evaluation goals were analysed and it was discovered that they could be classed under five headings, described below along with the percentage of projects which performed this type of evaluation.

- *Usability/User Experience* (17.86%). Papers that used traditional usability heuristics [8] or aimed to gauge user satisfaction are counted in this category.
- *Distraction/Pervasiveness* (14.29%). This category includes papers which evaluated the amount of attention the application demanded of the user and whether this level of user interaction constituted a distraction.
- *Technology Validation/Performance Analysis* (46.43%). Evaluations which were purely for the purpose of demonstrating a successful and efficient implementation of application requirements are counted here.

- *Social Acceptance/Appeal* (35.71%). Projects that used their evaluations to assess how much users liked an application are counted in this category.
- *To Understand Strengths & Weaknesses* (28.57%). Evaluations aimed at collecting data to either improve an application or determine requirements for new applications are counted in this category.

### 3.8 Use of Control Groups

A control group is a group of subjects that will typically be asked to perform a task without the aid of the technology being evaluated. Comparing against the results gained from a control group gives researchers more insight into the real benefits of an application. Our survey revealed that 14% of evaluations choose to use a control group.

## 4 Challenges in Pervasive Application Evaluation

As stated by Weiser, "applications are the whole point of ubiquitous computing" [25]. The common pervasive computing research lifecycle follows a development, evaluation and publication pattern. The evaluation phase is often the process which determines the published contribution. To make verifiable and quantifiable advances it is necessary to conduct user-centered evaluation, the standard of which must be such that the results and lessons learned contribute to a better understanding of how the pervasive computing vision may be realised.

We have identified a typical evaluation lifecycle for pervasive applications which consists of the following steps: Identify Goals, Select Metrics, Select Evaluation Approach, Gather Data and Analyse Results. We discuss each step, identifying the related challenges and problems. We illustrate our points, where appropriate, with data from the application evaluation survey and our own experiences. We follow the discussion of each step with recommendations for improving on the current practice.

### 4.1 Identify Goals

Before gathering data via user study it is important to be clear on why this data is being gathered. Stasko et al. [22] were critical of researchers misusing data gathering techniques by not clearly identifying evaluation goals before collecting data. “Questionnaires are like any scientific experiment. One does not collect data and then see if they found something interesting. One forms a hypothesis and an experiment that will help prove or disprove the hypothesis.”. This is an observation which is particularly relevant to the field of pervasive application development. Many of the studies included in our survey did not state formulated evaluation goals before conducting their study. Once the data was gathered they then drew conclusions, depending on what the data suggested. Without well defined evaluation goals it is not possible to design a study that facilitates the answering of your research questions.

The failure of many studies to successfully identify evaluation goals may be a consequence of failing to find a problem before building the application. Our survey observed a tendency to develop applications without first exploring the problem space and understanding user requirements. Only 43% of surveyed projects conducted some form of formative evaluation or ethnographic study. Neglecting to design the application with user satisfaction as the principle requirement can cause certain badly designed features e.g., user interface, to prevent the collection of user study data, something which we experienced on the Hermes project [9]. Poorly designed applications that do not address a real problem cannot be deployed in the long term and therefore cannot be fully evaluated. The low deployment rate of 36% may be a result of this shortcoming.

Application-led researchers must decide when an application is ready to be evaluated. Before concluding the application development phase and entering the evaluation phase, application verification, validation and testing is required. This form of evaluation is necessitated by the fact that the technology used to develop pervasive applications is often quite novel and not fully understood. Technical challenges must be overcome before any insights into user acceptance can be gained, often necessitating an iterative cycle of testing and development. The most common goal when evaluating an application is validation of the application from a technological perspective, with 35% of projects solely investigating this aspect of their application. A significant proportion of evaluations are simply proofs of a working application as opposed to more meaningful assessments. Evaluation of the more interesting aspects of a pervasive application i.e., pervasiveness and user and social acceptance cannot be evaluated without a fully validated, stable application to give to user study subjects.

In a field with the long term ambition to realise Mark Weiser's vision [25] it is necessary to evaluate how an application contributes towards the realisation of that vision. Therefore applications must be evaluated for pervasiveness. It is apparent that pervasiveness is not being widely evaluated. This is borne out by the fact that only 14% of projects surveyed evaluated pervasiveness.

**Recommendations**

In the case of hypothesis-led research, evaluation goals should be formulated before the study is conducted to avoid recording large amounts of data without a clear purpose. For example, Bellotti et al. stated their goals as qualitatively analysing user acceptance in authentic use conditions, verifying their experimental framework and gathering information for future design [4]. With these goals in mind they designed questionnaires to gather the information required to meet the specified evaluation goals e.g., they asked questions on topics such as usability and enjoyability to assess user acceptance. They then conducted a non-contrived study to gather information in a real-world setting.

Before developing an application it is important that the user requirements are sufficiently understood. This avoids developing an application that has features e.g., a poorly designed user interaction model, which preclude the collection of data. Consolvo et al. have described the use of intensive interviewing and contextual field research to conduct a formative evaluation [7].

Applications must first undergo verification and validation testing before user evaluation. This allows researchers to assess the more meaningful aspects of the application, e.g. pervasiveness, during user trials.

Evaluation of an application's pervasiveness should be a goal of all researchers undertaking application-led pervasive computing research. Burrell et al. illustrate how they evaluated their application's pervasiveness by measuring user distraction [5].

### 4.2 Select Metrics for Success

The absence of a common vocabulary with which to discuss pervasive application evaluations is a barrier to the sharing of evaluation results. There has been some work in devising metrics (see related work in section 2), but these metric sets remain incomplete. In the absence of a specific metric, researchers are left with the challenge of determining their own. Different researchers choose different metrics for their evaluations, making comparative application analysis difficult. This gives rise to ambiguity regarding the meaning and quantification of metrics. Without a common structure for discussing evaluation practices we are limiting the amount that can be learned from each others evaluations, resulting in similar proofs being repeatedly demonstrated. In other fields where common metrics exist there are standard benchmarks which facilitate the comparative analysis of application evaluations.

**Recommendations**

The first step in developing a framework for exchanging and comparing evaluation results is to divide the evaluation task into distinct sub-tasks. Evaluation areas have been proposed by other papers (see related work) but to date none have been adopted. Given that the metrics now exist it is important that they are adopted by researchers. This will aid the comparison of published application evaluations.

In order to divide the evaluation task into suitable sub-tasks for which metrics can be formulated we can look to current evaluation practice as demonstrated by our survey. It is possible to classify all the evaluation areas used by papers in the survey under five headings.

1. Usability/User Experience
2. Distraction/Pervasiveness
3. Technology Validation/Performance Analysis
4. Social Acceptance/Appeal
5. Understanding Strengths and Weaknesses

Using such a classification of evaluation areas it is possible to identify which of these evaluation categories are targeted, then use a common set of metrics to examine how effectively an application performs as judged by the measures in this category. Advice should be offered on how data should be gathered to populate each relevant metric so as to avoid different methods resulting in figures that cannot be compared.

### 4.3 Select Evaluation Approach

Since pervasive applications are typically user centric systems it is usually necessary to perform a user evaluation. We have identified four common approaches that researchers may follow to evaluate an application.

1. *Deploy the Application* e.g., Place Lab [14], Guide [6]. Deliver the application to representative end users and allow them to use it in any way they feel appropriate.
2. *Build a 'Living Lab'* e.g., Aware Home [13]. Develop a physical, instrumented space into which your application can be deployed and monitored.
3. *Conduct Lab Experiments* e.g., [18], [23]. Use a traditional lab in which experiments can be conducted in a controlled, scientific manner. The evaluation environment is typically unrepresentative of the actual deployment environment.
4. *Use Limited Deployment User Studies* e.g., [21], [24]. Select a sample of users and deploy the application to them for a limited amount of time. The subjects are aware that they are participating in a user study and work with the researchers by providing data on their application usage experiences.

Understanding the full impact of an application involves fully understanding the environmental effects both on and of the technology. Assessing applications in real usage scenarios deployment is essential as applications are designed specifically for use in daily life settings and can only be accurately studied in this context.

There is a trade-off between performing a full deployment, which is very expensive to conduct, and performing a lab-based evaluation, affordable to most researchers. Lab-based evaluations are of much less value than full deployments and make understanding the real motivations behind application usage difficult to determine. Abowd et al. share the view that an application must be "… subjected to real and everyday use before it can be the subject of authentic evaluation." [3]. There are many examples of successful applications that have only revealed their true worth when used outside of the lab. SMS did not reveal its full potential during lab experiments and its very short messages and difficult input mechanism were considered a hindrance. However, once deployed it was quickly adopted by users.

There is a very high monetary and man-hour cost in deploying pervasive applications. Contributing factors to this cost include research and development effort, raising awareness of the application, performing training and supporting the application once deployed. The cost is further raised by the inherently interdisciplinary nature of the field, with evaluations requiring the involvement of experts from a wide range of fields. Building a living lab and carrying out partial deployments can reveal more than lab-based studies but can increase the probability of some form of bias affecting the study results.

The challenge in deploying applications is not one simply of cost but one related to the nature of academic research. Applications developed in a university research lab are often of non-industry quality and are built by small teams of disappearing students. Such applications, which must already contend with the issues effecting cutting edge technology research, do not suit wide-scale deployment. It has been previously

noted that “a good portion of reported ubicomp applications work remains at the level of demonstrational prototypes that are not designed to be robust.” [2].

Although there seems to be consensus in the field that the best way to evaluate an application is to deploy it [20], this is often impossible. Over half the studies we surveyed were carried out in a contrived manner. This situation arises when a study is con-ducted as a series of lab experiments and in some cases where only a limited deployment is conducted. It is a challenge to minimise the bias introduced due to conducting a contrived study without resorting to an expensive full deployment. The challenge is to perform an evaluation that is not biased by the nature of the experiment. It is also a challenge to determine what questions can be answered by such experiments and what issues can only be convincingly resolved through deployment.

In order to compare how a pervasive computing application has affected the environment into which it has been deployed, it important to record data about both states. For example, if you have developed an application to regulate an air conditioning system in an office then you must record data both before and after the system had been deployed. In this scenario the task is trivial. However a challenge exists when the pervasive computing application introduces behaviour that is not possible to directly replicate in a non-pervasive computing environment. Some effort must be made to compare however by using a control group who functions without the technology. Only in this way can the advances made by the technology be assessed. Only 14% of evaluations choose to use a control group in their evaluations. Of the papers that did feature a control group, none of them used the control group for performing usability, user experience or pervasiveness evaluations. These are the features which should ideally be evaluated by means of control group.

**Recommendation**

The evaluation process should begin with the identification of the evaluation goals and the evaluation approach. The chosen approach will have an effect on the aspects of the application which can be evaluated. For example, it has been widely recognised that evaluating an application designed for use in the ‘real world’ in a lab environment is of limited value. “… in the soft sciences, the requirement for a ‘controlled situation’ may actually work against the utility of the hypothesis in a more general situation. When the desire is to test a hypothesis that works ‘in general’, an experiment may have a great deal of internal validity, in the sense that it is valid in a highly controlled situation, while at the same time lack external validity when the results of the experiment are applied to a real world situation.” [26]. When selecting an evaluation approach the trade-offs must be understood and where possible we must strive to evaluate applications in authentic, real use settings. When deployment scale is restricted the selection of suitable subjects and the minimising of bias must be goals.

In addition to the evaluation approaches we have considered, there are also alternative forms of evaluation which may lower evaluation cost without sacrificing result quality. Wizard of Oz prototyping has been shown to be effective in evaluating ubiquitous computing application interfaces. Other components can be evaluated in isolation but little work exists on evaluating systems as a whole [15].

### 4.4 Gather Data

Our survey has highlighted the use of a variety of methods of data collection. Each of these has an associated cost and is suited to usage in specific situations. The challenge is to apply the relevant techniques in the most controlled, non-biased manner possible.

With 64% of projects choosing not to deploy as part of their application evaluation, it is possible that the Hawthorne Effect [27] is impacting on the majority of evaluations. The Hawthorne effect, first observed at the Hawthorne plant of the Western Electric Company in Cicero, Illinois between 1927 and 1932, describes a phenomenon which sees productivity increase regardless of the environmental factors manipulated. This is a short term effect which sees people become more productive due to being monitored, regardless of the modifications made to their environment. In a short-term, non-deployment user study this factor can significantly skew results. For this reason we believe that short-term non-deployment studies are unable to make strong claims about user reaction to applications.

As described in [10], the “wow factor” can affect the evaluation of pervasive computing applications. Users are impressed and intrigued by the novelty of pervasive technology, notably the hardware, and are prone to favourably receiving the technology. We experienced this phenomenon firsthand during the evaluation of a prototype application developed as part of our work on the Hermes project. The application was deployed on a PDA with a GPS device serially attached. We expected that this unwieldy approach would be negatively reviewed by users but the results of our hardware evaluation were quite the opposite. Believing that the “wow factor” was the sole reason for this we conducted a further study about PDAs, specifically about peoples long-term usage of these devices. Our hypothesis was that if we had deployed the application for longer subjects would have had different opinions regarding the suitability of PDAs for running our application. Our study of 60 subjects showed the number of people using PDAs every day dropped from over half of the sample to just over a quarter in the time they acquired the PDA to the present day. The number of subjects using the PDA about once a year or never went from 0% of the population to over a quarter. Over half the subjects said they would not replace their PDA if it were lost, stolen or irreparable. These results lead us to believe that our study was clouded by the “wow factor”. It is a challenge for researchers to minimise this first impression response and the effect this has on their results. Other forms of bias that may inadvertently be introduced at the data gathering stage include 1) ‘Mortality Bias’ - is there an attrition bias such that subjects later in the research process are no longer representative of the larger initial group? 2) ‘Evaluation Apprehension’ - have researchers taken suitable steps to mitigate the natural apprehension people have about evaluations of their activities, and to diminish the tendency subjects have to give answers which are designed to make themselves "look good"?

**Recommendation**
In order to minimise the effect of the “wow factor” researchers must take steps to quantify subjects’ familiarity with the relevant technology. The best way to minimise this effect on results is to fully deploy the application for a lengthy period of time. To date no pervasive computing studies have attempted to discover to what extent the “wow factor” may colour subjects’ experience of an application.

Deployment remains the best way to minimise the other forms of bias mentioned above such as mortality bias and evaluation apprehension. Consequently, deployment should be considered the only reliable way to conduct evaluations from which results can be directly published. All other evaluations require the bias effects to be analysed and commented upon. The crux of the problem is the general lack of scientific methods being applied. Pervasive computing has become a soft science which has been defined as "Any of the scientific disciplines in which rules or principles of evaluation are difficult to determine" [16]. The field is now at risk of becoming a pseudoscience which uses the language and trappings of scientific inquiry but is not based on any empirical method. There is a need for pervasive computing to adopt a more scientific method of research with papers being published that enable independent corroboration of results and evaluations that reduce the influence of individual or social bias on scientific findings. Skepticism within the field and the questioning of truth and reliability of the current user study-based evaluations are necessary to raise the standard of pervasive computing evaluations.

### 4.5 Analyse Results

Researchers are faced with the challenge of interpreting results. If clear evaluation goals have previously been determined and metrics for success have been identified then this phase should be relatively straight forward data analysis. However, a number of issues remain.

The form in which results will be published must be selected. 25% of the studied evaluations published results in the form of statistics. Of these, the average sample size used in calculation of statistics was 25. In addition, over half the studies were carried out in a contrived manner as illustrated in section 4.5. It is evident that the results of user studies are generally not statistically significant.

It remains a challenge when analysing results for publication to include enough detail on methodology to enable repeatable evaluations. This is necessary if we wish to allow other researchers to validate our findings and thus prevent pervasive computing from being relegated to the status of a pseudoscience. All the projects in our survey stated which techniques they used but that was the extent of the detail given. In this environment it is not possible to actually compare applications in specific areas. For example, two projects may claim that their prototype applications were favourably received by users with 68% and 59% of survey respondents finding the respective applications useful. Without knowing the factors that contributed to this value it is impossible to know exactly what useful means and how the two applications actually compare to each other in terms of utility.

**Recommendation**
When analysing the results of user studies, it is necessary to understand the statistical significance of results so as to not misrepresent findings. This is of particular concern in the field of pervasive computing where we have shown that user studies often involve very low numbers of subjects.

In order for evaluations to be meaningful they must be fully understood and comparable by those doing similar work. To further improve the believability of results

being published in the field it is necessary to have baseline results with which user study results can be compared. “To obtain a quantitative evaluation, it is necessary to compare the results from one method with those from another. If a single version of a device is being evaluated, it can be compared with a control,” [11]. We believe that control groups should be used whenever possible to allow researchers to more accurately determine results on the real benefit the application is to users.

Improvements in result analysis and the quality of resulting publications can be achieved by peer review evaluation. The peer review process has been very widely adopted by the scientific community but can often be too permissive. However, when reviewers insist on scientific methods this will generally improve the quality of the scientific literature.

It is a common practice in other fields for scientists to attempt to repeat the experiments in order to duplicate the results, thus further validating the hypothesis. To facilitate this, detailed records of experimental procedures should be maintained and published so as to provide evidence of the effectiveness and integrity of the procedure and assist in reproduction.

## 6 Summary

The ubiquitous computing field often values novelty, creativity and innovation over the need to have a clear hypothesis or set of goals. As a result, researchers tend to conduct evaluations in order to see what emerges in terms of changes in users’ behaviour. Although this approach has value and has often been fruitful in the past it must be followed up with hypothesis-led research that can verify and exploit conclusions inferred through these observational studies. It is in this way researchers can build upon each others work and deliver comprehensive evaluations.

In this paper we have highlighted the need for comprehensive evaluation of pervasive computing applications. We surveyed a sample of application-led pervasive computing papers and explored how they conducted their evaluations. This work exposed a number of deficiencies in the state of the art in pervasive application evaluations, which are symptomatic of larger challenges in the field.

Challenges were identified in the areas of selecting suitable applications, identifying evaluation goals, selecting metrics, selecting an evaluation approach, gathering the data and performing analysis. In response to these challenges several recommendations were made to address each of the identified challenges.

The contribution of this paper is to identify the challenges that exist in pervasive computing evaluations and to illustrate how these challenges have affected published evaluations. A clear lack of systematic application evaluation has been demonstrated by our survey. We have proposed ways in which to improve current evaluation practices based on principals of scientific research and hope to aid researchers in conducting more comprehensive evaluations in the future.

## References


1. Abowd, G., et al. "Charting past, present and future research in ubiquitous computing". ACM Transactions on Computer-Human Interaction, 2000. 7(1): p. 29-58.
2. Abowd, G., et al. "The Human Experience". Pervasive Computing, 2002. 1(1): p. 48-57.
3. Abowd, G. "Classroom 2000: An Experiment with the Instrumentation of a Living Educational Environment". IBM Systems Journal - Special Issue on Pervasive Computing, 1999. 38(4): p. 508-530.
4. Bellotti, F., et al., "User Testing a Hypermedia Tour Guide". IEEE Pervasive Computing, 2002. 1(2): p. 33-41.
5. Burrell, J., et al. "Context-Aware Computing: A Test Case". In 4th International Conference on Ubiquitous Computing. 2002. Göteborg, Sweden: Springer-Verlag.
6. Cheverst. K., et al. "Experiences of Developing and Deploying a Context-Aware Tourist Guide: The Lancaster Guide Project". In 6th Annual International Conference on Mobile Computing and Networking (Mobicom 00). 2000. New York: ACM Press.
7. Consolvo, S., et al. "User Study Techniques in the Design and Evaluation of a Ubi-comp Environment". In Forth International Conference on Ubiquitous Computing. 2002. Sweden. Springer-Verlag.
8. Doubleday, A., et al. "A Comparison of Usability Techniques for Evaluating Design". in Designing interactive Systems: Processes, Practices, Methods, and Techniques. 1997. Amsterdam.
9. Driver, C., et al. "A Framework for Mobile, Context-aware Trails-based Applications: Experiences with an Application-led Approach". In Workshop 1 ("What Makes for Good Application-led Research in Ubiquitous Computing?") Pervasive 2005, Munich.
10. Fleck, M., et al. "From Informing to Remembering: Ubiquitous Systems in Interactive Museums". IEEE Pervasive Computing, 2002. 1(2): p. 12-21.
11. Goodman, J., et al. "Using Field Experiments to Evaluate Mobile Guides". In HCI in Mobile Guides, workshop at Mobile HCI, 2004.
12. González, V., et al. "Towards a Methodology to Envision and Evaluate Ubiquitous Computing". In Workshop of Interacción Humano Computadora. 2004. Mexico.
13. Kidd, C.D., et al. "The Aware Home: A Living Laboratory for Ubiquitous Computing Research". In Second International Workshop on Cooperative Buildings (CoBuild'99). 1999.
14. LaMarca, A., et al. "Place Lab: Device Positioning Using Radio Beacons in the Wild". In Pervasive 2005. Munich, Germany.
15. Mäkelä, K., et al. "Evaluating the User Interface of a Ubiquitous Computing system Doorman". In Workshop on Evaluation Methodologies for Ubiquitous Computing 2001. Atlanta, Georgia.
16. Popper, K. "Unended Quest; An Intellectual Autobiography". September 1, 2002, London: Routledge.

17. Ranganathan, A., et al. "Towards a Pervasive Computing Benchmark". In Third IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOMW'05). 2005. Hawaii, USA.
18. Rohs, M., et al. "A Conceptual Framework for Camera Phone-based Interaction Techniques". In Pervasive 2005. Munich, Germany.
19. Scholtz, J. et al. "Towards a Discipline for Evaluating Ubiquitous Computing Applications". Intel Research Seattle, 2004,
20. Sharp, R., et al. "The 2005 UbiApp Workshop: What Makes Good Application-Led Research?". IEEE Pervasive Computing, 2005. 4(3): p. 80-82.
21. Smith, I., et al. "Social Disclosure of Place: From Location Technology to Communication Practices". In Pervasive 2005. Munich, Germany.
22. Stasko, J., et al. "Questionnaire Design". 2005. Available from: http://csdl.ics.hawaii.edu/techreports/05-06/doc/Stasko.html.
23. Suzuki, G., et al. "u-Photo: Interacting with Pervasive Services using Digital Still Images". In Pervasive 2005. Munich, Germany.
24. Wasinger, R., et al. "Integrating Intra and Extra Gestures into a Mobile and Multi-modal Shopping Assistant". In Pervasive 2005. Munich, Germany.
25. Weiser, M., "Some Computer Science Issues in Ubiquitous Computing". Communications of the ACM, 1993. 36(7): p. 75-84.
26. Wikipedia, "Experiment". September 1st, 2005. Available from: http://en.wikipedia.org/wiki/Experiment
27. Wikipedia "Hawthorne Effect". August 31st 2005. Available from: http://en.wikipedia.org/wiki/Hawthorne_effect.