Adapting Evaluation to Study Behaviour in Context

Scott Sherwood, Stuart Reeves, Julie Maitland, Alistair Morrison, Matthew Chalmers

Department of Computing Science, University of Glasgow,
Glasgow G12 8QQ, UK
{sherwood, stuartr, jules, morrisaj, matthew}@dcs.gla.ac.uk

Abstract. We present a reflection on a series of studies of ubiquitous computing systems in which the process of evaluation evolved over time to account for the increasing difficulties inherent in assessing systems 'in the wild'. Ubiquitous systems are typically designed to be embedded in users' everyday lives, however, without knowing the ways in which people will appropriate the systems for use, it is often infeasible to identify a predetermined set of evaluation criteria that will capture the process of integration and appropriation. Based on our experiences, which became successively more distributed in time and space, we suggest that evaluation should become *adaptive* in order to more effectively study the emergent uses of ubiquitous computing systems over time.

Introduction

When working with ubiquitous computing (Ubicomp) systems, challenges and rewards arise from moving from the relative safety of the usability lab into the uncontrolled environment of everyday life. For example, unpredicted contexts of use and environmental features such as intermittent network connectivity may challenge traditional evaluation methods, and yet we gain the mobility, contextuality and appropriation that let users take full advantage of new mobile devices. As Carter & Mankoff (2007) put it, "Ubicomp systems [are] more difficult to evaluate than desktop applications. This difficulty is due to issues like scale and a tendency to apply Ubicomp in ongoing, daily life settings unlike task and work oriented desktop systems." Many of these challenges have already been faced by researchers studying the *use* (rather than *usability*) of Ubicomp technologies in the wild. Observational techniques founded in ethnography may be well suited in principle but in practice are often hampered because keeping up users' activity is difficult. Small devices such as mobile phones and PDAs can easily be occluded from view, and people's use may be intimately related to and influenced by the activity of others far away (Crabtree et al., 2006).

In this paper, we reflect on our studies of three mobile multiplayer games: Treasure (Barkhuus et al., 2005), Feeding Yoshi (Bell et al., 2006) and Ego, and of two everyday awareness applications: Shakra (Maitland et al., 2006) and Connecto (Barkhuus et al., 2008). The development of these systems has spanned the last five years, with user experience design and evaluation techniques evolving over this time. We show a progression from early trials lasting around a quarter of an hour and taking place

within a specific confined area, to trials several weeks in length that explore users' integration of technology into their everyday lives. Studying system use over longer periods of time and in less constrained settings provides greater opportunity for witnessing unanticipated behaviour as users take ownership of the system, but can leave the evaluator more detached from the trial. Additionally, while many have studied the effects of uncertainty with regard to positioning accuracy and network connectivity on the user experience (e.g., Crabtree et al. (2004)), the impact these factors have on evaluators is not usually explicitly acknowledged.

Here we discuss the strategies that we as evaluators employed to discover participants' reactions towards and experiences of our five systems. The studies are presented chronologically, as the challenges faced in one study often influenced design and evaluation of subsequent systems. We suggest methods for keeping evaluators informed of activity during a trial that might take place over an extended period of time and over a wide geographical area, and suggest that such information is of crucial importance to adaptation of an ongoing evaluation based on evaluators' continual involvement with it or, in more extreme cases, immersion in it. Such adaptation may be done in order to inform and improve ongoing and post-hoc analysis. To conclude the paper we discuss the temporal and geographic scale of each study as contributing factors to the complexity of running such studies, and of gathering and interpreting evaluation data.

Related work

Researchers have examined how a particular design (along with external factors) can encourage adaptive behaviour, and how that behaviour can in turn inform the design process. Vougiazou et al. (2006) discuss a number of systems in which evolving cooperation or the dynamic and ongoing creation of authored game rules feature. It is notable that the evaluation of each system, for the most part, was based on direct observation and was short-term in duration. While the authors identify the challenges of long-term studies, no strategies are offered to overcome them. Iterative participatory design practices for developing the usability of mobile devices offer to some extent more 'agile' evaluation techniques (de Sá et al., 2008), and there are some existing demonstrations of remote, in situ data collection systems (Carter et al., 2007; Consolvo & Walker, 2003; Froehlich et al., 2007). Froehlich et al. (2007) have explored context–dependent 'experience sampling' systems that prompt the user for explanatory input when a mobile device detects that it is in a context of interest. Carter et al. (2007) recently developed Momento, which supports experience sampling, diary studies, capture of photos and sounds, and messaging from evaluators to participants. It uses SMS and MMS to send data between a participant's mobile device and an evaluator's desktop client.

There is also a growing body of work examining performance and game-based systems that examines the often rapidly adapting practices of both participants in and authors of an experience. Researchers have noted, for example, how players may develop an "emerging etiquette" in mobile games that take place over relatively long periods of time (Grant et al., 2007). This appropriation and adaptation by participants may also be mirrored by those running a given game or performance. Systems involving

more performative settings for player experiences have shown how those running – or 'orchestrating' – an experience, may sit in an evolving relationship with participants, such as in the SMS-based game Day of the Figurines (Crabtree et al., 2007) in which the game narrative came to be an ongoing negotiated production by orchestrators interacting with players. Other mobile city-scale experiences have focussed on the uncertainty inherent in using GPS and wifi systems, and how orchestrators' approaches to running the performance (i.e., their tactics and strategies) adapted over time, building up a working knowledge of how to manage that uncertainty (Crabtree et al., 2004). Such practices of orchestration also frequently involve distributed teams, and, of particular interest to this paper, extensive monitoring of participants, leading to intervention when necessary.

By and large, however, this literature generally identifies and examines adaptation only within the bounds of users' practices and experiences, thus lacking any explicit consideration of adaptation of the practices of evaluators. For instance, existing frameworks for evaluation of Ubicomp, such as that proposed by Scholtz & Consolvo (2004), might provide a toolbox of techniques, metrics and guidelines for evaluators, making existing practices easier, but little is mentioned regarding new adaptive or emergent approaches to an evaluation.

Thus, our concern in this paper is to address how adaptation within the experience may be complemented by evaluation techniques that are adapted and changed in response to user experience. It is not necessarily the case that adaptive evaluation is new in itself, but in this paper we offer some examples of techniques we used and tools we developed to help take this relatively unacknowledged approach.

Ubicomp trials within semi-controlled environments

In this section we discuss the evaluation of Treasure, a mobile multi-player game (Barkhuus et al., 2005). Each game comprised two teams, each with two players, who competed in games lasting around a quarter of an hour within a fixed game 'arena' of ~7000m^2 on the edge of the University of Glasgow's grounds. Each player used a GPS and 802.11-enabled handheld PDA that showed a map of the game arena. The object of the game was to walk to the locations of 'coins' that players saw scattered around their maps, and to move in and out of areas of network coverage to upload their collected coins to a server in exchange for points.

Teams were asked to come back on several different days in order to play against different opponents. This enabled the players to discuss the game, and develop and refine tactics both in and between game sessions. In initial pilot studies, this was found to be very important; players would often spent their first game learning how to use the technology and hence how to play the game at a basic level, whereas in subsequent games they often developed more complex strategies that suited their style of play as well as the setting. Without these multiple plays, we suggest that much of the developing competence with and appropriation of the system we observed would not have come to the fore. Multiple plays or

long-term use became imperative to a deeper understanding of system use, and we have tried to maintain this evaluation principle in all of the subsequent studies discussed in this paper.

Errors in positioning technology and patchy network coverage are usually considered to detract from a user experience. Treasure was designed to exploit these factors, changing them from problems into resources for the game. However, these factors did still prove to be problematic when it came to collecting data for evaluation. Unlike lab-based experimentation, the log data gathered was often unreliable in the sense that the recorded position did not necessarily represent the actual location of the player when an entry in the log was recorded. Inaccurate positioning and intermittent connectivity meant that it was possible for there to be several versions of the game state at any one time – one for each player and one for the server. In order to make sense of all of these different streams of data, the information had to be synchronised. This is normally a very labour-intensive task (Crabtree et al., 2006), complicated by the need to explore circumstances of play and interaction by synchronising events captured by multiple data sources with multiple, sometimes conflicting, states of the system at any one time. Such challenges inspired the design of Replayer (Morrison et al., 2006), an analytic tool that integrates and synchronises log data from multiple sources to allow quantitative and qualitative forms of exploratory data analysis, an issue also explored by Greenhalgh et al. (2007). Like most evaluation, Treasure's analysis was conducted retrospectively – after each game or set of games. However, due to the limited space within which the game was played, evaluators were able to directly observe the play. This meant that they could use their observations to tailor the questions posed during the interviews that followed each game, enabling them to prompt participants to elaborate on areas of the trial that seemed significant.

Evaluating a system used in everyday life

Feeding Yoshi was also designed to run on handheld devices and exploit 802.11 (Bell et al., 2006). Rather than connection to a single wifi access point (AP), it used the distribution of secure and unsecured APs as a resource for the game, with players collecting 'fruit' which grew in unsecured APs and 'feeding' it to Yoshis in the secure APs. Feeding Yoshi was designed to be played over a much wider area and over a much longer time period than Treasure, with the intention that users would have a chance to fit it into the contexts and routines of their everyday lives.

The trial participants consisted of two teams in Glasgow, one in Derby and one in Nottingham, with the study lasting a full week. Unlike Treasure, the evaluation put no constraints on where the game could be played – participants could play anywhere that wireless access points could be readily found, such as office blocks, cafes and suburban areas. As a result of this, and of our interest in discovering how players responded to the contingencies of the technology and of the everyday world in which they were playing, the approach taken to evaluating Treasure was infeasible here. The game was not run within a semi-controlled environment, meaning that the main constraint put on game play was players' existing circumstances of work, leisure and home life. Therefore, evaluators were only occasionally able to observe players, as they were often spread out over different cities and there was no guarantee when or

where they would play the game. Capturing video was similarly difficult; since a main research question was examining where and when users would choose to play, there was little point in arranging contrived meetings to video system use. In consideration of these challenges, the employed evaluation strategy focussed on system logs and post-trial interviews.

This greater detachment of the evaluator from the system use made the evaluation of Feeding Yoshi more challenging than that of Treasure. Although every participant was interviewed following the week's trial, the duration of the game was such that particular instances of play or players' motivations at specific times were often forgotten. Other aspects were deemed irrelevant by the players and therefore went unreported. As the evaluators were unable to directly follow the events of the trial, they were less able to focus questioning on specific topics that emerged from witnessed behaviour. Another issue arose when some players reported using other forms of communication technologies to discuss and encourage play with other team members. This had not been anticipated in advance and most of this information was inaccessible for post-trial analysis. If unable to directly observe the participants, and not directly immersed in the game themselves, evaluators might find it difficult to establish how use of the system is developing. Opportunities to observe or log unanticipated activity can be lost, and without such behaviour being mentioned by chance during interviews, its occurrence might go completely undetected. We subsequently tried to address these problems in the evaluation of later systems.

Dynamic questioning: FlexiFill

The previous sections have shown that the evaluation of Feeding Yoshi was performed reflectively without any direct observation of the play. In subsequent systems such as Shakra (Maitland et al., 2006) and Connecto (Barkhuus et al., 2008), we tried to overcome this problem by introducing new techniques that enabled more informed reflective evaluation. The goal was to find a technological means of providing evaluators with a greater degree of insight into user activity during a trial and to allow them to embark on interviews with a better understanding of events in order to tailor questioning to each individual participant's experience.

Shakra was a mobile phone-based application that analysed patterns of fluctuation in GSM cell signal strength to provide summaries of the amount of time a user spends walking. Users could view their daily activity levels in comparison with the accumulated activity totals of their friends, with the intention of making people more aware of their activity levels and thereby hopefully encouraging them to achieve the recommended 30 minutes per day. In the evaluation of Shakra users were given a usage diary – a printed form that they were asked to fill in and return at the end of the trial. Shakra was piloted with three groups of friends over the course of one week, the aim being to examine the impact the activity tracking and sharing of activity levels had on users' self-awareness and to discover whether this motivated any change in attitude towards physical activity.

The evaluation of Shakra attempted to address some of the problems experienced in evaluating Feeding Yoshi. The usage diaries were an attempt to overcome the issues of delayed reflection, with participants encouraged to document any significant happenings that occurred each day. The diary had 19 questions and was returned after the trial but before the participant was interviewed, so that evaluators could familiarise themselves with the individual's experience and tailor specific questions to draw out particular events. However, it became apparent during interviews after the trial that the players would spend less time on the (static) diaries on each successive day, stating that they felt they were repeating the same things day in and day out. Additionally, our evaluations were focussed on how user behaviour changes over time, and a static diary is not a tool adept at uncovering such information. While the diary could capture some very common issues that arise in such experiences, getting at the nuanced behaviour in a particular experience was much harder using this technique. The evaluation of a second awareness application called Connecto attempted to address these difficulties through the creation of a more dynamic diary tool.

Connecto is a mobile phone application that displays contextual information about friends and contacts. Building on from the previous usage diaries, a new tool called FlexiFill was designed to make daily enquires about system use via a more dynamic web-based interface. Within this diary-style interface, information logged during the trial was shown to the participant, as a reminder of who they had communicated with. This helped to prevent users forgetting interesting events, and acted as a prompt for them to recall their motivation and actions around the communication. In each day's entry, users were asked about a single, randomly selected phone call and text message sent that day. Since all information about location and communication was logged and sent to a central server via GPRS, it was simple to use this information in the questions. From previous experiences, we were aware that participants are reluctant to write thorough diaries by hand, especially if they have to do it daily and the questions are repetitive. The diary was therefore designed to be flexible; participants could fill it in at a time and place that suited them. In order to help give incentives for answering questions posed in the diary, players were presented with the FlexiFill interface before they could gain access to the game's website. Since the trial of Connecto lasted two weeks, we also conducted interviews both at the midway point and end of the trial. In preparation for these interviews we were able to use the participants' FlexiFill answers in order to tailor the interview questions to that particular participant.

Awareness, orchestration and evaluation

Ego, a game that makes use of both mobile and online play, is the most recent of the systems discussed in this paper. During mobile play, the system captures aspects of the player's everyday life that can be used to present a profile tailored to the interest of the audience viewing the profile. The aim of the game is to boost one's 'ego' by being seen as the most 'popular', 'well-travelled' and 'coolest' person. To achieve this, players could gain points in three different ways: 1) when the players' phones detected proximal Bluetooth devices they earned 'popularity' points; 2) when the phones detected wireless access points players gained 'well-travelled' points; 3) each day players were asked to vote for co-players who had done the most interesting things the previous day, gaining the three most voted–for

players more ‘coolness’ points. This logged information was then streamed to the server via GPRS, slowly developing the profiles of the players.

The game was trialled over a month with two groups of five, where each player knew everyone else in his or her group. During play, relationships evolved between individuals belonging to different groups as well as between those in the same group. An example of this involved a conflict between two players who fell out over comments made to a mutual friend (one of the group’s lecturers, with whom the group were friendly and socialised with, but who was not part of the game). When one player criticised this lecturer, another player was offended, and supported the lecturer. The offended player then used Ego to express this feeling, through continuously taking points from the other player. This type of retribution is seen in the “he said she said” encounters discussed in (Goodwin, 1980).

During the trial each player was interviewed halfway through the game and then again at the end of the month. However, the design of the Ego system also enabled the evaluators to unobtrusively examine player profiles on the website throughout the month and thereby observe what was happening in the game. This continual and ongoing awareness enabled the evaluators to identify when players were having technical problems or when interesting interactions took place, with both activities then shaping and informing subsequent as well as ongoing analysis. A feature of this continual connection is that it enables data to be gathered in a less intrusive way than direct observation or shadowing. Such direct techniques are often impractical in experiences that span a large area over a long duration. In Ego, the evaluators were able to passively involve themselves in players’ experiences without heavily impacting them, and yet gain understanding about interactions between players.

However, on some occasions, the evaluators were moved to make more active interventions during the trial. For example, the users would sometimes encounter technical difficulties, which is to be expected with a system running over such a large period of time and geographical distribution. In past trials it had been difficult for users to understand when the system was not working correctly, typically resulting in frustration and disengagement from the experience. However, through evaluators’ continual awareness, the identification of such problems was no longer the sole responsibility of the user. In Ego, by being able to observe the system and immerse themselves in the game as if they were playing, evaluators could remotely identify occasions when patterns of activity were potentially unusual and intervene accordingly. For example, one participant who had been an extremely active player had only managed a very low score one day, which did not fit with his usual pattern of play. He had not contacted the evaluators since he was in the final few days of the trial and he assumed he had done something to break the device. Through their continued awareness of the system and individuals’ play, evaluators were able to identify and fix this problem. In this case the database on the mobile client had been corrupted, but it was then possible to fix this within an hour of the problem being identified. This prevented this user’s experience from being cut short prematurely and therefore generated more data for analysis.

It is worth noting that one possible side-effect of the continual attention paid to user problems by evaluators is an increased engagement in the trial by those users, therefore opening up findings to the potential criticism of an increased likelihood of the Hawthorne effect. “Proponents of the Hawthorne effect say that people who are singled out for a study of any kind may improve their performance or behavior not because of any specific condition being tested, but simply because of all the attention they receive.” (Rice, 1982). Such a view seems to indicate that the degree of attention paid to those participating in a study is positively correlated with any subsequent Hawthorne effect – a commonly held assumption being that the no human-centred study is completely free from the Hawthorne effect (Macefield, 2007). However, the generalisability of the Hawthorne effect has recently been called into question (Rice, 1982; Macefield, 2007). Macefield (2007) presents a full discussion on the limitations of such a generalisation with respect to usability evaluations. Similarly, Crabtree & Rodden (2004) propose that the Hawthorne effect is often overestimated when considering ethnographic studies in the workplace and home, simply because when in these environments people “have better things to do than impress or worry about the ethnographer”.

The Ego evaluation also featured a new version of the FlexiFill tool. Rather than relying on a relatively broad sampling of user activity over the course of the trial as the focus of questions, evaluators used the improved FlexiFill which permitted the addition and tailoring of questions in order to make specific enquiries at any point, once again moving a further step away from post-hoc static diaries and questionnaires. Any observed actions that seemed interesting or puzzling could be put to the user the next time they logged in, without waiting for a post-trial interview, reducing the risk of player accounts becoming skewed over time or, worse, forgotten.

One subtle example of this revealed a complex, but somewhat misguided tactic employed by a player to hide his activity from those in his team. Throughout the game, players had the ability to hide or reveal certain aspects of their profile to others, and either give or take points to or from other players. Players’ use of these abilities was displayed on an events page accessible to other players, making them accountable for their actions. Figure 1 shows an extract from the events page below.

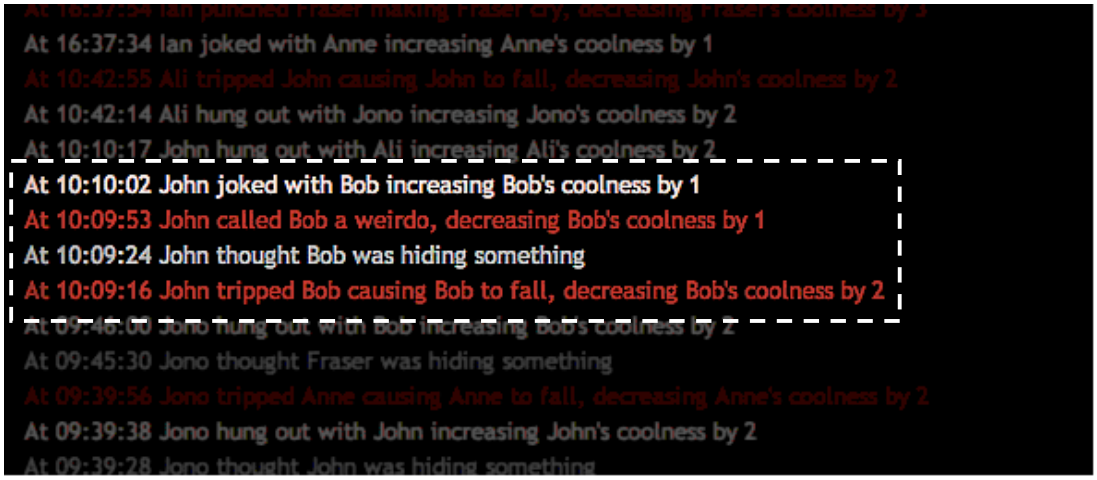

Figure 1: An extract from the events page of the Ego game's online component (events relevant to this paper are indicated by the dashed box). Players' actions during the game were made public to everyone else.

Figure 1 shows John being 'unfriendly' to Bob before he is then 'friendly' to him. Through their continued awareness of the game the evaluators noticed this pattern and felt it to be unusual. At first it was thought that the player did not fully understand the game mechanics. However, during the trial a question was added to FlexiFill specifically for this player, asking about this event. Explaining his actions in an interview, the player stated:

> Where it shows you on the side of what happens it goes in chronological order so the first thing [John] would have seen was that I joked with him and then he wouldn't have bothered to check the fact that I have went absolutely after him.

Although this reasoning was flawed, it reveals not only a tactic the player developed, but also how he viewed the events page: namely that it offered the potential to present a certain (in this case, incorrect) impression to others. Taken in the larger context of the game, this conduct led the evaluators to question the relationship between these two players who were, at face value, friends, which in turn led to insight into the dynamic behind the whole cohort of players as a social group. It is extremely difficult to say whether or not subtle events like this would have been uncovered and understood through the use of static diaries, post-trial interviews and reflective analysis. However, it is possible to say that with their continual awareness of the happenings within the game, evaluators were more attuned to the in-game events and therefore more likely to investigate what were at times quite nuanced interactions.

Discussion

As systems like Ego push further into the wild, away from controlled and constrained settings, the spatial distribution and temporal duration of users' experiences grows. Space and time play key parts in understanding the increased levels of uncertainty introduced by the evaluation of Ubicomp systems in

this way. Responding to this changing perspective directed us to exploit more agile evaluative techniques to tackle the rising uncertainty – i.e., the seemingly uncontrollable aspects of an experience – that naturally go hand-in-hand with embedding interactive systems into users' lives. It is only through increased awareness that we may manage and control uncertainty in evaluation procedures. Using time and space as a basis to characterise evaluated user experiences, we can distinguish between four rough categories and draw out distinctions between the kinds of evaluative techniques we found necessary for each. The table below summarises much of the prior discussion of our five systems, and serves to help with the subsequent discussion.

System	***Space***	***Time***	***Evaluator awareness and involvement***	***Data Collected***
Treasure	Small (Constrained)	Very short (minutes)	Direct observation of users' actions	Direct observation, interviews, video, system logs
Feeding Yoshi	Large (Unconstrained)	Short (days)	Indirect post-event / reflective methods	Interviews, system logs
Shakra	Large (Unconstrained)	Short (days)	Indirect / reflective	Interviews, usage diaries, system logs
Connecto	Large (Unconstrained)	Medium (weeks)	Indirect / reflective / adaptive	Interviews, semi-adaptive usage diaries, GPRS, system logs
Ego	Large (Unconstrained)	Long (Month)	Semi-direct / indirect / adaptive	Interviews, ongoing observation through GPRS uploading, system logs

Table 1: Summary of the evaluations of the discussed systems.

Treasure is an example of a system in which interaction is relatively *constrained with regard to both duration and space*. Although there were some difficulties in collecting data for evaluation, and tight experimental control was limited to some extent through the trial's openness to interruption by people outside the trial (such as car drivers passing through the area), Treasure's limited spatial and temporal extents permitted close and relatively comprehensive direct observation. On the other hand, broader issues of how Ubicomp technologies may be woven into everyday life are not addressed easily when time and space are so constrained.

User experiences of *longer duration but constrained space* tend towards more 'traditional' Ubicomp systems such as the Active Badge Location System (Want et al., 1992) and smart home environments (Demiris et al., 2007). Evaluations here have necessitated ongoing commitment to a static setting like a home or an office building. To a varying degree, the challenge for evaluators in studying these scenarios is in maintaining this continual involvement in the systems – rather than an intense, spatially

distributed involvement (see below). Thus the problem becomes one of determining when interesting interactions take place rather than where.

A *shorter duration but a less constrained space* suggests participants may move in a far less restricted way – with no set boundary. City-scale experiences such as Human Pac-Man (Cheok et al., 2003) and Uncle Roy All Around You (Benford et al., 2004) involved evaluating games that took place over a large yet flexible area, with player experiences lasting at most a few hours each. Direct observation was still possible due to the temporally focussed nature of player interactions. Although problems collecting this observational data are typically mitigated by the limited duration, such evaluation activity is often quite intensive due to monitoring or shadowing participants over such a large physical space.

Experiences involving fully *unconstrained spaces and long durations* perhaps present the greatest challenge analytically, even though systems like these have become increasingly common for researchers concerned with a key characteristic of Ubicomp – the appropriation of technology in everyday life. The player experience in games like MobiMissions (Grant et al., 2007) and Day of the Figurines (Crabtree et al., 2007) lasted a month or more, during which time players could roam wherever they pleased. Such systems are usually assessed without direct observation, instead exploiting mixed methods of interviews, questionnaires, usage diaries and so on. We note that such conventional, static evaluation techniques were employed in our earlier systems like Feeding Yoshi. Connecto and Ego followed a modified version of this approach due to the trial lengths and unconstrained spatial distributions of users, but, importantly, data streamed via GPRS enabled evaluators to view the moment-by-moment actions of the users and adapt both the systems' orchestration and their ongoing evaluations accordingly.

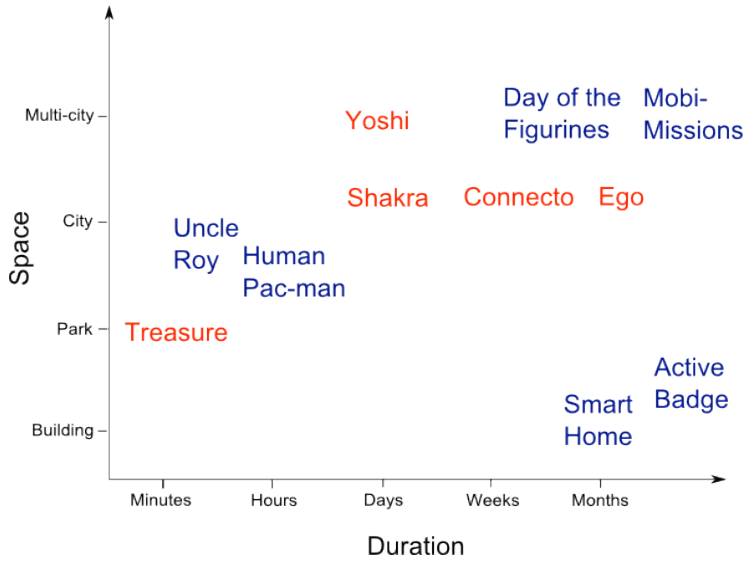

Figure 2: Categorisation of Ubicomp system evaluations by time and space.

Figure 2 attempts to summarise the systems mentioned above in terms of time and space. Difficulty in exercising experimental control increases as both geographic distribution ('space' in Figure 2) and

temporal duration increase. For Treasure, it tended to be easier to mitigate uncertainty about where and when interactions might occur due to 'park-sized' interactions occurring over minutes that could be covered by saturating the space with evaluators. Shakra and Yoshi introduced uncertainty over *where* interactions might happen due to 'city- or multi-city sized' interactions as well as extending data collection times from minutes into days. Finally, Connecto and Ego stretch the boundaries of evaluation to a greater degree by involving 'city-sized' interactions for weeks or perhaps months. Uncertainties over where and when interaction may occur is at its greatest in this region.

As we increase the temporal and geographic scale of user experiences, we gain a greater opportunity to see how ubiquitous computing may become embedded into everyday life, subject both to mundane routines of work and home, and to possibilities for serendipitous and opportunistic interaction. Through this we might explore more fully how competence, system appropriation and mastery, as well as strategies and tactics (particularly in the case of Ubicomp games) develop in use. As we have seen, however, this interest is in tension with our ability to evaluate such uses and environments. For summative methods, which as we note may be more practical in longer term trials of a greater spatial distribution, evaluators must assess much data collected in a post-hoc way, gleaning information from users during interviews and system logs. In contrast, ethnographic studies favour observation and rich description as a way of understanding system use. By engaging and immersing oneself in the experience, evaluators may be able to observe many of the more subtle interactions that take place. Such techniques have been key to understanding the nature of interaction either over a lengthy duration *or* wide physical space. We faced difficulties in adopting such ethnographic techniques on interaction both unconstrained in space and happening over a long time, and so we attempted to produce a synthesis of both forms of technique. As such, in the evaluation of Ego we attempted to claw back some of the properties of ethnography – such as ongoing observation – which were lost by adopting more summative methods. This resulted in evaluators continually observing participants' activities online in order to inform later face-to-face interviews or questionnaires, as well as carrying out a form of orchestration, intervening as and when required in order to fix technical glitches and keep the system running smoothly.

Adaptive evaluation can thus come to employ existing orchestration techniques applied to other systems, as described earlier in this paper. We note that in these more performance-based systems (such as Uncle Roy All Around You or Day of the Figurines), orchestration is geared towards maintaining the performance as well as ensuring the smooth running of technology. An adaptive perspective on evaluation then couples these two facets of orchestration, using the idea that the continual involvement with a system's execution – part and parcel of general orchestration duties – can then also be used to inform ongoing evaluation.

In summary, this paper does not explicitly argue for a particular evaluation perspective, as there are times when either quantitative or qualitative approaches are most appropriate, although different regions of the figure favour different evaluation techniques. However here it is worth characterising

our adaptive approach in the context of other recommended methods implied within the discussions of evaluation techniques employed for specific systems thus far. For systems that involve smaller, room- or building-sized spaces and shorter durations from minutes to hours, we would recommend constructing video record from multiple angles, perhaps also embedding sensors within the environment and also potentially conducting more formal experimental setups. Evaluator involvement within these spaces is increased. Within such evaluations, logging is important, although remote monitoring is less vital than for other regions. Remote monitoring is likely to be unnecessary given the increased evaluator access. For systems run in much larger (e.g., city-sized) spaces but over similar time periods, logging comes into its own as a vital tool, as does remote monitoring, however equally important is the use of mobile camera operators recording video when and where possible. These fragments are more feasibly pieced together after the event. Systems in smaller spaces but over longer durations in turn preclude more exhaustive video recordings, again favouring remote monitoring and logging as useful ways to enrich a video record which ideally takes short samples across the duration of the systems run. Finally, for systems in the more problematic larger space, longer duration region, we recommend our adaptive evaluation techniques.

The challenges of using adaptive evaluation

There are three overarching and inter-related challenges to consider in the design of adaptive evaluations: firstly, the appropriate triangulation of both evaluation methods and of evaluation data; secondly, capturing adequate amounts and types of data while avoiding data overload; and thirdly, maintaining scientific rigour. Although the triangulation of data from multiple sources and of mixed methods is an approach (Denzin, 1978; Mackay & Fayard, 1997; Wilson, 2006) rather than a problem, a challenge lies in the appropriate selection of methods and/or data to be logged in order to (a) answer the research questions of concern, while secondly (b) avoiding data overload, and lastly (c) to formulate reliable findings. Data overload may occur due to the sheer volume of data being captured, as could potentially be the case in any long-term or large-scale evaluation. It is exacerbated somewhat during adaptive evaluations when ongoing analysis of data is required in order to maintain an awareness of what is happening "in the field". While the question of scientific rigour should be of concern to any researcher, when employing adaptive evaluation methods researchers must strive to ensure that if an evaluation is adapted for a particular user or subset of users, the subsequent findings remain comparable with user group as a whole and that the nature of the adaptation over time is well-documented.

Conclusion

While we would not necessarily claim that what evaluators did on Ego – i.e., orchestrating the experience and modifying the research questions on-the-fly – is radically new, we suggest that it contributes toward the argument that more strongly adaptive evaluation is an appropriate strategy for overcoming the kinds of control problems faced when evaluating user experiences of large geographic and temporal scale.

A variety of methods have been discussed that enable the evaluator to maintain some degree of control and connection. Adaptive journals such as FlexiFill, like experience sampling, aim to capture elements of the evaluation that may be neglected if reflection is attempted only at a later date. Logged data can be visualised post-hoc (Greenhalgh et al., 2007; Morrison et al., 2006) or alternatively can be streamed in real-time, thus providing a continuous awareness mechanism for the otherwise isolated evaluator. Orchestration techniques commonly employed in performance-based systems and games (Crabtree et al., 2004; Vogiazou et al., 2006) allow evaluators to direct the course of the evaluation as their research questions change in the light of ongoing observations. We suggest that when combined, such a collection of techniques afford the *adaptive evaluation* of Ubicomp systems in the wild, and open up new directions for future work on novel tools and methods for evaluation. And, although many of the systems we have reviewed in this paper (either our own or others') are games-based, we would also argue for the relevance of adaptive evaluation techniques to a broad range of Ubicomp domains.

More generally, we see strong benefits for evaluators in taking advantage of the same design principles and technologies that we are developing for users, in terms of using wireless networks and distributed sensors and cameras as tools for maintaining awareness, and of building up models of context and information with use (and perhaps even with users). We see the potential to make evaluation more of a synchronous engaged experience despite the vagaries of geographic and temporal scale, shifting context, and the work of fitting Ubicomp evaluation into the routine of our own everyday lives.

Acknowledgements

Several people apart from the authors worked on the systems and evaluations described here. In particular, we thank Louise Barkhuus, Marek Bell, Barry Brown, John Ferguson, Malcolm Hall and Paul Tennent. This research was funded as part of the Equator interdisciplinary collaboration (UK EPSRC GR/N15986/01).

Author Biographies

Scott Sherwood is a Research Assistant in Computer Science at the University of Glasgow, UK. He is a member of the Social/Ubiquitous/Mobile (SUM) Group and works predominately on ubiquitous technologies and mobile systems. While his interest is varied his main focus is studying systems within their context of use and in particular how ubiquitous technology is used in everyday life to manage the impressions individuals give to others.

Stuart Reeves is currently a Research Assistant at the Department of Computing Science, University of Glasgow. He is interested in interfaces and technology situated within public or semi-public settings, with particular focus upon performance and spectatorship.

Julie Maitland is a final year Ph.D. student in the SUM Group in the Department of Computing Science at the University of Glasgow. She takes an interdisciplinary approach to explore the underlying dynamics of peer-involvement in health-related behaviour management and investigate how social

support can be effectively integrated into the design of health-related behaviour management systems. She comes from a healthcare background, having previously trained and worked as a Staff Nurse in various acute and community settings within the NHS.

Alistair Morrison is a Research Assistant at the University of Glasgow's Computing Science Department. His background is in information visualisation and his Ph.D, completed in 2005, focussed on methods for exploring high dimensional data sets. More recently his research has focussed on tools and techniques for analysing data collected from trials of ubiquitous and mobile computing systems.

Matthew Chalmers is a Reader in Computer Science at the University of Glasgow, UK. After a PhD at U. East Anglia on ray tracing and object-oriented toolkits for distributed memory multiprocessors. he worked at at Xerox PARC and EuroPARC. He ran an information visualisation group at UBS Ubilab, in Zurich, and then had a brief fellowship at U. Hokkaido, Japan, before starting at U. Glasgow in 1999. He mostly works in ubiquitous computing, leading the SUM Group, but also maintains an active interest in information visualisation. He is an associate editor for the journals Information Visualization and Pervasive and Mobile Computing, has been an Associate Chair for ACM CHI and on the paper committees of ACM UIST, ACM CSCW, ECSCW, IEEE Information Visualization, Ubicomp, Pervasive, ECIR, CKIM, PerCom and... others. He was one of the authors of the UK Grand Challenge on ubiquitous computing.

References

Barkhuus, L., Chalmers, M., Tennent, P., Hall, M., Bell, M., Sherwood, S. & Brown, B. (2005) Picking Pockets on the Lawn: The Development of Tactics and Strategies in a Mobile Game. In Beigl, M., Intille, S., Rekimoto, J., & Tokuda, H. (Eds.) *Ubicomp*, (pp. 358-374). Tokyo: Springer

Barkhuus L., Brown B., Bell M., Sherwood S., Hall M. & Chalmers M. (2008) From Awareness to Repartee: Sharing Location within Social Groups. In Czerwinski, M., Lund A., M. & Tan, D., S. (Eds.) *CHI*, (pp. 497-506), Florence: ACM.

Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown, B., Rowland, D., Benford, S. & Hampshire, A. (2006) Interweaving Mobile Games with Everyday Life. In Olson, G., M. & Jeffries, R. (Eds.) *CHI*, (pp. 417-426), Montreal: ACM Press.

Benford, S., Rowland, D., Flintham, M., Drozd, A., Hull, R., Reid, J., Morrison, J. & Facer, K. (2005) Life on the edge: supporting collaboration in location-based experiences. In van der Veer, G., C. & Gale, C. (Eds.) *CHI* (pp. 721-730), Portland: ACM.

Benford, S., Seager, W., Flintham, M., Anastasi, R., Rowland, D., Humble, J., Stanton, D., Bowers, J., Tandavanitj, N., Adams, M., Row-Farr, J., Oldroyd, A., & Sutton, J. (2004) The error of our ways: The

experience of self-reported position in a location-based game. In Davies, N., Mynatt, E. & Siio, I. (Eds.) *Ubicomp* (pp.70–87), Nottingham: Springer.

Carter, S. & Mankoff J. (2005) Prototypes in the Wild: Lessons Learned from Evaluating Three Ubicomp Systems, *Pervasive*, 4(4), 51-57.

Carter, S., Mankoff, J., & Heer, J. (2007) Momento: Support for Situated Ubicomp Experimentation. In Rosson, M., B. & Gilmore, D., J. (Eds.) *CHI*, (pp. 125-134), San Jose: ACM Press.

Cheok, A. D., Fong, S. W., Goh, K. H., Yang, X., Liu, W., & Farzbiz, F. (2003). Human Pacman: a sensing-based mobile entertainment system with ubiquitous computing and tangible interaction. In Jamin., S. (Ed.) *NetGames* (pp. 106-117), Redwood City: ACM Press.

Consolvo, S. & Walker, M. (2003) Using the Experience Sampling Method to Evaluate Ubicomp Applications. *Pervasive Computing,* 2(2), 24-31.

Crabtree, A., Benford, S., Capra, M., Flintham, M., Drozd, A., Tandavanitj, N., Adams, M., & Row-Farr, J. (2007) The Cooperative Work of Gaming: Orchestrating a Mobile SMS Game. *CSCW* 16(1-2), 167-198.

Crabtree, A, Benford, S., Greenhalgh, C., Tennent, P., Chalmers, M., & Brown, B. (2006) Supporting Ethnographic Studies of Ubiquitous Computing in the Wild. In Carroll, J., M. (Ed.) *DIS* (pp. 60-69) Pennsylvania: ACM.

Crabtree, A., Benford, S., Rodden, T., Greenhalgh, C., Flintham, M., Anastasi, R., Drozd, A., Adams, M., Row-Farr, J., Tandavanitj, N., & Steed, A. (2004) Orchestrating a mixed reality game 'on the ground'. In Dykstra-Erickson, E. & Tscheligi, M. (Eds.) *CHI* (pp. 391-398), Vienna: ACM.

Crabtree, A. & Rodden, T. (2004) Domestic Routines and Design for the Home. *CSCW* 13(2), 191-220.

de Sá, M., Carriço, L., Duarte, L., & Reis, T. (2008) A framework for mobile evaluation. In In Czerwinski, M., Lund A., M. & Tan, D., S. (Eds.) *Ext. Abs. CHI* (pp. 2673-2678), Florence: ACM.

Denzin, N. (1978) Sociological Methods: A Sourcebook, NY: McGraw Hill, 2nd ed

Demiris, G., Oliver, D., Dickey, G., Skubic, M. & Rantz, M. (2007) Findings from a participatory evaluation of a smart home application for older adults. *Technology and Health Care*, 16, 111-118.

Dourish, P. (2004) What We Talk About When We Talk About Context. *Personal and Ubiquitous Computing*, 8(1), 19–30.

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. (2007) MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In Knightly, E., W., Borriello, G. & Carceres, R. (Eds.) *MobiSys* (pp. 57-70), San Juan: ACM.

Grant, L., Daanen, H., Benford, S., Hampshire, A., Drozd, A., & Greenhalgh, C. (2007) MobiMissions: the game of missions for mobile phones. In Swanson, J. (Ed.) *SIGGRAPH 2007 Educators Program*, San Diego: ACM.

Goodwin, M.H. (1980) He-Said-She-Said: Formal cultural Procedures for the Construction of a Gossip Dispute Activity. In *American Ethnologist*, 7 (4). 674-695.

Greenhalgh, C., French, A., Tennant, P., Humble, J., & Crabtree, A. (2007) From replay tool to digital replay system. In Online Proceedings of *E-Social Science*, Ann Arbor: ESRC/NSF.

Macefield, R. (2007) Usability Studies and the Hawthorne Effect. *Journal of Usability Studies*, 2(3), 145-154.

Mackay, W.E. & Fayard, A-L. (1997) HCI, Natural Science and Design: A Framework for Triangulation Across Disciplines. In Coles, S. (Ed.) *Designing Interactive Systems*. (pp. 223-234), Amsterdam: ACM.

Maitland, J., Sherwood, S., Barkhuus, L., Anderson, I., Hall, M., Brown, B., Chalmers, M. & Muller, H. (2006) Increasing the Awareness of Daily Activity Levels with Pervasive Computing. In Bardram, J., E., Chachques, J., C. & Varshney, U. (Eds.) *Pervasive Computing Technologies for Healthcare* (pp.1-9) Innsruck: IEEE.

Morrison, A., Tennent, P. & Chalmers, M. (2006) Coordinated Visualisation of Video and System Log Data. In Andrienko, G., Roberts, J., C. & Weaver, C. (Eds.) *Coordinated and Multiple Views in Exploratory Visualization* (pp. 91-102), Zurich: IEEE.

Rice, B. (1982). The Hawthorne defect: Persistence of a flawed theory. *Psychology Today*, 16(2), 70-4.

Scholtz, J. & Consolvo, S. (2004) Toward a Framework for Evaluating Ubiquitous Computing Applications. *Pervasive Computing*, 3(2), 82-88.

Vogiazou, Y., Reid, J., Raijmakers, B., & Eisenstadt, M. (2006) A research process for designing ubiquitous social experiences. In Mørch, A., Morgan, K., Bratteteig, T., Ghosh, G. & Svanaes, D. (Eds.), *NordiCHI*, (pp. 86-95), Oslo: ACM Press.

Want, R., Hopper, A., Falcao, V. & Gibbons, J. (1992) The active badge location system. *Transactions on Information Systems,* 10(1), 91-102.

Wilson, C. E. (2006.) Triangulation: the explicit use of multiple methods, measures, and approaches for determining core issues in product development. *interactions* 13(6), 46-ff