Further into the Wild: Running Worldwide Trials of Mobile Systems

Donald McMillan, Alistair Morrison, Owain Brown,
Malcolm Hall, Matthew Chalmers

Department of Computing Science, University of Glasgow, UK
{donny, morrisaj, owain, mh, matthew}@dcs.gla.ac.uk

Many studies of ubiquitous computing systems involve deploying a system to a group of users who will be studied through direct observation, interviews and the gathering of system log data. However, such studies are often limited in the number of participants and duration of the trial, particularly if the researchers are providing the participants with hardware. Apple's App Store and similar application repositories have become popular with smartphone users, yet few ubiquitous computing studies have yet utilised these distribution mechanisms. We describe our experiences of running a very large scale trial where such a distribution model is used to recruit thousands of users for a mobile system trial that can be run continuously with no constrained end date. We explain how we conducted such a trial, covering issues such as data logging and interviewing users based in several different continents. Benefits and potential shortcomings of running a trial in this way are discussed and we offer guidance on ways to help manage a large and disparate user-base using in-application feedback measures and web-based social networking applications. We describe how, through these methods, we were able to further the development of a piece of ubiquitous computing software through user-informed design on a mass scale.

Keywords: Evaluation Techniques, Large Scale Deployment, Trial Methods

1 Introduction

It is often considered beneficial to conduct trials of ubiquitous computing (ubicomp) systems 'in the wild' i.e. in uncontrolled contexts and environments that are typical of everyday use of many modern technologies [1]. In contrast to the lab-based environment of more traditional usability-style studies, it has been argued that experiments carried out *in situ* can help evaluators gain insight into how people fit systems into their existing practices and contexts of use, and how people change their contexts and practices to accommodate or take advantage of new systems. While this approach has its benefits, the staging of ubicomp system trials may give rise to a number of practical issues that inhibit evaluators' ability to draw substantive conclusions on system use. For example, many trials involve providing each participant with a mobile device on which to run the system under investigation. This in itself can introduce biases into the trial: participants are dealing with a device with which they are not familiar, and there will likely be a period of acclimatisation during which they might not use the new technology as naturally, or with the same degree of skill, as more experienced users. Merely having to carry around an extra device during a long-term trial might have an effect on some participants—it is likely that they will already be carrying mobile phones and perhaps also cameras, so the obligation to carry around additional hardware might affect participants' perceptions of the system, or they may simply not always carry the trial device around or use it as much as experimenters hope or expect.

Another limiting factor that arises when providing participants with new hardware on which to run a system under investigation is the number of devices that can be supplied, and therefore the number of participants available for the trial. Most research projects have a specific budget for trial hardware, but this rarely stretches to pay for thousands or even hundreds of devices such as smartphones, and so the size of experiment that can be conducted is necessarily limited to a relatively small number, e.g. 10–20. Such hardware may be shared by several experiments in the same project, or in several projects, and this may create pressure to keep trials short so that different experiments can take place.

Furthermore, if participants are supplied with devices by researchers, it is common practice to recruit these users from the researchers' local area. Many university-based research teams will use student volunteers as participants, for example, or other participants who reply to adverts placed around the campus. Although many interesting findings are of course possible from such a user-base, an evaluator could not realistically extrapolate these insights into conclusive statements in a global sense; how a group of university undergraduates adopt a particular technology may not be typical of the wider community in the same urban area, and communities in a different continent may be even more different. So, not only does a local participant set give rise to the dangers of basing findings on a very narrow subset of a technology's potential user-base, it also leaves no possibility for studying cultural differences by comparing many geographically distant groups of users.

A step towards addressing some of these issues is running a trial of a ubicomp software system not on experimenter-supplied devices, but on devices the participants already own and use daily. Only in very recent years have we seen mobile phones that are both numerous enough to afford a large trial as well as advanced enough to support downloading and installation of researcher-supplied software. Market research firm IDC [2] suggests that, at the end of 2009, 15.4% of the mobile phone market consisted of smartphones, an increase from 12.7% in 2008. So, while still not the predominant type of handset, we suggest that smartphones have been adopted into mainstream use. While running a trial solely with smartphone owners may not be selecting a user-base that is representative of the population at large, it is not now using only the most advanced 'early adopters'. By recruiting smartphone owners, we may be able to avoid or reduce some of the issues outlined earlier, in particular the small number of hardware devices that a research project can generally supply and the length of time that a trial can last for.

In this paper we describe our tools and techniques for recruiting smartphone owners for ubicomp trials, deploying systems amongst them, directing questions to users and encouraging social interaction among them. We document our experience of a system's deployment among 8676 active users. A key element of our recruitment and deployment was our use of a public software repository rather than directly supplying software to trial participants. Although a recent phenomenon, such repositories are a well-established means of deploying software to smartphone users. Apple's App Store has proved to be a very popular and effective means by which iPhone users can access new software, and several other mobile platforms now have similar repositories. Despite their popularity, the potential for such repositories to be used as a distribution mechanism for research prototypes, while having been touched upon in [3], has not yet been explored and documented, and yet several potential benefits of such a mechanism are apparent, e.g. such repositories already offer means for users to browse and find software they are interested in, and so researchers can effectively advertise a system and recruit participants for a trial by putting the system into a repository.

This paper describes our experiences of making a free application available in this way, a mobile multiplayer game called Hungry Yoshi. This is a new version of Feeding Yoshi, a seamful game that we ported to the Apple iPhone and updated. Feeding Yoshi's main trial was described in [4] as a "long-term, wide-area" trial "being played over a week between three different cities" in the UK. We wished to scale up our deployments and trials as part of a project, Contextual Software, that explores system support for collaboration with communities of users in the design and adaptation of software to suit users' varied and changing contexts [5]. Distribution in the App Store style, along with our new tools and infrastructure, allowed for a trial that involved a much larger number of participants than before, who were far more geographically dispersed than we could previously handle, and which lasted longer than any trial we have ever run. At the time of writing, the current trial of the new Yoshi system has been running for five months and has involved thousands of users from all around the world.

The following section describes work related to this and other examples of large-scale trials, as well as outlining the original Yoshi system and trial. This is followed by a description of the re-design of Yoshi for use on the Apple iPhone and wide-scale distribution. Thereafter we describe the processes involved in distributing the game to a global audience, managing a trial involving a large and widely distributed user-base, and involving those users in development of a new system feature. We then discuss some methodological and practical issues before we offer our conclusions.

2 Related Work

Several ubicomp projects have featured data collected from large numbers of people via mass-scale sensing. An example is the Cityware project [6], which collected data from scans of Bluetooth devices detectable by static recording equipment at various locations around a city in order to measure densities and flows of people in particular urban areas, which in turn were to be used in architecturally based models of those areas. In a related theme, abstractions similar to those of the Cityware work but at an even larger scale were shown in [7], which involved the generation of coarse-grained city-scale maps of people's density based on concentrations of mobile phone signals sampled from GSM infrastructure. While this work undoubtedly exhibits great scale, it is different to the area we are investigating in that sensor data is collected and aggregated, rather than data on the use of applications. Also, such techniques do not directly feed into qualitative investigations of social and personal behaviour, a useful combination that we aim to support.

In 2008 Nokia Research Centre released Friend View, a "location-enhanced microblogging application and service" [8] via Nokia's Beta Labs. This site allows its community members to contribute feedback to in-development and experimental software, but this study reported only on statistical analysis of social network patterns based on anonymised log data representing 80 days' use by 7000 users. Like Cityware, this serves as an example of many quantitative studies in which potentially interesting analyses were carried out, but no interaction with users is described that would allow analysis to be contextualised with user experience, or determine how users' opinions, behaviour or systems might change in the light of such analyses.

One of the early landmarks of large-scale deployment of ubicomp applications was *Mogi Mogi*. As reported by Licoppe and Inada [9], this location-based mobile multiplayer game was released commercially in Japan, and in 2004 had roughly 1000 active players. Some basic aggregate analyses involved system profiles, e.g. gender and age group, but almost all the presented analysis is based on more ethnographic interviews and observations of ten players who were, apparently, strangers to each other. This method afforded rich detail of the ways that they fit the game into urban contexts and lifestyles, based on months of game play, including occasional social interactions between players.

In 2006, we trialled Feeding Yoshi, running what we called "the first detailed study of a long-term location-based game, going beyond quantitative analysis to offer more qualitative data on the user experience" [4]. The participants consisted of four groups of four people and, as mentioned above, the main study lasted a week. The participants in each group knew each other before the trial, and collaboration and social interaction was observed during the trial. The study drew on participant diaries and interviews, supported by observation and analysis of system logs. Somewhat like the study by Licoppe and Inada, it focused on how players "interweaved the game into everyday life" and how wireless network infrastructure was experienced as a 'seamful' resource for game design and user interaction.

Observational techniques founded in ethnography may be well suited in principle to studying ubicomp systems, but in practice they are often hampered because keeping up with the 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 [10]. Several video cameras may be used to record activities in several locations set within some larger activity, but this brings the practical problem of synchronisation, and how to gain an overview of this material and combine it with other relevant data, such as system logs gathered from the mobile devices. Furthermore, network connectivity may be intermittent or costly enough to hamper attempts to keep in continuous contact with users and their devices, e.g. to stream log data back to evaluators or developers. Consequently, some researchers have explored 'experience sampling' methods, in which a questionnaire appears on-screen when the mobile device detects that it is in a context of interest [11]. Carter and Mankoff developed Momento [12], 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. *Replayer* [13] similarly combined quantitative and qualitative data in order to offer a more holistic view of systems in use, and to let researchers study users acting in greater numbers, and at larger geographic and time scales, than they can directly observe. In particular, it used spatial properties within quantitative log data so as to make analysis of qualitative data less time-consuming and therefore allow larger trials to be run.

3 Hungry Yoshi

Feeding Yoshi [4] was a mobile multiplayer game for Windows Mobile PDAs. It was re-implemented for the Apple iPhone and renamed Hungry Yoshi. It uses wireless networks infrastructure as part of its design. Players' mobile devices perform regular scans of the WiFi access points visible in the current area, classify each of these access points according to its security settings and display it to the player. Each password-protected access point is deemed to be a creature called a 'yoshi' whereas a network without password protection appears as a 'plantation' growing a certain type of fruit. Yoshis ask players for particular fruit, and players score by picking these fruit from the correct trees and taking them to the yoshis. Yoshis also provide seeds that enable players to grow new fruit in empty plantations. A research objective of the 2006 study of Feeding Yoshi was to establish how players could interweave playing a long-term game with their everyday lives. Four teams of four players were used in the trial, each being paid for taking part, with a competitive element introduced such that the members of the team with the highest combined score received double the standard participation fee.

Hungry Yoshi has some differences to Feeding Yoshi. Perhaps the biggest change is that, with the availability of the iPhone's data connections over cellular networks, the system can generally maintain a globally synchronous game world. In the old game, yoshis and plantations visited and their contents were stored only on players' mobile devices, so two players might visit the same plantation and see it containing different contents. By storing such details on a centralised server, one player can seed a plantation with a fruit type and another can pick the fruit when they grow. A new piece of functionality in the iPhone version of Yoshi is the ability to change pieces of fruit for a small cost. Players are able to insert fruit into a fruit swapper (Figure 1-right) that returns a different type of fruit at random. To use this swapper, players are charged tokens, which can be earned by performing tasks. Section 4.3 explains why this task mechanism was important for helping us interact with the users during the trial.

Another difference from the original trial is that the game no longer has any explicit team element: each player participates as a solo entity. However, the score table is retained as a form of motivation for players, though with the difference that now there is no prize at the end of the trial and indeed no defined end to the playing of the game. Separate score tables are maintained for overall score, score this week and score today, the latter two being used because new players might join in at any time and could be months behind the early users. The table screen is divided into a top section showing the top players, and underneath, a section showing the players around the user's current position.

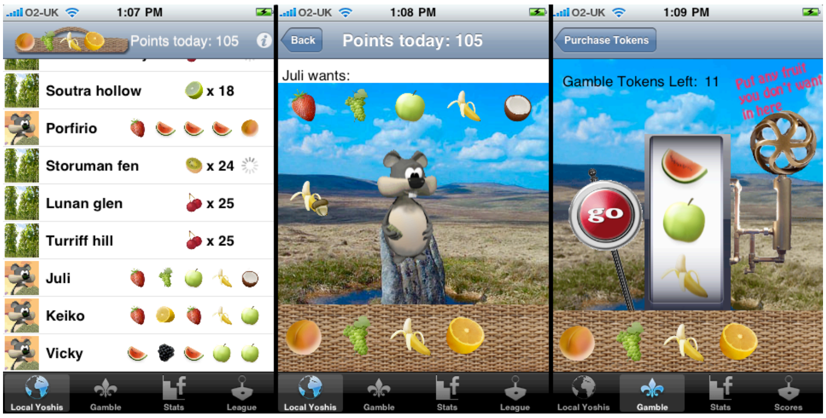

Fig. 1. The list of nearby yoshis and plantations (*left*), a yoshi screen (*centre*), and the fruit swapper (*right*).

4 Engaging Users Worldwide in Iterative Design

This section describes how Yoshi was evaluated and modified in the course of our trial. It discusses our approach to distribution, management, data gathering, analysis and redesign, coupling them together in a form of iterative design suited to the large scale of our trial. We outline how we interacted with trial participants, how users interacted with each other, and how these interactions fed into a new version of Yoshi so as to begin another design iteration.

4.1 Distribution

Hungry Yoshi was released in early September 2009. At the time of writing it has been publicly available for five months. Distributing software via a public repository means using a mechanism that users are already very comfortable with, again possibly leading to more naturalistic interactions than with a more contrived physical meeting and handover of a device or software. Yoshi appears in the 'games' section on the store, and so benefits from recruiting users who deliberately seek out this type of application and who will hopefully therefore be more keen to engage with the game. An unanticipated but welcome benefit to this form of distribution is free advertising outside of the store and beyond our own announcements of the trial, e.g. in interviewing one of Yoshi's users, we learnt that she first heard of the game in a review in an Italian technology blog. In releasing a research prototype through a public marketplace, we harness some of the enthusiasm of amateur and professional writers who regularly scour the store for new applications to try and discuss.

Figure 2 charts the number of downloads of the game over the time that the game has been available. When the game was first released, and when updated versions are made available, it features near the top of the store's "most recent" lists, providing a boost in the number of downloads that day.

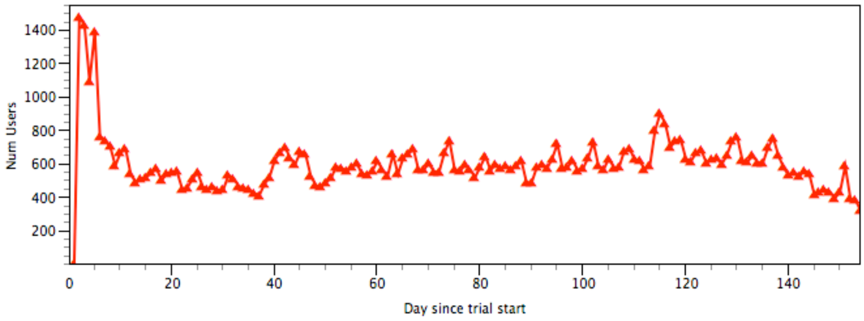

Fig. 2. Number of downloads of Yoshi per day since release

It can be seen that there was a peak of interest in the first few days following the game being launched, after which download figures were around 600 per day. There appears to be a gradual trend upwards, perhaps falling off only in the last month or so. Occasional spikes, such as that at 40 days, correspond to the release of new versions. At the time of writing there have been 137367 downloads in total. This figure includes people updating to new versions of the game; we recorded 94642 unique downloaders. Figure 3 shows the proportion of players of the game each day who are playing it for the first time, as compared to those who have played the game before. It can be seen that by the end of this period, the proportion of returning players is increasing although around 25% of players are playing for the first time each day.

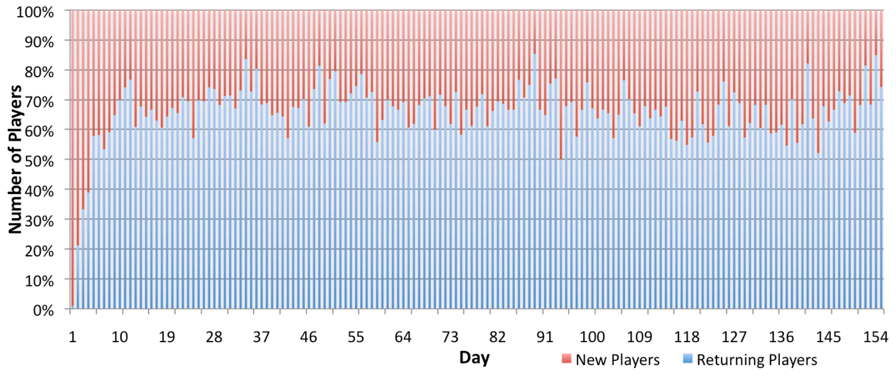

Fig. 3. Proportion of new and returning players per day of the trial.

Having installed the game and on opening the Yoshi application for the first time, users are presented with an information page, written in English, French, German and Japanese, that explains that the system is created as part of a research project and that details the various forms of information that will be logged during interaction with the game. The page also states that researchers might contact users to enquire about their use of the software, but that these communications can be ignored and would cease on request. Only by agreeing that they have read and understood these terms can players proceed into the game. Further to this, we state that log data will be stored securely, that users can opt out at any time, that we will destroy the data logged about them on request and that all the data will be destroyed following the end of the project. To date, no such requests have been received. Links are provided to an email address for the trial organisers, and to a public Web forum where users can either chat amongst themselves or seek clarification on any aspect of the game or research trial from the organisers. The data logging process is described below in Section 4.2. At the time of writing, 24408 out of the 94642 downloaders registered with the game and agreed to be part of the trial. This reduction may be because people were wary of having their data logged in the manner described, were perhaps apprehensive over being contacted by researchers or were deterred by having to register a user account. Of those 24408, many only briefly interacted with the game, but 8676 played for long enough to produce log data that could be studied. Although this represents only around 9% of the total number of downloaders, the number of players is still very large.

Quantitative analysis benefits from having such a large user-base. Having information gathered from thousands of users allows many inferences to be made with a much higher degree of confidence than if an experiment had been run with, for example, the 16 participants we had in 2006. Results of our quantitative analyses are covered in the following section, and we offer some reflections on this scaling up in the later Discussion section.

4.2 Quantitative Analysis

To aid our evaluation of Yoshi, system log data is generated from every trial participant's phone. The system makes use of our SGLog logging framework (described in detail in [5]), which manages data collection on the phone and periodic uploads to a server using the same data connection required to run the game. The data logged includes activities within the game, such as feeding a particular yoshi, and general contextual information. Uploaded data from each user is timestamped and stored on a database on a central server. To protect the privacy of participants, this framework uses TLS to encrypt data sent between phones and the server.

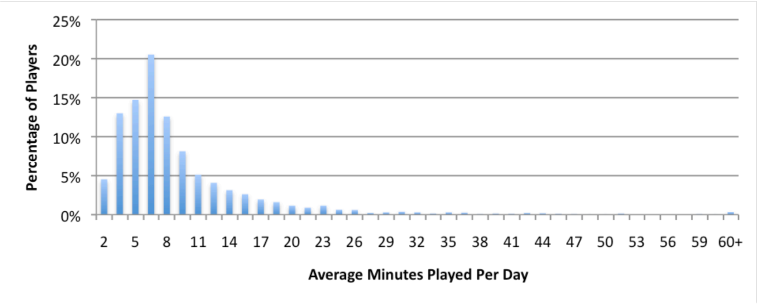

Fig 4. The distribution of players' average system use per day, with a mode of 6 minutes (20.5% of players).

Figure 4 shows the distribution of the average amount of time each user played the game each day. This time was calculated by looking at timestamped game events registered on the server, rather than simply the times at which the application was running, so times when the device was sitting idle do not contribute to the figures. It can be seen that there is a range of levels of activity, with several players playing for over an hour a day on average.

One player's average daily play was significantly longer than the rest. Over the first two months of the trial, she had an average of more than 2.5 hours of play per day and at the time of writing has played the game for over 200 hours. She is the game's top player, and has been at the top of the overall score table since the early days of the trial, with around double the overall score of the second highest-scoring player. In any trial it is probable that researchers will observe a variety in the level of engagement shown by users. In running an experiment with hundreds or thousands of participants, it is likely that this spread will be wider, and that some of these users will be more enthusiastic. For example, in the original trial of Yoshi [4], the longest time a player spent playing the game in any one day was 2.5 hours, whereas here this figure is almost 7.5 hours.

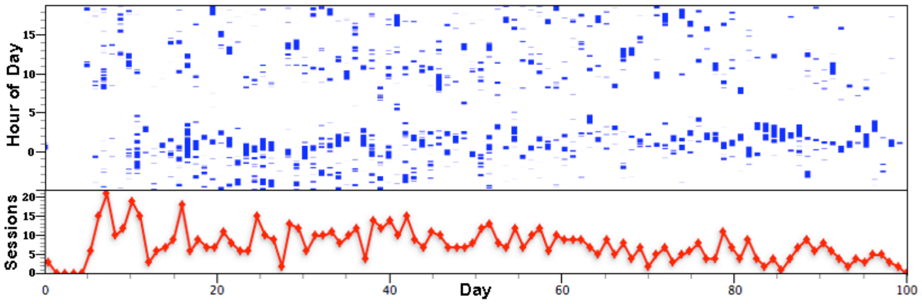

Fig. 5. A visualisation of one of the most active players' use over the first 100 days of the trial. In the upper section, the *x*-axis shows days since the trial began and the *y*-axis shows the 24 hours of the day, with blue shading showing the periods at which the participant was playing the game. The lower section shows the number of 'sessions' per day for the same user, with 'session' meaning a period with less than five minutes between each user action.

Figure 5 shows one of the top-scoring players' activity in greater detail. The number and lengths of lines give a quick impression of the amount of activity this user has engaged in, and the length of these lines shows whether the user favours long sessions or quicker games, squeezing a short burst of play into a spare few minutes. We also see daily patterns, e.g. finding a strongly shaded row in the plot would indicate regular play around that time of day. Quantitatively-based visualisations such as these were useful both in themselves, in letting us see basic patterns of use, but also in feeding into qualitative analysis, e.g. in selecting participants to interact with more directly, and in preparing for such interactions—as described in section 4.3.

4.3 Interacting with participants

One of the challenges of conducting a worldwide system trial lies in managing interaction with participants: maintaining a presence with participants, harnessing feedback and supporting qualitative analysis. As we would not be able to regularly meet participants, as one might do in a more standard trial with locally sourced trial subjects, alternative means were sought to keep in contact with our users. We used two mechanisms: tools for communication within Yoshi, and communication via a social networking web site.

In-game communication with users

Rather than have communication with users happen in a way that clashed with the user experience, we built bi-directional communication into the functionality of the Yoshi game. Section 3 introduced the fruit exchange mechanism, which users were charged tokens to use. These tokens were earned by players performing tasks, set by researchers throughout the course of the trial. In this way, we could relay messages to participants, ask specific questions and receive feedback as appropriate.

The tasks set to users in this manner took a number of forms. Simple factual questions such as age, gender and continent of residence were asked, with users selecting answers from drop-down lists. This provided a simple means for us to build up demographic profiles of the user-base. More open-ended questions which allowed free text responses were also set, such as what a player liked about the game, and whether he/she had any suggestions or bugs to report—as we detail later. This system proved to be of particular benefit because the tasks could be dynamically updated in real-time during the course of the trial, and because specific questions could be targeted towards a particular user or set of users in response to some interesting activity we observed in their log data or our interactions with them. The tasks available to a player are downloaded from the server each time the player visits the task list screen. Therefore, although the system is deployed to a worldwide user-base, and we could not access devices to update the software on them, we could alter the questions at any point during the trial. Once edited, the new task set becomes live immediately, thus supporting adaptation of our research interests.

The task and token-earning functionality proved popular with users, with 28442 responses in total. Before the trial, we were unsure whether players would use this feature 'honestly' or would provide dummy answers. As no checks were in place, free text answers could be submitted as empty or with a few random characters, and players would still be rewarded tokens by the automated system. However, results proved that users were willing to engage with this feature, providing answers of varying length, but in the main making an attempt to answer in a useful way. As an example, a task asking demographic information from the user was completed 2406 times, with all but 73 being sensible answers to the question. While the tasks themselves were in English, care was taken to ensure that where possible the grammar and vocabulary used fell within the Common European Framework of Reference for Language's A2 level bounds, a level achievable by most attending public school in westernised countries where English is taught as a second language [14].

Interacting with participants through Facebook

Although the task system provided a basic communication mechanism between researchers and participants, more powerful external tools were also used in order to facilitate more in-depth dialogues and to support communication between participants themselves. We elected to use Facebook, a popular online social networking application, as a means of supporting such interactions. Facebook has more than 300 million active users, 50% of whom log on to Facebook in any given day [15], making it an appealing choice of platform for this task. Also of benefit was Facebook Connect, a service with an iPhone API that allows users to verify themselves and log in to third party sites and applications using their Facebook account. On starting Yoshi, players are required to log in to their game account in order to track their score across devices, and to allow multiple people using the same device to have individual accounts. This can be done either through Facebook Connect or by creating a username and password specifically for the game (which we called 'Lite mode').

Though we still wanted non-Facebook users to be able to play the game, we sought to encourage users to login through the Facebook Connect method to provide the benefits outlined above. As such, we limited use of the fruit swapper described earlier to only Facebook-logged in users; users logged in via Lite mode were prompted to login to Facebook when attempting to access this functionality. Additionally, we allowed users to post their Yoshi progress to their Facebook Feed (which shared their scores and rankings with all their Facebook contacts). This served both as an enticement to use the Facebook version, and as further user-generated advertising for the game. Of our 8676 users who agreed to the terms and played the game, 6732 elected to use the Facebook login, including 44 of the top 50 scorers.

In addition to providing a login mechanism, we also used content on the Facebook site itself both to provide features for the user and in contacting users to aid in the management of the trial. We created a Facebook application—a series of PHP-based web pages displayed within Facebook—that showed the ranked scores in greater detail and provided statistics on the players' game play, such as their most visited yoshis. More importantly, Facebook has a set of well-established means of communication both in one-to-one and one-to-many models. For example, as players had provided us with their login IDs, we could send emails to their Facebook accounts, and we set up a forum for players to communicate with each other and discuss potential new ideas.

4.4 Qualitative analysis

Section 4.2 described quantitative analysis performed on log data. With such data, gaining an in-depth understanding of individual player behaviour is challenging. While we could visualise various aspects of play, this did not necessarily make a player's motives and reasoning comprehensible. We now describe allied forms of qualitative analysis, centred on interviews that let us explore and clarify issues more adaptively than if, for example, we had used an on-line questionnaire to gather qualitative data. As will be discussed, some of the processes already described such as visualisations and Facebook tools were useful resources for this form of analysis.

Interview process

The process of interviewing participants worldwide is not quite as straightforward as in a more traditional experiment involving locally based users. Whereas in a traditional setup researchers are likely to have met participants before the trial begins, perhaps to deploy the system or to explain the trial, we had no direct contact with users at the beginning of our qualitative analysis process. Although all the users had agreed to a series of terms before playing the game, that explained that we might try to contact them, they had also been informed that they could feel free to ignore this communication or to tell us that they were not interested in participating. More positively, having over 8500 users gave us an opportunity to focus on interviewees that we deemed the most relevant to a design issue or potentially significant in terms of use and user experience. For example, we could choose the most active players, i.e. those who had accumulated the most game time, those who had answered a particular in-game question, or those who had a particular pattern of use in their system logs.

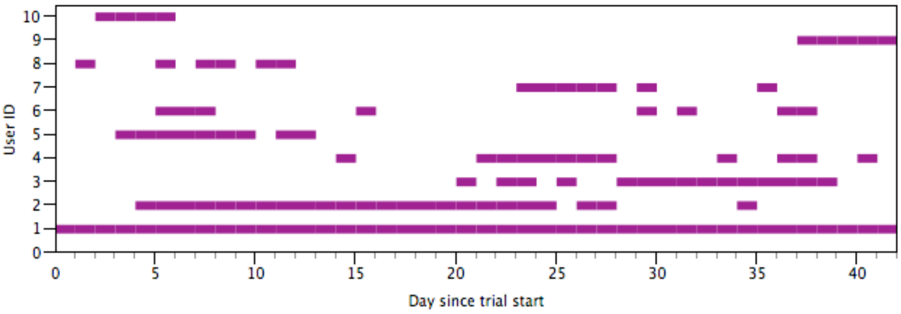

Fig. 6. A snapshot from a tool used to select participants for interview, showing which days each participant used the application. From a chart showing all the trial participants, ten users have been selected who exhibit contrasting patterns of use. This figure shows the filtered set.

Selecting participants for interview began by using the information publicly available to application developers via the Facebook API to first filter our participant database to show only those over the age of 18. Thereafter, we used visualisations of log data to examine participant use over time. For example, Figure 6 has a separate row for each participant, and shades days, shown on the *x*-axis, if the user played the game on that day. We were interested in speaking to a set of users with a diverse amount of engagement, so used the visualisation to select rows with contrasting patterns. Figure 6 shows such a subset of users. As can be seen, there is a wide variety of use, with user 1 playing every day, user 10 playing for a few days near the start of the trial before giving up and user 9 joining the game later than the others, but having played every day since starting until the current time. Simpler methods of selecting interviewees would obviously have been possible, such as choosing the highest-scoring players, but we were interested in using our methods of selection to interview a set of players which showed a broader range of activity in playing the game.

Having identified our interviewee set of choice, Facebook again proved a useful tool in making contact. We emailed users, enquiring whether they were interested in taking part in an interview over VoIP or telephone in exchange for $25 of iTunes or Amazon vouchers. We interviewed 10 of these players, from 5 different countries and 3 different continents. 5 were male and 5 were female, and they ranged in age from 22 to 38. As the trial progressed, we noted a shift in participants' willingness to be interviewed. In the first months of the trial, requests for interviews were met with enthusiasm and even pleasure at being given the opportunity to participate. However, the number of users willing to take part in the feedback process seemed to drop as the trial continued. We speculate that this is perhaps a result of the self-selection of participants involved in this type of trial; early adopters willing to download software on release and persevere through initial versions seem to have more interest in the process and greater willingness to participate than those who joined the trial later.

In addition to the visualisation shown in Fig. 6, other tools were useful in interview preparation. The tool seen in Fig. 5, showing activity across different hours of the day, helped to make the interviewer aware of the overall level of engagement the player had shown, as well as raise specific items of interest to ask about, such as if a player seemed to be using the application for five minutes every day at lunch time. Similarly, the answers that the interviewee had submitted to tasks were surveyed before the interview commenced, again to prime any potentially interesting aspects to ask about. Each interview began with another explanation of the trial and of who we are, and typically lasted between 15 and 45 minutes. All were transcribed afterwards.

Findings from analysis of interviews

In order to explore our research issue of running trials at a distance, we asked as many questions about the trial mechanism itself as about the application. Unsurprisingly perhaps, many of the same game experience themes arose in interviews as had been reported in the original Yoshi trial. For example, several participants mentioned that awareness of other players' scores, as shown in the table on Facebook and within the game itself, was a strong motivating factor for them. It should be noted that, unlike the first trial of Yoshi, no prizes are awarded to the best players; presenting names on the score table and the ability to share their success with all their friends, players and non-players, via Facebook, proved to be enough of an incentive for many players.

Our use of Facebook also afforded more direct interaction between players. By having full names visible on the scoreboard, and as the game had clear links to Facebook, users appeared to have a 'ticket to talk' to each other. For example, one participant (A) reported seeking out another player (B) on Facebook, to ask about what was perceived as unusual scoring patterns. From seeing B's name, A made assumptions about where B was likely to be based in the world and was confused about the times of day that B appeared to be accumulating points: "*I really couldn't figure out how they could have all those points when I was asleep*". After exchanging a few emails with each other, A discovered that B lived in a different continent. Their email conversation has continued, and they now consider themselves friends.

Turning now to the user trial mechanisms, interviewees were enthusiastic about the range of feedback mechanisms made available to them. In particular, players we interviewed were very positive about the task mechanism, with one saying "*I think it's a pretty good idea that I can answer certain questions for [tasks] so I can give feedback there. Even free-text feedback. And it's really good.*"

This trend is in accord with our analysis of task response rates and the number of sensible answers received. One interviewee specifically addressed having noticed that it was possible to just get tokens from submitting empty responses, but still felt he should give proper answers: *"Sometimes, you scan through, and just try and hit the submit button … you're just like, gimme these tokens, I wanna get on with it… But most times, I answer honestly, about 98% of the time."*

This enthusiasm for the task response mechanism extended to being emailed over Facebook to request an interview, with all interviewees responding positively when asked how they felt on being contacted, with one commenting: "*I find it really nice that [you are] contacting me and asking me my opinion. I guess it's a really nice thing*." Indeed, at the end of their interviews, two of the ten interviewees actually declined the payment that had earlier been agreed, saying that they were happy to participate. We speculate that this is maybe because we had provided a free game that these players evidently value. Of course, players who do not enjoy the game stop playing it and are not available for logging or interview—thus potentially biasing our 'sampling' of users. By targeting users with the task mechanism, Facebook messages and email we were able to quiz those who declined interview requests on their reasons for doing so. The response rate was low but those we did receive fell evenly into categories of general refusal, e.g. 'I don't have time.', and refusal based on perceived lack of language skills, e.g. 'I don't speak English.'

Users are playing of their own free will rather than perhaps feeling obligated by having agreed to participate in a system trial, and so their play is more 'natural' than those who use the system out of a sense of obligation or for financial benefit. As a result, compared to our experience of earlier trials of other systems, we observed that players had more good will towards 'giving something back' than we have observed in more traditional trials.

Although time-consuming to arrange and conduct, these interviews offered valuable insights into player behaviour and their reactions to the trial process and provided a valuable, rich communication channel through which detailed contextual understanding of logged data could be sought.

4.5 Redesign

Given the flexibility of the tools for interacting with users and studying log data, we were able to use the tools to ease the task of redesign. This reflects one of our research goals, which is developing means to quickly and appropriately adapt software to suit the changing contexts and interests of users.

For example, as an answer to the task "What could be improved about Yoshi?", one user (anonymised here as Helen) commented that plantations were often too full. Helen was invited for interview, and the interviewer then raised this point to obtain further detail. Helen explained that, as plantations auto-generate fruit at a rate of one per hour, they would often be full, which she felt was to the detriment of the game. In particular, Helen described a situation where she would empty a plantation before leaving for work in the morning, and wanted to collect a seed from work to plant when she got home. However, by this time the previously empty plantation would have around 10 pieces of fruit in it again, which would have to be picked first and fed to unwilling yoshis, leading to a points penalty.

Following this interview, the game designers agreed that this was a valid criticism that should be addressed if it reflected a common concern or problem among users.

We again used the task mechanism to consult our user-base at large. A question was added as a task in the game, in the form of a vote as to whether to introduce this feature, and exactly what form it should take. We presented three options: (A) leaving the game unchanged, (B) players could burn empty plantations to stop them re-growing (as suggested by our interviewee) and (C) even full plantations could be burned, which would also destroy all the fruit that had grown. 17% voted in favour of leaving the game as it was, while 29% were keen to see option B and 54% selected option C. The chosen feature was therefore implemented and distributed in a new Yoshi version, thus beginning another iteration in our design process.

On detecting that Helen had installed the new version, we contacted her again to gauge her reaction towards the new feature and she replied positively, agreeing that the version implemented, although not the design she had suggested, was the better of the new options. Around the same time, we included another vote on the new feature, consulting the opinion of the user-base at large after they had had a chance to use it. Users responded with approval, with 94% agreeing that they liked the new feature. This demonstrated to us a significant benefit in this iterative approach of conducting design by engaging with users at both a micro and macro-scale, and letting the results of one feed into the other.

System bug handling was dealt with in the same way. One user was having stability issues that were reported in-game through the task mechanism. Upon contacting the user for more information, the problem was narrowed down to be specific to his model and operating system version combination in areas of high access point saturation. This problem was resolved and the next update to the game was released. Over the five months the software has been live, seven versions have been released to the public.

By having interaction with evaluators integrated into the game dynamic, users are able to report issues directly within the relevant context of use. While these reports are generally brief, they provide a hook back to the context of use they were created in. In this respect, the log data was an invaluable tool for helping the user recall the context of use and therefore the detail and qualitative texture of the problem or suggestion he/she had reported previously. Placing the user at the scene of the problem or suggestion by discussing their location, the game actions they took leading up to the report, and how their pattern of play had evolved to the point where they noticed a problem gave interviewers a valuable means to elicit the detail necessary to pinpoint problems and ground suggestions.

5 Discussion

The tools and techniques described in the previous section let us carry out a relatively normal iterative design process but at an unusually large scale. Methodologically, when we compare our approach to more standard trials, we see both advantages and disadvantages. The large number of users is helpful in statistical terms, but the volume of data can inhibit the move from quantitative aggregates to qualitative detail. While we sometimes used common database query tools to work with the ‘raw’ log data, we found it beneficial to also develop our own visualisations to better understand patterns and detail in the data and to choose where to focus requests for interviews. Compared to more traditional trials that involve local participants who are paid to use our software in a trial, we suggest that our process of ‘recruitment’ led to more realistic conditions in that users were using software that they themselves chose to use—without inducement from us or obligation on their part to keep using it even though they did not want to. However, this advantage has to be weighed against issues such as our inability to gather data from those who dislike the application, and our reduced knowledge of local context and culture.

In practical terms, our methods incurred expenses in terms of development time and interviewer effort. The language skills of the group were put to the test as we created French, German and Japanese internationalisations—giving greater access to those for whom English is not their first language. Initial worries about our ability to interview users with limited English and a first language outwith the skill set of the team proved to be irrelevant, however. The nature of the interview selection process meant those who were not confident in their language skills were less likely to volunteer to be interviewed. Again we note a potential bias: potentially significant interview subjects could decline to be interviewed due to their lack of confidence.

Communicating across time zones can cause delays and sometimes involved the scheduling of out-of-hours interview times in order to fit in to the daily schedules of our users. Taking into account the time differences when considering the rapidity of response from users is another aspect; users generally expect ‘timely’ responses to any messages they send—no matter how many time zones away from the developers they are. We found that taking careful note of the sender’s time of day when a message was created, and addressing their perception of the passage of time until we responded, was important in building relationships with users, e.g. a reply which will not be read until the ‘next day’ in the user’s time zone should be phrased to take into account the user’s likely perception of a slow response.

In our trial we found that relative wealth scales also played a part, with the level of entry to our trial set at having an Apple iPhone—still a relatively expensive item which is not price-normalised to match local incomes. In rough terms, and taking into account countries’ populations, we observed that as the average income of a country decreased so did the density of Yoshi users there. We suggest that this pattern may not appear with software for more widespread, price-normalised mobile phones—potentially leading to a larger proportion of users in countries with lower average incomes taking part in trials. Similarly, although the trial software was developed on the latest iPhone hardware, firmware and OS, care was taken to ensure that the game was backwards compatible with older versions to try to maximise potential user-base. In practical terms this meant compiling for the earliest possible OS version and ensuring that features relying on later OS versions degraded gracefully.

As explained in Section 4.1, users were prevented from starting the game without first stating that they had read and understood the terms and conditions, which explained the nature of the trial and the data that would be logged about their use of the system. However, when speaking to participants, it emerged that none of those interviewed had understood the game was part of an academic trial. A task was subsequently presented to users, further explaining the nature of the trial and asking whether they had understood this, with 70% responding that they had not. This mirrors findings by the Federation Against Software Theft [16] where the percentage of users who reported reading EULA’s on the desktop was 28%, with 72% routinely agreeing to them “without taking any notice of exactly what they are agreeing to.” This highlights a potential ethical issue for all researchers distributing software in this manner as, opposed to a traditional face-to-face handover where participants’ understanding can be gauged and explanations repeated or re-worded as necessary, the understanding of remote participants is assumed on the basis of clicks on a checkbox and can only be verified after they have become involved in the trial.

6 Conclusions and Future Work

We have described running a worldwide trial of Hungry Yoshi, and the means by which this application was distributed to users, on a scale beyond that usually found in ubicomp field trials. A central aim was to push the upper limit on the number of participants as far as we could while still combining quantitative and qualitative approaches in ways that usefully and efficiently fed into the redesign process. We used a distribution method that made the system available to the general public, comprehensive system logging, a means of interacting with users that was integrated with the user experience of Yoshi, and interaction via a social networking web site. The benefits of such mechanisms include a significant reduction in the effort and monetary cost of field trials—particularly the cost of devices for such field trials—as well as an increase in the numeric and geographic scale of the user-base.

The worldwide nature of the trial meant that we had to adapt our tools and methods to maintain awareness of participants. We described how we used quantitative and qualitative assessments to assess the activity and engagement of our user-base, and how we used this to perform targeted interaction with participants, how that interaction took place on a variety of scales, and how we embedded feedback mechanisms within the system and encouraged their use. The Facebook social networking site served as a means to contact users, to give them awareness of other users’ activity, and as a means for them to interact with each other. In combination, these features let us run a trial involving a very large number of participants for a long period of time, and yet have relatively quick redesign cycles set within that process. We offer these summarising points for researchers taking a similar approach:

- *Expect low percentages of uptake and participation.* Software on mobile devices has become a ‘disposable’ form of entertainment; expect your software to be treated in the same manner as any other.
- *Be inclusive.* In order to maximise user engagement, lower technical and social barriers to participation not relevant to research issues.
- *Stay in the application.* Communication within the bounds of the application is more acceptable to users, and therefore achieves a much greater response rate. We found an order of magnitude less participation for every step ‘out of the game’ users were asked to take.

In our future work, we will be exploring further ways to enhance users’ engagement in reporting problems, proposing new design suggestions and discussing game play and game development. We will offer means for a user to use Facebook to gain access to data collected from his/her device, and consequent analyses and comparisons, and thus create a resource to change his/her own system use and to participate further in the design process. We are also aware of the way that some of our mechanisms may be particular to games, such as tasks that users are motivated to carry out for game advantage. We are therefore considering how to generalise this mechanism to other application areas. Finally, we are exploring breaking up our applications into software components that can be flexibly combined, so as to support finer-grained updates and to support users' customisation of software configurations appropriate to their contexts, preferences and behaviours. We see such customisation as particularly appropriate given the variety of contexts and uses that become open to study as techniques such as those presented in this paper allow us to increase the scale of system trials beyond the limits of currently standard methods and techniques.

Acknowledgments. We thank Louise Wright for translation services, and the other members of the Social/Ubiquitous/Mobile Group for their collaboration: John Ferguson, Phil Gray, Stuart Reeves and Scott Sherwood. This research was funded by UK EPSRC (EP/F035586/1).

References

1. Rogers, Y. et al.: Why it's worth the hassle: The value of in-situ studies when designing UbiComp. Proc. Ubicomp, Springer LNCS 4717, pp. 336—353 (2007).
2. IDC, Worldwide Converged Mobile Device Market Grows 39.0% Year Over Year in Fourth Quarter, http://www.idc.com/getdoc.jsp?containerId=prUS22196610
3. Zhai, S., et al.: Shapewriter on the iPhone: from the laboratory to the real world. In: Proc. ACM CHI Extended Abstracts, pp. 2667-2670 (2009).
4. Bell, M., et al.: Interweaving Mobile Games with Everyday Life. Proc. ACM CHI, pp. 417-426 (2006).
5. Hall, M., et al.: Adapting Ubicomp Software and its Evaluation. In: Proc. ACM Engineering Interactive Computing Systems, pp. 143-148 (2009).
6. O'Neill, E. et al. Instrumenting the city: developing methods for observing and understanding the digital cityscape. Proc. UbiComp, Springer LNCS 4206 pp.315-332 (2006).
7. Reades, J., et al.: Cellular Census: Explorations in Urban Data Collection. Pervasive Computing, 6 (3), pp. 30-38. (2007)
8. Chin A.: Finding Cohesive Subgroups and Relevant Members in the Nokia Friend View Mobile Social Network. Proc. Social Computing, pp. 278-283, (2009)
9. Licoppe, C., and Inada, Y.: Emergent Uses of a Multiplayer Location-aware Mobile Game: the Interactional Consequences of Mediated Encounters. Mobilities 1(1) pp. 39-61, (2006)
10. Crabtree, A. et al.: Supporting Ethnographic Studies of Ubiquitous Computing in the Wild. Proc. ACM DIS, pp. 60-69 (2006)
11. Froehlich, J. et al.: Voting With Your Feet: An Investigative Study of the Relationship Between Place Visit Behavior and Preference. Proc. UbiComp, 333-350, Springer (2006)
12. Carter, S., et al: Support for Situated Ubicomp Experimentation. In: Proc. ACM CHI, pp. 125-134 (2007)
13. Morrison, A., et al.: Using Location and Motion Data to Filter System Logs. In: Proc. Pervasive, pp. 109-126 (2007)
14. Common European Framework of Reference for Languages, http://www.coe.int/T/DG4/Linguistic/CADRE_EN.asp
15. Facebook Statistics, http://www.facebook.com/press/info.php?statistics
16. FAST, Federation Asks: Do you know what you're agreeing to?, http://www.fastiis.org/resources/press/id/304/