Mass Participation User Trials

Donald Calum McMillan

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Computing Science

College of Science and Engineering

University of Glasgow

June 2012

Abstract

This thesis investigates how researchers can take advantage of the rapid adoption of mobile technology that has brought with it transformations in social and cultural practice; the expectations of what computers are, what they can do, and the role of digital objects in everyday life.

In particular this thesis presents and discuses the use of new App Store style software distribution methods to reduce the cost, in terms of researcher time and hardware, of recruiting a large group of participants for a trial 'in the wild' while increasing the potential diversity of users is becoming an attractive option for researchers pursuing the ubicomp vision. It examines the procedures for running large scale trials with the deployment of three applications released to a combined user base of over 135,000 in such a way as to keep the qualitative detail necessary to inform design while gaining the diversity of users for claims of generalisability. More generally, it discusses the results that can be expected from this 'mass participation' approach, and the ethical responsibilities they place upon researchers.

The contributions of this thesis for mobile HCI show that in large-scale trials, relatively rich qualitative data can be collected along with substantial quantitative data, and that a hybrid trial methodology combining a large-scale deployment with a local trial can be a powerful tool in addressing shortcomings of trials that are either solely local or solely global.

This thesis also contributes guidelines for researchers running large-scale user trials that give consideration to the established research norms and practices, in an attempt to strike a new balance between invasiveness and utility.

Acknowledgements

Having been lucky to be part of the Social Ubiquitous Mobile group at Glasgow, I have been fortunate to work collaboratively with a number of excellent researchers, and had the opportunity to present work throughout the world.

Specifically, I would like to thank the group members past and present with whom I have collaborated over the past 4 years; Dr. Alistair Morrison, Dr. Malcolm Hall, Dr. Marek Bell, Dr. Scott Sherwood, Dr. Barry Brown, Dr. Julie Maitland, Dr. John Ferguson and Owain Brown.

I would like to offer special thanks to my supervisor and the leader of the SUM group, Dr. Matthew Chalmers, for his support, guidance, and motivation through the course of this research.

I would like to add another thanks to Dr. Malcolm Hall, for persuading me to start the process of moving half way around the world and entering academic research in an incredibly eloquent 11 characters.[1]

Finally, I would like to thank my family. Donald and Carolyn, my parents, for always encouraged and supported me. My brother and sister, Alexander and Samantha, for taking the slack and allowing me to concentrate on my thesis. My grandmother, June, for all her help and advice. My aunt, Ngaire, and two cousins, Curran and Asha, for making me smile and reminding me that no matter how much I was struggling I could either be, or be the parent of, a teenager.

Credit must also be given to Bruce Peninsula, Professor and Maryann, Sara Vaughan, Billie Holiday, The Amsterdam Royal Concertgebouw Orchestra, Modest Mouse, Chet Baker, Isobel Campbell, Mark Lanegan, The Folk Implosion, Alabama 3, and Sons and Daughters for helping keep me sane during the writing process.

The work described within this thesis has been funded by EPSRC grant EP/F035586/1.

[1] Dr. Donny? - M

I dedicate this thesis to all my imaginary friends.
Thanks for keeping me sane, guys!

Table of Contents

List of Figures

Associated Publications

- A. Morrison, D. McMillan, S. Reeves, S. Sherwood, and M. Chalmers, "A Hybrid Mass Participation Approach to Mobile Software Trials," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1311–1320.
- M. Chalmers, D. McMillan, A. Morrison, H. Cramer, M. Rost, and W. Mackay, "Ethics, logs and videotape: ethics in large scale user trials and user generated content," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*. ACM, 2011, pp. 2421–2424.
- A. Morrison, O. Brown, D. McMillan, and M. Chalmers, "Informed consent and users' attitudes to logging in large scale trials," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*. ACM, 2011, pp. 1501–1506.
- D. McMillan, A. Morrison, and M. Chalmers, "A Comparison of Distribution Channels for Large-Scale Deployments of iOS Applications," *International Journal of Mobile Human Computer Interaction*, vol. 3, no. 4, pp. 1–17, 2011.
- D. McMillan, A. Morrison, O. Brown, and M. Hall, "Further into the wild: Running worldwide trials of mobile systems," in *The 8th International Conference on Pervasive Computing*. ACM, 2010, pp. 210–227.
- A. Morrison, S. Reeves, D. McMillan, S. Sherwood, O. Brown, and M. Chalmers, "Mass Participation in Evaluation and Design," in *Proceedings of Digital Futures 2010*. BCS, 2010, pp. 1–2.
- M. Bell, S. Reeves, B. Brown, S. Sherwood, D. McMillan, J. Ferguson and M. Chalmers, "EyeSpy: supporting navigation through play," in *Proceedings of the 27th international conference on Human factors in computing systems*. ACM, 2009, pp. 123–132.

Chapter 1

Introduction

The growth in the popularity of smart mobile devices has been rising as swiftly with their power and usability. In 2006 the number of mobile phones shipped worldwide topped 1 billion [9] with 64 million of them smartphones [14]. In 2011 the number of smartphones shipped was over 488 million, including tablets this number rises to 551 million mobile networked devices — 200 million more units than the combined global shipments of desktop, laptop and netbook PCs [17].

Smartphone ownership had risen to 42% of mobile subscribers in the USA and 44% in Western Europe by the end of 2011 with the UK and Spain above 50%, and is set to continue increasing [200].

While still not the predominant type of handset globally, with the number of smartphones shipped forecast to reach 1 billion units annually by 2016 [15] it can be seen that smartphones have been adopted into mainstream use and are no longer the solely for the highly technically competent or the early adopters.

During this time the ease with which end users can find and install new software has increased with mobile 'app stores' allowing software to be discovered, installed and reviewed without needing access to a PC. This has not only influenced the type of application being developed for smartphones, with the relatively low price points of mobile applications — or 'apps' as they have come to be known — leading to consumers devaluing apps and creating a 'throw away' app culture [40], but has also built demand to create a global app market across multiple platforms that, for 2011, consisted of over 1 million applications, 30 billion downloads and US$6 billion in revenues [238].

This rapid adoption of mobile technology has brought with it transformations in social and cultural practice; the expectations of what computers are, what they can do, and the role of digital objects in everyday life have evolved considerably [72], leading many researchers in ubiquitous computing to voice the opinion that the mobile phone should be the primary platform for deploying new information services [217, 190].

Mobile applications and services were integral to Weiser's vision of ubiquitous computing (ubicomp) as first outlined in the scenario of Sal's computer supported weekday [246]. However, Weiser envisaged computation disappearing into the walls and furniture with interaction achieved through a variety of dumb or semi-dumb devices. The wealth of storage, bandwidth and processing power that has found its way into our pockets today is opening the door to aspects of that initial vision with a less dramatic change in infrastructure than was initially thought.

Using the new App Store style software distribution methods to reduce the cost, in terms of hardware, of recruiting a large group of participants for a trial 'in the wild' while increasing the potential diversity of users is becoming an attractive option for researchers pursuing this vision.

1.1 Research Context

The research presented here has been done within the Social, Ubiquitous and Mobile research group at the University of Glasgow as part of the Contextual Software[1] project and in collaboration with the Designing the Augmented Stadium[2] project.

The Contextual Software project's aims were to make user experience, design and evaluation in ubiquitous computing more integrated and synergistic than before and in doing so address two key problems in ubicomp: adaptation and evaluation.

This goal was achieved by building and trialling systems using Mobile Ad Hoc Networks (MANET) in collaboration with the EU Haggle project [192] and software component infrastructures to support swift adaptation, treating developers, evaluators and users as collaborators in the ongoing process of increasing the software's fit with the context of the user.

[1] EPSRC, EP/F035586/1, June 2008 – Dec 2011, £869K
[2] EPSRC, EP/E04848X/1, Oct 2007 – Mar 2010, £532K

Headed by Matthew Chalmers, the Contextual Software project employed five RAs: Marek Bell, Malcolm Hall, Alistair Morrison, Stuart Reeves and Scott Sherwood, and two PhD students, Owain Brown and the author.

The Designing the Augmented Stadium project was based around the concept of a new form of stadium that connects spectators, the game being watched and those watching at home. This project combined ethnographic studies with the design, building and trial of prototype systems. The ethnographic component explored in depth the nature of sport spectating and participation, documenting a new area for technology design. Design concepts drawn from the studies had the aim to experiment with a role for stadiums not simply as containers of spectators, but as an augmented environment that support the active participation of spectators in making watching games an enjoyable, sociable and valuable experience. These concepts were built and tested in with spectators, at a host of live events both in the stadiums and in other venues where fans were socially enjoying the act of spectating.

As the investigation into developing systems to support spectators continued, the size of spectator groups led the research towards supporting more and more spectators as participants in an attempt to enhance these wide, varied and complex social events. This led development in two directions. The first of which was to develop systems which afforded interaction from large numbers of users at once.

This resulted in the development of the Digital Flag [173], a bluetooth and GPS enabled artefact that a lead supporter would carry during the pre-match buildup, into the stadium, then to the post-match celebrations or commiserations. This device not only scanned for bluetooth devices to generate density maps of crowds, later versions allowed supporters to send images and text messages to the device which would be opportunistically uploaded and shared with other fans online.

In the other direction there was an attempt to put new affordances into the hands of more spectators, done with the development and distribution of the application FanPhoto [42]. This application allowed groups of fans to share a communal bank of photos to which any of them could post comments or add new images.

Deployed to 4 different groups of local participants in trials running 2 and 4 weeks, these 20 participants where each supposed to be given one of 10 project iPhones. But at that time the quickly rising popularity of the iPhone

made them difficult to get hold of, only two were able to be bought at any one time with a long lead time on delivery and a proscription on overlapping orders.

However that same popularity began to turn into an opportunity as members of the prospective participant groups started to take delivery of their own iPhones, allowing the software to be deployed and the trial to start earlier than had been feared possible.

This experience made an impact on the proposed work-plan for the newly starting Contextual Software project.

One goal of the project was to help make systems more adaptable, seen as crucial to ubicomp's progress because contexts, needs and uses are often more dynamic and subtle than any designers' predictions can fully encompass when, for example, mobile users are interacting in the uncontrolled environment of city streets.

It was recognised that it is prohibitively difficult for evaluators to be with users all the time in their everyday lives, observing and recording where they go, what they do, and their interaction with people nearby and – via networked mobile devices, for example – other people in quite different contexts. Another goal of the project was to develop practices and procedures to allow evaluation of software in those uncontrolled environments and to understand users' changing contexts and uses in a manner both practical for the researcher and valuable in the timely design of new components or configurations.

A final goal of this project was to demonstrate that software developed using a component infrastructure and supported by evaluation of its changing use and the user's changing situation would be able to adapt to a user's context and better provide for their needs. The plan was to demonstrate this within a number of user trials conducted in the same manner as those of FanPhoto, lending participants iPhones for a number of weeks at a time. However with more and more of the devices being used within the group for research finding their way into the pockets and purses of those that had been seen before as potential trial participants, the opportunity to recruit them all presented itself.

With a potentially much larger number of participants, the number of different identifiable contexts available to study would increase, as would the number of users in the most popular — potentially leading to more data and a greater understanding of those contexts into which the software was to be

adapted to fit.

Many of the same problems for evaluators identified for the originally planned trials would still need to be faced using participants recruited from the App Store, such as the challenge of understanding a context of use and a user's intention without direct observation.

New challenges, both ethical and motivational, presented themselves with respect to the lack of a researcher-participant relationship. However, the perceived benefits of greater generalisability and a better understanding of the nature of contextual adaptation caught the imagination of the group.

The first question that had to be addressed was how, and indeed *if* the quantitative and qualitative data collected from a remote participant set could give enough information to drive the generation of new components and the adaption of the software, in an inclusive way we call 'mass participation'.

Conducting a large scale, mass participation user trial did not preclude subsequent, or simultaneous trials with a local participant set. This begged the question of whether the increased cost, in data plans for trial devices and in researcher time, of conducting a pair of trials side-by-side would provide a worthwhile return in terms of actionable understanding of use and context.

The responsibilities of the researcher towards their participants in this type of trial was also undefined. The opportunity presents itself for detailed, long term logging of user activity leading to a modelling of participants' current context to a fidelity possibly greater than they would be willing to actively share. With the way that the information is collected and transferred given to the researcher is much less apparent than in traditional internet mediated research, this combined with the erosion of the traditional researcher participant relationship built up through recruitment, briefings and interviews raises many ethical questions.

However this wide range of unknowns was able to be distilled to the two research questions that follow:

> **RQ1:** Can qualitative and quantitative data be effectively collected, analysed and fed back into the redesign process from a large scale user trial?
>
> **RQ2:** What ethical issues arise in the approach to answer RQ1, and what ethical guidelines should such research follow?

These questions are used to structure and show the relevance of the research contributions of the thesis.

The first question encapsulates issues related to practical — often technological — aspects of the design and evaluation of large scale ubicomp trials.

The second question encapsulates overarching ethical concerns that should also lead to practical action by designers and evaluators, but which are discussed separately in this thesis. This apparent partition is done for presentational and discursive reasons, rather than to suggest that the methodological issues and concerns related to each of these questions are cleanly distinct or independent.

1.2 Thesis Structure

Following this introduction the Background chapter (Chapter 2) describes the movement of mobile user trials from the lab to the app store (§2.1 & §2.2), the use of games as a research vehicle (§2.3), and discusses the use of ethnography, ethnomethodology and activity theory to methodologically ground user studies in this area (§2.4).

Chapter 3 describes the environment in which the applications in subsequent chapters were released, with an exploration of the processes and consequences of releasing through the traditional app store or through a third party repository.

Chapter 4 covers the release of the mobile location based game Hungry Yoshi, and the processes and features built into the design and evaluation of the software to allow researchers access to higher fidelity qualitative data than would be possible when relying on usage logs and store-based reviews alone.

Following on from this, Chapter 5 examines running a hybrid trial – concurrently investigating a local participant group in a traditional manner and using them to inform, and be informed by the global remote participant group.

The ethical challenges faced by researchers when running such trials are explored in Chapter 6. Problems arise when the traditional dialogue between researcher and participant that negotiates understanding of the nature and purpose of the trial is replaced with online or in-application pre-

sentations of text. The nature of these problems and possible steps that researchers can take to mitigate any potential ethical conflicts are discussed.

The final chapter summarises the findings of the intermediate chapters, shows their relation to the research questions above and outlines areas for further research.

1.3 Division of Labour

The systems and trial described in Chapter 3 were primarily created and conducted by the author with assistance from Alistair Morrison.

The re-implementation of Hungry Yoshi for iOS described in Chapter 4 was led by the author with input from Alistair Morrison, Marek Bell, Malcolm Hall and Owain Brown. The trial was run by the author and Alistair Morrison with assistance from Owain Brown.

The initial pilot study conducted in Chapter 5 was conducted by Stewart Reeves and Scott Sherwood with assistance from the author in both the implementation of the system and the group interviews. The implementation of the release version, including the additional peer to peer functionality, and the trial were conducted by the author, Alistair Morrison and Scott Sherwood.

The ethics workshop proposal was drafted and coordinated by the author. The workshop was ran by the author, Henreiette Cramer and Wendy Mackay. The results of this workshop form the inspiration for Chapter 6.

Chapter 2

Background Literature

In this chapter there are several areas of background work that will be explored. In the first section, the tradition of user trials in HCI and CSCW will be charted, explaining the move that current research has taken into the large-scale user trials examined in §2.2. As noted in the introduction this field is in its infancy, with the work covered in Chapter 4 being based on one of the very first publications in the area. The work gathered in this chapter may seem fragmented simply because it is a collection of exploratory 'first steps' into this field by researchers from many different backgrounds.

§2.3 explores the use of games, specifically mobile games, as vehicles for research into topics other than games and gaming. This section covers the properties that enhance the researchers' opportunities, such as engineering engagement and participation being a reward in itself as well as those that can work against the researcher like raised expectation of participants and the difficulties of designing a compelling game mechanic.

§2.4 begins by charting the birth of ethnography, how it is carried out within HCI and the issues around the adaptation of these methods to suit large-scale trials. There is then a consideration of the merits of using ethnomethodology, cultural historical activity theory or a combination of the two in the interpretation and analysis of the data collected from large scale trials in order to inform future research and feed back into design.

2.1 The Evolution of User Trials

As computation has moved into more and more areas of people's lives, trials to understand the use and usability of these devices have followed – even

into the most hostile environments, such as burning buildings [65].

Evaluation of the use of ubiquitous or mobile computing systems has, as recommended by [1], continuously expanded in scope to encompass conducting evaluations outside of the laboratory and in the wider world, with all the complexities and challenges that brings. There have been arguments against the utility and cost-effectiveness of this move where there is a possibility to construct a more 'life-like' laboratory setting that incorporates the important aspects of context and mobility as identified by thorough field work prior to designing any system under investigation [143]. However when there is a desire to understand how novel pervasive technologies are appropriated in real world settings, and where the level of mobility is greater than a laboratory setting can comfortably contain in-situ, or 'wild', studies can add great value [211].

In contrast to the lab-based environment of more traditional studies, it has been argued that experiments carried out in situ in this manner 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.

A great deal of data can be gathered from creating 'lifelike' situations within the lab to evaluate systems, such as at NCR [35] where the lab was made to resemble an office workplace or NTNU [221] where the researchers spent time simulating a hospital environment for the testing of a patient record system. This method is widely used to gain high fidelity data on usage and usability, with usability labs available for rent to companies in major cities around the world from companies such as Cimex[1] and in use in a number of universities, however as the behaviour of new mobile systems incorporate more and more contextual and social cues these laboratories can begin to constrain the user.

Another tactic was taken by research institutes such as Xerox PARC — here the lab was expanded to encompass the office environment of the researchers working there and the technology distributed and embedded throughout. This gave the opportunity to observe use on a larger scale, both spatially and temporally, and in the more 'wild' environment of a working office environment without the costs of supporting an offsite deployment in a less technically tolerant environment. With this move necessarily came a loosening of the researchers' control of the environment in which the re-

[1] www.cimex.com

search took place.

With the release of more powerful hand held devices like the iPaq and the Nokia N70, researchers were able to move their trials out of the laboratory and into 'wild' environments prepared in advance with the infrastructure necessary, such as data connectivity and recording devices as in REXplorer [20], George Square [41], Voices of Oakland [74] or Treasure [50].

As the devices became less expensive and more user-friendly, the opportunity presented itself to give users more time to adjust to the new technologies presented to them by researchers such as Feeding Yoshi [29] which gave the users 3 weeks to use the devices and Mopix [155] in which the 10 devices were left in situ for 2 months.

While this approach has its benefits, the staging of ubicomp and mobile system trials in this way can 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 familiarisation 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. Trials such as that of the system MoMENTus [152] used users' own devices with local deployments to raise the number or participants to 43 over one week and would have avoided this problem.

Another limiting factor that arises when providing participants with 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.

2.2 Large Scale Studies

With the aforementioned sudden and ever increasing market penetration of smartphones and tablets, incorporating many of the esoteric sensors seen only a handful of years ago in nothing but lab-built devices as used in [149] or low-volume off-the-shelf research hardware such as the SHAKE [122], researchers have been able to increase the numerical and geo-cultural scale of their trials to levels unattainable by conventional means.

Some applications were released to explore what to expect when releasing applications, for research or otherwise, through these app stores in the manner of ZooEscape and Packer, discussed in Chapter 3.

Riccamboni et al. released 5 localised botanical identification apps [207] for iOS covering different locations: two countries, a geographical region, a nature reserve and a botanical garden. The apps covered different prices points, from free to €10 to determine the interest and the market for such reference applications. They found that for the highly localised applications the demand was small, unless the app was free in which case users from outside the area would browse the application. The nation wide keys for identifying flora and fauna had high demand irrespective of the price. No analysis was done beyond counting the number of purchases per application.

In [242] a free, commercially developed application for a leading Swiss insurance firm was used to deploy logging that allowed the researchers to examine the location accuracy of iOS devices. Using 2,289 location logs gathered in a four month period in 2010, they were able to compare the accuracy of the location data returned by two generations of iPhone, iPod Touch and the initial release of the iPad. The researchers were able to show that the later generation iPhone 3GS gave a lower accuracy in urban areas than its predecessor the iPhone 3G.

Other research applications have been released to a wide audience without taking full advantage of the number of users they attracted to deepen the understanding or widen the scope of their conclusions.

One of the earliest large-scale deployments of a ubicomp application was Mogi Mogi [153]. This location-based mobile multiplayer game was released commercially in Japan, and in 2004 had roughly 1000 active players. The distribution was done through a corporate partnership with the mobile phone carrier KDDI in Japan and growth relied upon word of mouth, and the biannual month-long no-cost promotions that the carrier ran to increase uptake of many of the applications they offered. The application cost ¥210 per month to play, not including data costs. The presented analysis is based the interviews and log data of ten players who were selected to fit within the demographic spread of the overall user population. This method afforded rich detail of the ways that the game fit into their urban contexts and lifestyles of players, based on months of game play and social interactions between players within the chat function of the game. However the large number of users was not leveraged to inform the analysis as it is presented in [153] beyond informing whom to recruit for interviews.

Shapewriter [254] was released, very soon after the Apple App Store was launched, to test a novel form of text input. While the focus of the paper is on the reviews written by end users about the software, the approval delay for submitting to the App Store is mentioned, as is a link between a positive blog entry on Time's website and the number of downloads they achieved, peaking at 30,000 per day. However the reported analysis focuses solely on the 556 reviews left by users on the App Store and doesn't fully take advantage of the high user numbers they achieved by, for example, statistical analysis of system log data and/or in-depth interviews with users.

Museek [146] is a music player, released on Android, to analyse and arrange a user's music library by similarity. The evaluation was based upon

quantitative data gathered from 128 highly engaged users. As the algorithm takes time to learn about the preferences and collection of a user the majority of users who browsed the application were not useful subjects. Qualitative data was collected from the reviews posted on the store to back up the claim that as 64% of the use of the heavy users took advantage of the app's content derived playlists, those features were "very accepted" by users. The validity of using voluntary, public comments to explain the use of the most engaged users without any correlation between the users that provided feedback and the those that provided data is not discussed.

Text Text Revolution [212], is a game that helps users improve their typing on mobile touchscreen keyboards by providing targeting practice, highlighting areas for improvement, and generating targeted training data while they are playing the game. Although released through the Android store and downloaded more than 25,000 times the effectiveness of the application in improving typing skill was only evaluated using six local participants, the release and download statistics are used only to verify that the game is engaging – and therefor could, in future work, be used to provide data to inform a predictive key-resizing algorithm for use at a large scale.

Mirage Money [95] was released on Symbian phones, testing the use of accelerometers to detect tilt as input to a 3D game. Although released in 2008, before Symbian's app store launched in 2009 [138], it achieved 10,000 downloads by being distributed to a number of freeware websites. Despite the large number of users, the only publication of an analysis of this system [52] is based upon a local trial with 11 observed and interviewed participants and focused upon the ease and 'fun' of the interaction technique and the game itself.

Ocarina [245] is an iPhone-based instrument allowing users to play music and listen to others playing around the world. This highly polished application focused upon the novel interaction of blowing on the device, while the accelerometer and the position of the fingers upon the virtual keys manipulated the tone. It was also released very soon after the Apple App Store was launched and quickly gained over a million users with 20 million snippets of music shared. Detailed analysis of the use or distribution was not the focus of this release. Rather, the number of users was used to reinforce the claims that the application provided a compelling and expressive music-making experience, that identity is not crucial to building a community, and that with the right approach and application it is possible to create a global community 'overnight' and encourage a large population to

engage in expressive music making. The Stanford Mobile Phone Orchestra has since moved away from large deployments towards performance based studies [184] before adding credence to such claims.

Walk'n'Play [44] measured the amount of calories burnt by a user by determining the number of steps taken using the accelerometer on the iPhone combined with several self reported variables that allowed the calculation of a Body Mass Index. With 6,000 users out of 8,000 showing continued engagement, it proved to be an incredibly 'sticky' application in comparison to the others discussed here. While the publication does not directly take advantage of the number of users the authors state the intention to explore using the usage data collected to inform interventions with the goal of increasing the health of their participants.

Using a global release to gain a large number of users from whom enough data on the use of the application can be gathered and analysed to find, prove or disprove hypotheses has been a popular mode of research in this area.

In 2008, Nokia Research Centre released Friend View, a "location-enhanced microblogging application and service" [54] via Nokia's Beta Labs. The authors report on statistical analysis of social network patterns based on anonymised log data representing 80 days' use by 7000 users. In the analysis of this trial the volume of data was used to examine the type and frequency of user activity with respect to the structure of their friend network.

Product Empire [43] is a social network game that motivates users to scan barcodes, enter corresponding product information, such as product name, brand, category and to upload a picture of the product. The are also requested to verify other users' uploads. This Android application was released in 2010 and, within 17 days, 244 users generated more than 990 product descriptions and cross-checked product data 1230 times. The quantitative analysis of Product Empire focused upon aggregate data of the types of products scanned, the number of interactions performed and the location of the users. Qualitative data was collected from 15 store-submitted reviews, an email feedback form used for suggesting updates to the game and, the researchers manually scouring the Facebook accounts of top users for publicly available data. They note that by requiring users to log in using Facebook their accountability was increased to the point where less than 1% of product descriptions that were entered were deliberately wrong. However

this may have been an artefact of the game mechanic rewarding 'correct' information through the verification procedure.

My2Cents [134] is a mobile application for reading and sharing comments and ratings on retail products by scanning their barcodes on a mobile device. Initially released for Android, an iOS version was made available for download shortly afterwords[2] and the cross-platform system had gathered 2656 comments on 1640 products by May 2012. The initial publication focuses on 1122 active installations of the Android version, looking at the aggregate data on the scanning of, and commenting on products by users. The qualitative analysis of the 16 reviews the application received in the store was used to base the conclusion that the application needed polish in its interaction design, aesthetic, user motivation and reliability to build a large community of users. However, the authors note that planned steps include more detailed user studies to evaluate the app.

Bargain Finder [133] is an Android utility application released by Comparis, a national provider for price comparison of services and products in Switzerland. The application allows the user to see a list of current top bargains for local retailers ordered by the percentage of potential saving. The team that produced My2Cents negotiated access to six months of usage logs covering more than 33,000 users, making over 1.6 million geo-located requests for product information. They used web usage mining processes and the Knime[3] analytics software to analyse user sessions, with the variety of the interactions providing a 'focus' metric and highlighting different patterns of use between app browsers and heavy users. They conclude that this approach holds "little value for research" as they were unable to analyse the determinants for using the app and the effects on shopping behaviour. The authors suggest expanding the information logged and combining it with survey data to provide a clearer picture of the use of the application.

Twiphone [98] is an application for Android mobile phones which shares a user's call logs and text messages on Twitter. The 2700 downloads in less than two weeks equated to an active installations count of 372 at the time of publication, with 316 users sharing at least one communication and 22 sharing at least one communication every two days. The analysis looked at what was shared by, and to, whom. However, the authors note that in order to gain an understanding of why users share, and how acceptable this is to the wider community, they plan to perform qualitative interviews.

[2] www.my2cents.mobi
[3] www.knime.org

Halos [112, 116] compared visualisation techniques for off-screen objects using a release on the Android Marketplace to conduct a 'controlled' experiment comparing three conditions, exploring how this type of trial and deployment can be used to complement traditional lab based HCI experiments. This application was released in April 2010 and within 10 weeks had recorded over 5,000 installations, with useable data being returned from 3,934 users. They also note that this application was run on 40 different devices with varying versions of the mobile operating system — as, for most users, upgrades to the Android operating system are initiated by the mobile carrier at its discretion. In this trial, the large number of users was used to compare the performance of the three different conditions on two different mobile devices. The authors note that one needs a large data set for each condition to smooth the problems with running such an experiment without being able to control the environment or ensure that users complete the task properly. This caused the authors to discard the data received from 38 of the 40 device types they encountered due to the low number of data points.

PocketNavigator [197] is a navigation application that gives coded vibrations as the user reaches a waypoint to indicate which direction should be taken at the next turn. During the release of PocketNavigator, published shortly after the study of Hungry Yoshi discussed in Chapter 4, the authors identified similar issues with recruiting users in competition with commercial applications, analysing the data, and the issues with experimental validity.

World Cupinion [213] was an Android application released and evaluated alongside the 2010 FIFA World Cup in South Africa, as was the World Cup Predictor application discussed in Chapter 5. World Cupinion allowed users watching a live football match to share one of nine emoticons relating to the refereeing decisions on the pitch and their enjoyment of the action. The number of presses each button received by co-supporters and those supporting the opposition was displayed to the users. A web based version was also available which accounted for 29% of the 925 users seen over the 64 matches in the World Cup. Use location and the buttons pressed were logged for analysis. This was enhanced by an optional survey on their viewing habits around football and the application, which 46 users participated in. The timing of usage during a proportion of matches was mapped to the timed updates from online sports feeds to show that the audience saw the same events as important as the professional commentators did, although the influence of them listening to commentators while they used the ap-

plication was not discussed. It is noted that while the conclusion is that users felt more connected to other fans and had more fun using the application than they would have otherwise, the self-selection nature of the post-experiment questionnaire where users who did not like the application would simply delete it and never be presented with the survey makes it harder to claim generalisability.

Hit-It [117] was used as a delivery mechanism to put six different trials into the hands of end users, and shows the research benefits that can be gained by culturing a strong user-base. This is a game where users tap moving targets as quickly and as accurately as they can. The validity of Fitts' Law for mobile touch screens [114] was tested across 992 users on a 4.0in touch screen device and found to give implausibly high indexes of performance (IP) with weak correlation of the data. In order to test the hypothesis that this was due to the small size of the screen meaning that the user only had to move their finger as opposed to their whole hand the calculations were also carried out using data from 352 users of a 7.0in tablet device but this still resulted in high IP and low correlation leading to the conclusion that Fitts' Law is not a good model for this task.

The data from Hit-It was also analysed to determine the error rates for different target sizes and screen locations [117]. The amount of data enabled meant that they were able to show that touch positions are systematically skewed. In a follow up experiment in the same paper, a compensation function that shifts the users' touches to reduce the amount of errors was derived from the data and evaluated by publishing an update of the game. This experiment reported that from the 15,326,444 touch events collected from 12,201 installations the function reduced the error rate by 7.79%.

The effects that different forms of tactile feedback had on error rates was also explored [113]. By giving 27,102 different users different conditions on when the phone vibrated, every touch, correct touches, incorrect touches and never, is was shown that users were fastest when the phone only vibrated if a target was hit, but made the least errors when the phone only vibrated when a target was missed. It was seen to be better to provide no feedback than to vibrate whenever the screen was touched.

Updates were also published for Hit-It to determine how best to ask the users for consent to take part in the study with the aim to gain the most participants [195]. In this experiment, users were randomly given one of four different styles of confirmation dialog and the number of affirmative

responses counted. Unsurprisingly the dialog with only an affirmative button, the alternative being to quit out of the application, gained the highest percentage of positive responses but what was of interest was that 13% of users quit the application when presented with the dialog. A pre-checked dialog box vs. one that had to be tapped to confirm assent gave starkly different results of 57% acceptance against only 7% where the extra user input was required. While this is interesting, it does not address the problem of ensuring that the users understand the agreement they are entering into with the researcher, detailed in Chapter 6.

By combining the data regarding the times at which Hit-It was installed and run across all 157,438 recorded users with the recording of 24,647 game releases across 8 categories in a three month period [115] a peak in user activity that coincided with a drop in the rate that games are released was seen on Sunday evenings. This suggests that the greatest visibility on the 'new' list of the store, and a consequently greater number of downloads would be achieved by publishing then.

A third use of such mass-deployments of research applications has been to gain data of smartphone use out-with the application, by installing background loggers on participants' devices, for example.

One of the most long-term deployments of such a mobile research applications is Cenceme [46, 171, 172], an application that uses context sensing to automatically update social networking sites with each user's current activity. Initially developed for the Nokia N95 and trialled among 30 locally based participants, the software was then ported to the iPhone and released in July 2008 when the App Store was first launched. Cenceme went through cycles of deployment, use and refinement based upon the feedback gathered via email and through an app specific discussion board. Here it was noted that the university's ethical review board required them to add a second, more explicit notification to the users of their participation. The contrast to the notification that required the user to quit the app to decline, as was released with versions of Tap-It [195], shows the vast range of ethical guidelines and regulations across research institutions, and highlights the need for discussion and consolidation of these throughout the research community using such trials.

AppAware [96, 97] allows users to share, via existing social networks or within the application itself, location-tagged information as to which applications they are installing, removing or updating. In doing so, users are able

to explore applications popular in their current location. AppAware was released for free on the Android Marketplace and from 24,000 users recorded more than 1,400,000 installations, updates and deletions of 18,740 third party applications. AppAware calculates a rating for an app based on the assumption that if a user uninstalls an app, they don't like it. These ratings were mapped against the user star ratings given on the store and it was shown that ratings in the store were skewed to the higher end of the scale. It was also proposed that the software provides more accurate ratings where the number of downloads is low.

Appazaar [38] is similar to AppAware in that it detects users' app usage and location to recommend applications that may be of interest. With data of use from over 4,100 devices they detected over 22,000 different applications in use and examined the average amount of use of each category of application experienced, the time of day it was used and the number of apps used in a session. They found that users spent almost an hour a day using apps but less than 72 seconds with an app at a time.

AppJoy [253] is yet another Android based app recommender system designed for casual discovery of interesting applications. With 4,600 users, its recommendations were followed in 5% of cases, but where they were followed it predicted the users usage with more than 80% accuracy. While AppJoy also stores the locations of use, it has yet to incorporate this data into the recommendation algorithm.

Energy consumption and charging behaviour on mobile phones has also been investigated using large scale deployments of logging applications. Falaki et. al. [83] provided 33 Android devices and 222 Windows Mobile devices to participants with unlimited talk time and data for the course of their investigation of smartphone network and energy usage. While not distributed in the same manner as the other trials mentioned here, and incurring a significantly higher cost, the scale of this research is comparable.

Oliver [187] released a logger to investigate how users interact with and consume energy on their portable devices via the Blackberry distribution system, with 20,100 users providing over 1150 years of cumulative data, and was able to use this dataset to build an Energy Emulation Toolkit that allows developers to evaluate the energy consumption requirements of their applications against real users energy traces [188].

OverCharged [85] gave users information on their devices current battery level, for how long the phone had been running on battery and other mis-

cellaneous information, such as temperature and voltage. Logs of battery use were collected over a four week period of use by the 4,035 users who opted to give anonymous statistics out of the 17,000 people using the application at the time the study was conducted. This data was analysed to determine common charging patterns, and calculate the optimal scheduling of battery intensive tasks such as uploading large amounts of log data to a server.
Several ubicomp projects have featured data collected from large numbers of people via mass-scale sensing without deploying mobile phone software. An example is the Cityware project [189], 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 [203], 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.

The trials above that *used* the large scale of their deployments for research did so by, in general, gathering enough data from enough participants to draw statistically based conclusions avoiding the danger of local biases present in gathering enough data in a traditional manner. Most directly mention the greater confidence in the generalisability of the conclusions drawn based on the number of data points used.

Investigation into the actions and motivations of users was only shown to be done in the trial of WorldCupinion [213]. However, limitations in the collected qualitative data, such as the self-selected nature of the participants taking part in the post-trial surveys and the inability to contact users that didn't like and uninstalled the game, were identified.

In HCI research there is a tradition of employing usability labs and formal experiments to isolate variables and control for effects [2, 127]. In many cases these approaches offer the best method to answer the research questions posed, however there are a number of areas where moving beyond the lab is desirable in the pursuit of valid results. The systems described above where the research falls into one of the areas in the following list would not have provided the results they have had the researchers chose to study their system in a lab or a semi-scripted living-lab environment [201].

- Aspects that develop over time, e.g. adaptation.
- Systems requiring long term usage and habituation.
- Natural usage patterns.
- Interaction in and with context, e.g. location-based services.

The benefit of having large numbers of users using a mobile system lies in how ingrained mobile usage has become into our everyday lives [123], by changing, augmenting or simply monitoring an artefact through which so much of a person's life is organised and recorded researchers have the opportunity to understand in more depth than before, and across a greater demographic spread than before, how the technological impacts the social.

While there have been attempts to move traditionally lab based studies from the lab into the lounge, the benefits of doing so to gain large numbers of trial participants has yet to be shown to outweigh the increased costs, increased complexity of analysis and reduced confidence in the gathered data. Henze et. al.'s work with Hit-it to collect target acquisition data would generally be done in a lab setting in order to reduce confounding variables. While the touch correction matrix generated from this large data set was shown to improve the accuracy of users, it was not compared against a matrix generated from a small number local participants or a matrix generated from a small, random sample of the global users.

Until more research has been done to compare the utility of lab based studies with conducting such experiments out-with a controlled environment, large scale trials of this type of research will continuously have to justify the quest for large numbers of users.

This is still, however, an area ripe for research: the investigation of methods to gain reliable, detailed and cost-effective qualitative data from remote participants to allow greater understanding of the large volumes of quantitative data available to harvest.

2.3 Using games as a research vehicle

In choosing to develop mobile games as vehicles for research, the choice is also made to embrace the freedom to design game rules, goals and interactions in ways that exploit and evaluate features of whatever is under

investigation, be that a new interaction technique or the altering of traditional social communication boundaries.

Play is one way to learn and experiment with negotiating our cultural norms, modes of interaction and societal boundaries [121]. By its very nature this experimentation involves mutually negotiated and dynamic rules and relationships. Indeed it can be said that it is this dynamism which provides the fun [145].

This means that research exploring dynamic, negotiated systems would, if using games as its medium, allow the participants in the experiments to use all the built up knowledge they have regarding the social and practical expectations of play. Experienced game masters know that "If it is the right time, we can change anything." [64] but, conversely, those designing the game must conform to the expectations of the users — the game must be fun.

There are many identifiable pitfalls in game design but most stem from the observation that "Games have rhythms and patterns: The job of a gamewright is to reinforce patterns and dampen dissonance." [214]

Recognising and learning these rhythms and patterns is part of what makes a game fun, what makes someone play again and again, and continue to derive enjoyment from the seeming repetition. As Koster [145] notes, games will not be seen as fun if these patterns are revealed too slowly (giving the impression that the game has no depth), too quickly (giving the user no chance to learn them), or too subtly (appearing just as noise to the player).

But rules and patterns are not the only factors in fun. There are other, more social aspects of play which must be addressed to ensure continued motivation of participants in any prospective trial.

> Games aren't much fun when rules, rather than relationships, dominate the activity, when there is no attention to the 'flow,' 'fairness,' 'respect' and 'nice.' (L. Hughes, *Beyond the Rules of the Game: Why Are Rooie Rules Nice?* [121])

Keeping in mind these factors when creating games, and spending the time and effort necessary to produce an enjoyable, fun gaming experience, will allow researchers to leverage this store of knowledge and fondness associated with dynamic, interactive systems.

Although not always to great success the majority of pervasive games do attempt to follow these guidelines. Magerkurth et al. [158] segment pervasive

games into five different categories: smart toys, augmented tabletop games, affective games, location games, and augmented reality games.

There are a number of smart toys represented in current literature. Smart toys are generally reminiscent of traditional toys with additional affordances provided by the integration of technology — reacting to the toy's physical state and displaying graphical information, playing sounds, or encouraging the child to play with the toy in a certain manner. They range from small, individual toys like the Zowie playsets [68] through larger collaborative play works like Lila [89] up to full tabletop story/game systems allowing a number of children to interact at once, like The Coloured Farm [129].

Augmented Tabletop Games are only a step or two from Smart Toys. Traditional table top games, such as chess and go, continue to be highly popular in spite of the arrival of computer based entertainment technology. The direct interaction between players, and the shared social bond created from all the non-verbal communication allowed by the setting are generally lacking in standard computer based gaming. Augmented tabletop games promise to provide the best of both worlds — mobile, interactive, games with a dynamic number of players and rules combined with complex simulations, impartial judging, evolving environments and the ability to save game state. Examples can be seen in games like False Prophets [161] which used projection onto a touch sensitive playing surface combined with PDAs and interactive game pieces, and Tarboard [150] which works by taking a traditional card-based battle game and — with the help of cameras, a computer and a projector — augmenting it with animations representing the battles represented by the cards.

Augmented reality (AR) games take this further. As the name suggests, instead of augmenting a small number of devices, like smart toys, or a location, like an augmented tabletop, AR games attempt to augment a wide range of locations, objects, and interactions. They strive to track users' full range of motion and interactions with everyday objects while adding virtual elements to this combined reality. These virtual elements are affected by actions taken in the real world, and — while it is technically more difficult and confined to a pre-defined subset — real world objects are affected by actions in the virtual world.

Although the technology is in its infancy, games like ARQuake [194] and Human Pacman [53] allowed people to play traditionally virtual games using the physical world, folding all its variance and dynamism into the games-

pace.

Location Games are probably the most prevalent type of pervasive game — possibly due to the availability of cheap GPS and WiFi enabled mobile devices. Games like Treasure [27], Eyespy [30], and Heartbeat [159] as well as Hungry Yoshi, covered in Chapter 4, use physical location as an input to a game played out in a shared virtual space. Can You See Me Now? [32] and Uncle Roy All Around You [33] combine virtual and physical domains — with online players using standard home computers interacting with players in the physical world using text messages and streamed audio.

Affective Games are those that relate to, arise from or deliberately influence emotions [193]. Currently the games that are available use a physiological signal as an explicit input to a game environment, be that from an EEG or Galvanic Skin Response [158]. While affective systems are very much in their infancy, compared by Picard to very early speech recognition systems able to understand only a small pre-defined set of words in a controlled environment, they offer the promise of an incredibly rich, engaging contextually aware games in the future when combined with techniques used in some, or all, of the other pervasive gaming categories.

Building engaging mobile and contextually aware games and games which integrate into everyday life gives the opportunity to conduct engaging and transferable user trials. Another advantage is that game play can be designed to hide limitations in the current implementation of a system while exploring its potential [227], or to generate desired behaviours [226], or emotions [222]. The CHI 2010 workshop [45], and subsequent special issue [168], on games as vehicles for research in HCI exposes the breadth and depth of this practice within the community.

The study and design of games has added diversity to many areas of ubicomp research. Games have wide social and financial impact, and form an interesting application area in themselves [214], however the challenges that deciding to base a trial upon a game presents should not be ignored. The challenge of designing enjoyable experiences, implementing robust distributed ubicomp systems and having the software produced by a small number of researchers evaluated against competitive mobile games with an estimated average budget of US$100,000 [235] by end users must be taken into account at the planning phase of any project.

2.4 Descriptive Theories for HCI

Descriptive theories seek to explain the behaviour of the world. They tend to provide more conceptual models as opposed to predictive theories, that seek to provide models able to predict outcomes based on changing values of component variables, and generative theories that generate guidelines and principles providing applicable knowledge and models.

Descriptive theories allow greater understanding the social practices of people engaging with mobile applications, and how design intervenes and opens up possibilities for new practices is of central importance not only to being able to design better systems but also in finding research methodologies that better deal with the challenges of large scale user trials, in much the same way that workplace studies have informed the design of CSCW systems.

This examination of descriptive theories for mobile HCI was also driven by the previous work of the research group, which has a strong social emphasis: both in the collaborative nature of the technologies that are developed and in the broader orientation towards the work of Heidegger, Goffman, Garfinkel and Wittgenstein.

However with the movement to remote and mobile participants the practicalities of studying situations using traditional ethnographic immersion becomes incredibly difficult, if not impossible. In the following sections the history and use of ethnography is explored, along with how its methods can be extended to meet the challenges faced in the following chapters. The use of both ethnomethodology and Activity Theory, specifically Engeström's form of Cultural Historical Activity Theory, to analyse the results of this ethnography and how they fit with the restraints placed on researchers by large scale user trials are also outlined.

2.4.1 Ethnography

Ethnography has its roots in anthropology and sociology and has, in some forms, become a poplar model for research in the HCI community [70, 169, 120]. Ethnographic research methods are being applied to a variety of projects, including the evaluation of mind map software [84], a proximity based mobile game [154], the working practice of nightclub DJs [3], the

results of severing IT workers from their email [162] and the analysis of a mobile, interactive performance [236] at this year's ACM CHI alone.

Unlike many other scientific research strategies, the ethnographer as researcher is not typically a detached or uninvolved observer. The ethnographer collects data and gains insight through firsthand involvement with research subjects or informants.

From the standpoint of ethnography, the only plausible way to study social and cultural phenomena is to study them in action. [175]. The formalised ethnography in anthropology is generally seen to have grown from the foundations of the then mainstream practice of ethnology, comparative analysis of different cultures using observational data, in the late 1910's [210] along with the rise in modern fieldwork.

Where previous ethnographic researchers relied on pidgin or interpreters and augmented their data sets with third party accounts from sailors and travellers the new wave, like Malinowski and Radcliffe-Brown, guided by Boas lived among their subjects for extended periods of time. They learned the local language, recorded local myths, customs and ceremonies is much greater detail than had been done before [210]. Although it has been argued [206] that earlier 'travel writing' such as as Fanny Wright's 'View of society and manners in America' published in 1821 and Harriet Martineau's 'Society in America' study from 1837 should be considered ethnography as they side-stepped these barriers by focusing on America – living as part of their new culture using a language they already understood after emigration from europe.

The influence of Boas, Malinowski and Radcliffe-Brown led to this sort of intensive fieldwork being viewed as a 'right of passage' for subsequent generations of anthropologists [234]. It must be noted that the desire to follow in the footsteps of Malinowski, Radcliffe-Brown and their contemporaries was not the only driving force behind the move from the practice of explaining the difference between cultures and to documenting and understanding a culture from within — "to grasp the native's point of view, his relation to life, to realize *his* vision of *his* world" [160].

The values of an incumbent theory, determining if it will be embraced by a research community, can be seen to be either epistemic or non-epistemic values. Epistemic values are those related to the theory as a means towards greater understanding and the advancement of research, such as simplicity [252] or predictive power [216]. Non-epistemic values are those connected

with the scientific context in which the theory evolves, for example its alignment to current moral [167] or political values [119].

Boas challenged the basis of the evolutionary ethnology of the time on the premise that it was based upon a moral foundation, not a scientific one — the progressivist belief that human groups evolved at different rates and that cultures which were at a higher 'phase of development' achieved this as a result of the superior inventiveness of the race, or by borrowing the inventions of others. [210]

The move of such a politically influential figure – Boas was key to anthropology being recognised as a field in its own right – away from ethnology to align with his own morality [228] and the concurrent rise of political egalitarianism helped align the context in which new research was being done more closely with the non-epistemic values of ethnography than ethnology. With its solid epistemic values, ethnography was in fertile ground that lead to its swift adoption in the UK and with Boas' former students in American universities.

Ethnography developed as the study of cultures. Originally, the idea of a culture was tied to the notion of ethnicity and geographic location, the culture of the Trobriand Islands in [160] for example, but the areas in which ethnography has been applied has broadened this definition to include virtually any group or organisation. Using ethnographic methods researchers are able to study the 'culture' of a business, a sports club or the users of a particular system.

As explained by Reason and Rowan [204], the strengths of ethnography stem from the fact that good research of this sort is involved, committed, relevant and intuitive.

Ethnography must go beyond mere storytelling, it must encompass the elicitation and documentation of cultural knowledge. Seminal examples of ethnography such as Clifford Geertz's 'Deep Play: Notes on a Balinese Cockfight.' [94], the detailed investigation of patterns of interaction of football hooligans in Marsh *et. al.*'s 'The Rules of Disorder' [163] and the analysis of life within the Moonies by Barker in 'The Making of a Moonie: Choice or Brainwashing?' [25] all demonstrate the vitally important principle, as argued by Glaser and Strauss [100], that in ethnographic research theories must be developed and tested during the process of the research itself.

2.4.1.1 The Tools of Ethnography

The formalised multi-method form of ethnographical research used today reduces the risks stemming from reliance on a single kind of data and makes triangulation possible, allowing the researcher to compare data collected by different methods to aid understanding [231].

This type of research is concerned with the interaction of events, actors and system — the study of any one of these hold very little meaning without the others, and the research itself is embedded in the social, interconnected world under investigation. Researchers practicing Ethnography therefore recognise that they are part of the world they are studying, and that they will have an effect on the subjects under investigation.

Two key issues in any ethnographic study are those of access and of field relations [22]. In the type of trial covered in this document, the depth and breadth of access, and therefore the researchers' relations with their participants, is inherently different than a traditional ethnographic enquiry. Where, in other types of ethnographical observation, the researcher will negotiate access to the setting and begin the relationships with the actors then, defining what is and what is not to be part of the research, here the researcher intervenes in the user's the setting by introducing, and potentially insisting upon the use, of an application — the application within and through which the actions and actors will be observed.

Blomberg et al. [36] characterise ethnography with four principles and three main techniques: it takes place in natural settings; it is based on the principle of holism, that is, particular behaviours must be understood in their respective contexts; it develops descriptive understanding; and it is grounded in a member's point of view. The main techniques they use are observation, video analysis and interviews.

In this section each of the main techniques is discussed and the relation it has to the large scale user trials in Chapters 4 and 5 explored.

Observation

Observation is the primary means by which a researcher can examine the actions of a participant and the broader context in which the actions take place. In purely observational studies the actions can be open to possible misinterpretation by the researcher, a risk reduced in ethnographic studies by the researcher's long-term immersion in the environment.

There are two modes of observation, direct and indirect. In direct observation, the researcher is present in the subject's environment and watches the subject go about their everyday routine or perform a particular task. With no mode of recording the events that are being observed, one limitation is that events of interest may be missed by the researcher [224] and that there is no way to revisit the data [198].

The most commonly stated limitation of observational studies is 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 behaviour not because of any specific condition being tested, but simply because of all the attention they receive [208].

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 no human-centred study is completely free from the Hawthorne effect [156]. However, the generalisability of the Hawthorne effect has recently been called into question [156, 208]. Macefield [156] presents a full discussion on the limitations of such a generalisation with respect to usability evaluations. Similarly, Crabtree and Rodden 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" [59].

Observational studies were not conducted as part of the trials in Chapters 3, 4 and 5 because the global spread of participants increased the monetary and time costs far beyond the perceived benefit. Interviews, both remotely as practiced in Chapter 4 and locally as in Chapter 5, gave insight into the experiences of the participants and complimented the recorded data.

Video and other Recorded Data

There are a number of advantages to video recording in ethnographic research. One advantage is the density of data that a visual recording provides [104]. In an ethnographic approach to research, the goal is to study real people in real situations, doing real activities. Video can provide more contextual data than audio data alone [93, 125], and the addition of further sources of data as are available on modern smartphones even more [205],

such as the usage logs, audio recordings and mini-surveys used in chapters 4 and 5. Indirect observation also helps alleviate the problem of the Hawthorne Effect.

This can give a more complete sense of the participants, the setting in which they function and the types of activities they engage in.

Detailed analysis of the skills and methodologies needed for successful analysis of recorded data, covered in [58, 80, 81, 82, 102, 101] are beyond the scope of this document. Here only the most relevant points are highlighted.

The greatest advantage of recording is permanence [104]. This allows an event to be experienced repeatedly, and with each repeated viewing, the observer can change focus to things not noticed at the time of recording or on previous viewings [81, 80, 86]. Replaying the event also allows more time to contemplate before drawing conclusions, and hence serves to ward off premature interpretation of the data. Even a rare event, when captured, can be replayed repeatedly for a thorough analysis and intensive study. Real time observation does not have this advantage [81].

A recording contains very little information on how typical an event is [62]. Whether the event is frequent, unusual or unique must be supplemented by the ethnographer, by drawing on the time spent in the field as a participant-observer, or triangulating with other methods of data collection such as usage logs and survey responses [58, 81, 107].

The unspoken thoughts and feelings of a participant cannot be probed while watching a recorded event. Tacit knowledge and influential experience cannot be accounted for when relying only on observation [223]. However it can be played back to the participants [125, 124, 58, 87] in order to attempt to get them to recall and describe their thoughts, feelings and reactions at different points in time during a given event, thus giving information about the unobservable. A version of this technique was performed with the data, specifically location of use, recorded from participants in chapter 5 presented back to them to explore the nature of their understanding of the trial process.

Recordings only allow the events to be experienced vicariously, as opposed to participant observation where the ethnographer can test out emerging theories in the field [81].

No recording can show every observable thing that happened, but only that which was occurring within the range of the camera lens [75], or the

equivalent limit on the accelerometer, GPS or magnetometer sensors being recorded. The camera can no more provide accurate observations in the dark than a GPS can provide accurate locations inside a building. More of the context can be understood by recording a larger area and for a longer time than the specific event under investigation strictly requires [75], however a balance must be sought between the participants needs and expectations of privacy and the researchers desire for greater fidelity of data which is discussed in Chapter 6.

The ever reducing cost and increasing fidelity of recording instruments combined with the increasing complexity of tools available to synchronise and examine their output [174, 180, 135, 181] point to recorded data becoming more and more influential in the analysis of ethnographic research.

Interviews

There are many limitations to interviewing as an investigative technique, the most obvious being the widely acknowledged discrepancy between what people do and what people say they do. Interviewees may also tailor their answers to suit what they think the interviewer wants to hear [198] or to maintain their presentation of self to the interviewer [128].

A less obvious limitation is that interviewing relies on a degree of reflective expertise on the part of the subject, and the ability to articulate their thoughts, feelings, and experiences even though one purpose of interviews is to gain insight into the thoughts, feelings, and experiences of subjects that may not otherwise be easily observed [198].

Additionally, the questions that are asked are limited by the assumptions of the researcher. While this may be useful in situations where the research question has a narrow focus, in more exploratory studies this may delimit the subsequent scope for potential and valuable findings. The implication of incorrect assumptions is most damaging in structured interviews, in which the researcher follows a script predetermined questions with no opportunity for deviation, clarification or explanation. Semi-structured interviews offer some purchase on this problem in that the researcher enters the interview with a loosely defined schedule and willingness to let the course of the interview be guided by issues that are raised as relevant by the subject.

The choice was made to use such semi-structured interviews in the work described in this thesis. The determining factor behind the choice to use semi-

structured interviewing was the impracticality of employing observational methods to investigate geographically and temporally disparate phenomena. The problems involved in arranging and performing semi-structured interviews, and their efficacy in large scale user trials, is discussed in Chapter 4.

2.4.1.2 Ethnography in HCI

The emergence of ethnographic enquiry as a method of choice within HCI can be attributed to the fields of Computer Supported Cooperative Work (CSCW) and Participatory Design (PD), which imported ethnographic methods from anthropology and sociology to study the use of technology in situ. The use of ethnographic enquiry within HCI has been argued to be a method of rich requirements capture [70, 223, 224]. Apart from the insight gained into social practice, its merits with respect to the design process include providing "a useful contrast to traditional methods of requirements capture" [223] and the engagement of users in the design process [224].

The difficulty ethnography faced when attempting to influence design has been highlighted [70, 223], and the suitability of ethnography for the task of generating implications for design has been questioned [70, 223, 224]. It is not suggested that this in unobtainable, but simply that implications for design don't necessarily follow from ethnographic findings [224], or that ethnographic findings are more suited to identifying how people cope with existing technologies rather than inventing new ones [71, 196, 223]. The most influential ethnographic [219] studies in CSCW did not provide design recommendations "but instead tried to uncover, in minute detail, the ways in which social order is produced in cooperative work settings." [219]

More recently the notion of evaluating the merit of ethnographic work carried out within HCI by the presence or absence of implications for design has been challenged [70, 71]. Dourish [70] suggests that HCI needs to distinguish between ethnography to inform system design and ethnography to study human computer interaction.

While ethnography purists argue that employing methods such as quasi-experimental statistical tests or dependent and independent variables results in 'dead knowledge' [204] and that 'it is much better to be deeply interesting than accurately boring' [204] when human action and interaction are the subject of the research these tools are often employed in HCI

research to support the claims of ethnographic enquiry and add confidence to claims of generalisability.

An ethnographic description may contain a large amount of information with direct value to design and evaluation but it is still a largely unconstrained and personal narrative account. This raises problems of abstraction, generalisation and comparison and leads to a lack of cumulative research results [179].

Ethnomethodology [73] (EM) has also been used in the analysis of ethnographic data. EM is a particular analytic orientation to the practical study of social order, which can be seen woven through the fabric of everyday activity. Ethnomethodologists view human social action as something that is reflexively accountable.

Accountability not in any moral or political sense, but in the way the action is observable and reportable – in other words, the very way in which the action is orchestrated by the actor or actors provides others the means to recognise it as what it is. It is seen to be reflectively so in that it is accountable through its own production rather than within a frame of a set of predefined, global social rules.

Although there is some resistance to the evolution of EM-inspired ethnography [60] there have been attempts to move forward with hybrid approaches such as Technomethodology [73]. This methodology has the goal "to understand how ethnomethodological understanding of human social action and interaction can be used, directly, in designing interactive technologies." looking at an implementation practice called "Open Implementation" that aims to change the standard abstractions used in all programming languages for ones where the relationship between "what is done (the implementation behaviour), and what is done by what is done (the achievement of application ends)" is articulated in such away that it is accountable. By providing a development and design space that ties in with ethnomethodological understanding of action, then that understanding can be directly applied to design without being translated, as it were, by the ethnomethodologist.

MUST[4] [140] is another child of EM, although firmly in the camp of Participatory Design, providing a framework for the design process and combining ethnography with intervention, intervention meaning to deliberately set up activities that change the norm in order to learn from the actors' reactions to

[4]A Danish acronym for Theories of and Methods for Initial Analysis and Design Activities.

that change — as opposed to ethnography, where ethnographers strive not to change what they are studying. MUST provides guidelines for iterative ethnographies and interventions, allowing for interventions that happen "in the mind of the designer or though conversations among designers" [140] to some extent but stressing the importance of 'real' interventions in cases where the imagination is not enough.

Resonating with the earlier arguments relating to the merits of ethnography, Shapiro [223] calls for a more balanced view of what the various disciplines offer. This view has since been voiced by Cockton [56] who suggested that combining computing, sociology, and psychology is essential for the further development of the field. According to Cockton, computer science is a "source of innovation", sociology "provides the understanding required to shape innovations to usage contexts", and psychology "supports design refinement and evaluation of interaction".

Dourish [70] also highlights the disparity between the overall goals of ethnographic enquiry and what he refers to as "discount ethnography" techniques (for example, contextual enquiry and cultural probes), the primary outcomes of which are implications for design. Similarly, rapid ethnography [169] seems at odds with the ethnographic orientation towards understanding from the perspective of the population being studied. One of the three key ideas of rapid ethnography offered by Millen [169] is to narrow the focus of research "before entering the field" and to "zoom in on the important activities". The questions must be asked as to who determines what the important activities are and how they are determined.

By looking at such ethnographic studies though the lens of Activity Theory, discussed in the next section, a common vocabulary and structure of reporting along with methods to include participatory design techniques and iteratively re-focusing on the important activities in a manner rapid ethnography but informed by participants may be possible.

2.4.2 Activity Theory

Activity theory offers a set of perspectives and a set of concepts for describing human activity. It has been argued [148, 79, 130, 176] that HCI research would benefit greatly from the incorporation of this theory as it struggles to describe and understand concepts such as "context," "situation," and "practice." It has been recognised for some time that the use

of technology is not simply an input-output relation between a person and a machine; a much richer depiction of the user's situation is needed for design and evaluation. However, without moving beyond the ethnographic approach above it is unclear how to formulate that depiction in a way that is not purely ad hoc.

Activity theory can bridge this gap by providing orienting concepts and perspectives. As Engeström explained [79], activity theory does not offer "ready-made techniques and procedures" for research, its conceptual tools must be "concretized according to the specific nature of the object under scrutiny."

As in cognitive psychology, activity theory can be seen as a "real, natural sciencelike," theory [131]. Yet, unlike traditional cognitive psychology, activity theory analyses human beings within their natural environment in the manner of Gibson's theory of affordances [19]. Activity theory also takes into account the cultural and developmental aspects of human consciousness [37, 151, 243].

The fundamental principle of activity theory is that consciousness and activity are inseparable. "Consciousness" being the human mind as a whole, and "activity" the interaction with the objective reality. This asserts that the human mind emerges from, and exists only, as a consequential component of interaction with the environment. Mind is seen as a special "organ" resultant from evolution in order to help the species survive. In this way it can only be analysed and understood within the context of activity.

The second principle is that of object-orientedness. This differentiates activity theory from the theories of Gibson by including the social and cultural properties of the environment and treating them as objectively as the physical, chemical, or biological ones. In this way the statement that "the object is a key" is no less an objective property of something than "the object is made of a metal alloy." This is in contrast to the cognitive approach that relies on the assumption that the human mind's only contacts with reality are through low-level input-output processes.

The activity theory approach defines the environment with which human beings are interacting as one that is in itself full of meaning. This environment consists of combinations of objective features, including the culturally determined features that define the affordances for people interact with each object.

The third basic principle of activity theory is that activity is structured in

a hierarchy. Activity theory describes groups of levels into which processes are differentiated by taking into consideration the objects to which these processes are oriented. Activities are oriented to motives, where each motive is a material or idealogical object that satisfies a need. Actions are the processes performed in the pursuit of a specific conscious goals and are functionally subordinated to activities. This differentiation between objects that motivate human activity and the goals to which this activity is immediately directed is fundamental to activity theory. Actions are performed through operations determined by the conditions in which activity takes place.

In using activity theory as a foundation for understanding behaviour, the importance of differentiating among motives, goals, and conditions becomes apparent. When operations are obstructed (by a change in familiar conditions, for example), people adapt readily to the new situation. When a goal is obstructed, it is necessary to reassess the situation and set a new goal. This is often done without much effort or frustration. It is often possible to predict what the new goal will be given an understanding of the current motive. But when a motive is frustrated, people are upset, and their behaviour is more unpredictable [130].

The fourth principle of internalization-externalisation [243] describes how mental processes develop from external actions through the course of internalization, which by its nature is a social process. Internalization provides a means for people to try potential interactions with the physical world without performing actual manipulation with real objects but instead using mental simulations or considering alternative plans. Externalisation transforms internal activities into external ones and is often necessary when an internalized action needs to be examined and its validity reassessed, e.g. using pen and paper to check mental arithmetic. It is also important for collaborative actions as coordinated between several people requires their activities to be performed externally.

The fifth principle is mediation. Human activity is mediated by tools, both external (like a hammer or scissors) and internal (like concepts or heuristics). The modes of operation of these tools are developed over the history of the society. These culture-specific tools, by their embodiment of societal norms in their affordances, shape the way people act and, through the process of internalization, influence how people think. Tools in this way are an embodiment of cultural knowledge and social experience. The mechanism underlying tool mediation is the formation of “functional organs,” that

combine human abilities with the capacities of external tools to perform a new function or improve on an old one. For example, human eyes combined with glasses form a "functional organ" that provides better vision.

The last principle is that of development. In Activity Theory the development of a system is not only an object of study, it is also a research methodology. The basic research method combines active participation with monitoring of the developmental changes of the participants. Ethnographic methods that track the history and development of a practice are used to understand a phenomenon by understanding how it developed into its existing form.

These principles are components in an integrated system representing human activity as a whole. The application of any of the principles will often makes it necessary to apply the others as well [132].

Analysing a disruptive technology using an activity theory perspective would mean first identifying the various activities and the related objects (object-orientedness), the hierarchical structure surrounding these activities (hierarchical structure), how the activities are being changing through the use the technology (mediation), and how transformations of external components of activity are related to changes of internal components (externalisation-internalization). This should be analysed over time (development).

2.4.2.1 Engeström's activity system model

The foundations of Engeström's activity theory can be traced back through the Vygotskyian concept of tool mediation, Leont'ev's notion of activity and Marx & Engels' concept of labor.

Vygotsky determined that human interaction with the environment must be mediated through objects; be they physical such as tools or virtual such as language and culture [243].

This mediation should not be seen as a flow of intention from the subject, through the artefact to effect the object. The artefacts available restrict and inform what can be done with the object, providing motivation for the creation of new tools, and their use influences the subject and their future decision making in turn, by learning to manipulate the artefact with greater skill for example. This mediational model of human interactions with the environment is usually portrayed as in Figure 2.1. This model highlights that the relationship between the Subject and the Object is mediated through the use of Tools.

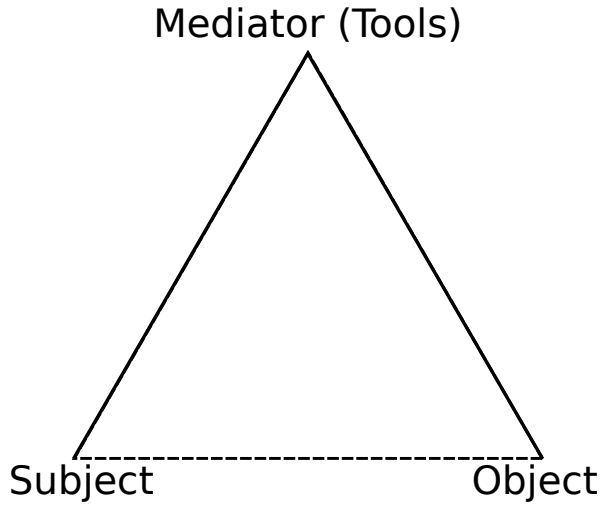

Figure 2.1: Vygotsky's Mediational Model of Activity

Leont'ev's theory of activity, another important step towards Engeström's approach, is predominantly concerned with activities of individual human beings. While Leont'ev, building upon the mediational model above, explicitly mentions that activities can be carried out not only by individual human beings but also by social groups, there is no conceptual model of collective activity — Although this omission could have been a politically motivated decision [132].

Engeström proposed his activity system model of collective activity as an extension of Leont'ev's concept of activity. The most significant revision was adding a third node, "community", which resulted in a structure comprising a three-way interaction between "subject", "object", and "community". Each of the three particular interactions within this new structure was mediated in different ways. For the relationship between subject and object, tools and instruments mediate, in the same manner as in Leont'ev's model. The relationship between subject and community is mediated through rules, and between the community and object the mediation factor is the division of labour. This new model also included the outcome of the activity system as a transformation of the object produced into an intended result, which in turn can feed into another activity system. This model is shown in Figure 2.2. The following chapters focus upon the evaluation and design of mobile systems, so as an example consider the user of a mobile, social game centred around the dynamic of passing digital artefacts down different paths on the user's social graph in order to score points. These artefacts come in different 'flavours' describing how the type of connection the user has with the person they pass the artefact to impacts the number of points they

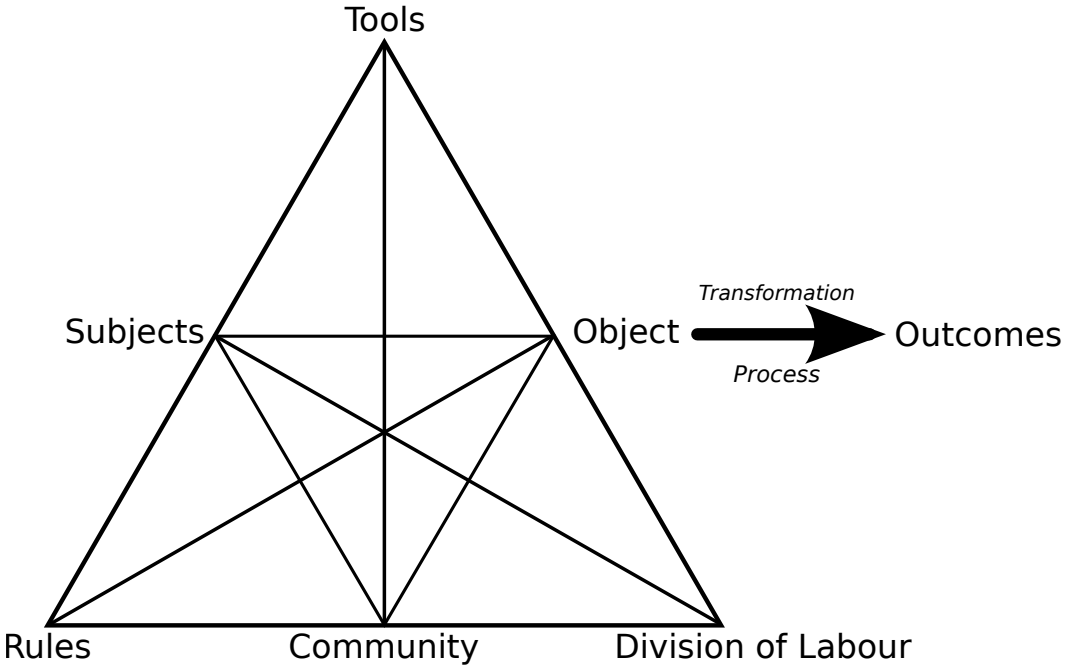

Figure 2.2: Engeström's Activity Triangle Model

receive, individually and for their team.

In this example the activity under examination is the user receiving an artefact and passing it to a third party. The object of the activity is this digital artefact, and the expected outcome the transfer of this artefact to a third party.

The user employs a variety of tools on the object, such as the game software itself to inspect its affordances and techniques she has learned or devised for successful play [121].

The community comprises of her friends on the social networking site that the game is based, and the members of her in-game team. Her relationship with the community is mediated by explicit and implicit rules, e.g. social commitments, presentations of self, perceived engagement with the game, etc. Beyond that the artefact transfer, and therefor the individual and team score achieved, is a product of the effort of the user in identifying the potential targets to transfer to and their subsequent negotiation of the transfer, the coordination of which is achieved by employing a devision of labour which mediates the relation between the community and the object.

When studying complex real-life phenomena, applying one activity system model is often not sufficient. Such phenomena need to be represented as networks of activity systems. In this example the two different communities, that of the user's friends on a social network and the team, which may or

may not be a subset of this larger group, lends itself to a network where the devision of labour between the team may influence the object's suitability for one transfer path over the other. If the team quota for transfers to those in their twenties had not been filled that day, the user may choose to spend more time negotiating with a college to convince them to play the game and accept the transfer of an object that must be passed to a female player rather than pass it to her mother who is already a keen player and would accept quickly for a lower points score.

It is key to this framework that the activity systems are constantly developing. This development is explained as a process driven by tensions. Engeström identifies four types of tension in activity systems:

Primary tensions are inner contradictions of each of the nodes of an activity system. For example, a medical practitioner's mediating means include medications. The medical effects of these must be weighed against availability, costs, legal regulations, etc.

Secondary tensions arise between the nodes of an activity system. For example, certain medication may be unsuitable patients with high blood pressure.

Tertiary tensions describe potential problems in the relationship between the existing forms of an activity system and the more advanced object and outcome. The improvement of an activity system can be jeopardised by the resistance to change in the existing organisation.

Quaternary tensions are those within a network of activity systems between one activity system and other producing of a joint outcome.

Activity theory does not support the creation of predictive models that only need to be given appropriate data to determine in advance the actions people will take. It is a tool to help researchers orientate themselves in complex real-life problems, identify key issues, and inform the search for relevant evidence and suitable solutions. Activity theory helps researchers ask right questions rather than providing answers.

2.4.3 Activity Theory and Ethnography for Large Scale Trials

Activity Theory, with its multitude of interpretations and methodologies, poses more problems than other theories when applying it to practical situations [63].

Additionally there are the challenges presented by applying a methodology to a nascent style of field trial, one without a large body of work to call upon to ground and inform methodological practice, which apply to both ethnography and activity theory based methodologies.

Large scale field trials' most apparent problems stem from the barriers it places between the researchers and their participants. As identified in §2.2 in almost all cases of published large scale field trials the communication between researcher and participant has been through store reviews, with very few using other technologies such as forums, email lists or social networks [43, 213].

For most practically applied versions of Activity Theory, the Action Workshop [148, 79, 176, 57], a directed discussion within a self-selected group of participants around the problems with the current system, is an integral part. At a distance this would have to be done with asynchronous communication channels. Asynchronous remote interaction can lead to useful decisions and products, as recently shown by Fisher et al. [88], but it is also clear that such interaction involves forms of communication that are, relatively speaking, detrimental to group decision making [31] and collaboration [183] — both of which are integral to the running of a successful Action Workshop.

Ethnographic approaches are, by their very nature, heavily reliant on observational techniques and immersing the researcher in the culture and environment of the participants. Without face-to-face interactions before, during and after the trial the norms of social research must be bent or abandoned altogether. Traditionally, pre-experiment briefings not only discharge the researchers ethical obligations, discussed in detail in Chapter 6, they also negotiate the expectations of engagement and imposition between researcher and participant [103].

The results of this lack of negotiated obligation on the part of the participants, with respect to the relative costs of contacting participants and the collection of log data, is discussed in Chapter 4

However, despite these challenges, there would seem to be significant potential rewards if a useful synthesis of theories could be developed that would give deep understanding of use and context, as described in this section, and large scale user trials that give wide demographic as well as geographic coverage. The goal of developing such a synthesis drove a significant amount of the work covered in the following chapters.

2.5 Conclusion

This chapter has shown that the trend for user trials in HCI has seen them spreading into more and more real-world situations of use to incorporate more aspects of context and mobility, and for the length of such trials to increase. While there is no doubt that trials conducted in controlled and semi-controlled environments will continue to provide the most reliable answers to some types of questions asked by researchers, such as those involving detailed analysis of interaction on a mechanical level or the effects of certain types of distraction on performance, the expansion of the accepted techniques for the evaluation of the effects of mobile technology on the everyday lives of end users, so as to include large scale user trials, seems inevitable.

The chapter continued by discussing the suitability of using games as a vehicle for research, pointing out the increased engagement and the inbuilt reward systems available to influence participant behaviour in order to further research goals but also the dangers of higher expectations of 'fun' and quality than developers of research applications may be used to providing.

The use of good game design, as described above, to motivate users to not only download and use applications, but to incentivise certain behaviours such as the participation in trial aspects of the application to a greater degree than they would otherwise, is a goal which will be better achieved in Chapters 4 and 5 in ways based on the understanding gained from work described in §2.3.

This chapter's focus was then turned towards the methodological challenges of conducting research involving large scale trials of games that serve as vehicles for research issues not specific to game design. Starting with the history of ethnography and its move into HCI, the practicality of using an ethnographic approach was discussed. Activity theory was explored as an alternative approach to gaining a deeper understanding of systems and their use with the goal of redesign.

There is clearly a tension or opposition between, on the one hand, methods based on ethnography and activity theory, so appealing in the depth of understanding and richness of detail they obtain and, on the other hand, methods that are practical when working with large scale deployments so as to acquire large numbers of remote participants in varied geographical and cultural settings. As discussed in later chapters, especially Chapter 5, this

tension led to new approaches and techniques that offer a useful synthesis of these apparently opposed methods.

The following chapter begins the main body of this thesis by describing the nature of the environment into which the applications in Chapters 4 and 5 were released and gives a detailed description of the software release patterns and deployment choices available on the iOS platform.

Chapter 3

Distribution Methods

This chapter begins the main body of work in this thesis. As discussed in §1.1 the author's research group decided to use iPhones and the iOS platform. Early in the design of any research application destined for release a decision must be taken as to the route by which the software will be distributed to end users, and as shown in this chapter there are a number of aspects of the distribution method which must be taken into account when deciding on which path to take, and resulting consequences. Here the decision was made to use iOS devices, however within this platform two distribution mechanisms are available for use; the official App Store and 3rd party APT repositories. The decision of which mechanism to use for distribution is in many ways similar to the decision of which platform to develop for as each determines the software APIs available, the content restrictions enforced, the demographics of the user base and the visibility of any application in the store.

In this chapter, these choices and consequences are discussed within the context of the two distribution mechanisms available for iOS devices. They were explored using deployments of two purpose-built applications. Guidelines for researchers are proposed, to help them choose the distribution method most suitable to their research.

The majority of the work published in the last couple of years on mass deployments of research applications, as outlined in §2.2, have used Android devices. However, researchers working in this area in 2008 had a difficult choice — while promising a more open development environment, the first Android based commercial mobile phone was only released in the second quarter of 2008 by HTC [7] and by the end of 2009 had captured less than 5% of the smartphone market [247].

The original iPhone was released in June 2007. Within weeks of its launch, the development community had produced a method of distributing software directly from the developers to the end users' handsets. In July 2008, Apple launched their App Store and addressed many of the traditional difficulties users experienced in downloading third party applications to their smartphones — all the applications were available in one place, compatibility was easy to ascertain and the process was made as painless as possible for the end user to the point where it was possible to install new applications without the need for a desktop PC. Faced with these opportunities, a number of research groups made the time and monetary investments necessary to move their development to this platform.

At the time of writing, the iOS platform, which runs on iPod Touch and iPad devices as well as iPhones, has a larger installed user base of over 315 million units [170] than either Android OS, with 300 million, Blackberry OS, with 77 million [139] and Windows Phone 7 with an estimated (as Microsoft does not release activation figures for the operating system) installed base of 10 – 12 million [17, 28].

Oliver [186, 187] noted in relation to the iPhone that "Out of the box, iPhone is a substandard research platform; however, unlocking it exposes a rich set of APIs from its Mac OS X foundation." Due to the restrictions on APIs and direct access to the hardware components when complying with Apple's requirements for an application's entry into its store.

This, however, only took into account the development of the applications and not the difficulties in distributing applications to end-users. Applications taking advantage of restricted features are not eligible for distribution via Apple's App Store and therefore must be released via third party software repositories only available to those users who have 'unlocked' their devices. This exclusivity based upon APIs used, among other considerations, means that the decision as to which distribution method to use has to be taken early in the design and development process.

3.1 Distribution methods

The large scale deployments mentioned in §2.2 are all influenced by the manner in which they distribute the software to users, which is itself determined by the hardware platform the researchers have chosen to develop on and, in the case of iOS, whether they are developing for inclusion in the App

Store or not. There are two primary methods for the large-scale distribution of iOS applications; the Apple App Store and the community of third party APT-based repositories. Experiences with both will be discussed after an examination of the popularity and the submission practices of each.

3.1.1 Apple's App Store

The Apple App Store is arguably the best known and most popular mobile software repository in terms of applications available for download and number of applications downloaded, with more than 550,000 applications available for download and a download total topping 25 billion [170].

Each application must go through an opaque review process by Apple in order to be approved for distribution via the store. To pass this review it must be seen not only to comply with the 37 page iPhone Developer Program License Agreement and the 136 page iPhone Human Interface Guidelines documents [8] but must also fit within the positioning of the store in the wider market context.

The review process itself runs on a sequential failure method, meaning that although an application may break two or more guidelines it will be rejected for one, edited, resubmitted and then rejected for the next. The time between submission and review is not guaranteed, although an estimate of the current load is given on submission. This currently averages under 8 business days for 97% of applications [220], although Apple has recently made available a service whereby the author of an application can request an expedited review [12] in certain circumstances lowering the average time to 4 business days — however there is no guarantee that such a request will be granted. Apple implies that repeated requests from a single developer will be given less priority.

3.1.2 APT Repositories

Only 12 days after the initial release of the iPhone, a consumer level method to allow third party software and unrestricted access to the file system was made available online [209]. Called 'unlocking' in [187], this process is generally referred to as a "Jailbreak", with the process of 'unlocking' being popularly associated with removing mobile carrier restrictions.

The security model in iOS is an implementation of the FreeBSD jail mechanism, which is a form of OS level virtualisation to compartmentalise the system, both its files and its resources, in such a way that system users can only access their own compartments, or jails. In order to access files and services outside the jail in which user level programs are run, bugs in privileged applications or the OS itself are exploited to escalate the privileges of the user.

Here the term 'to jailbreak' will be used as it is by the iOS community not as a description of the exploit used to gain control, but of the process of taking advantage of one of these exploits to modify the operating system to accept applications from sources other than Apple and installing a repository manager on the device for this purpose. Devices that have had their operating system modified in this manner will be referred to as 'jailbroken'.

Initially homebrew software, a generic term for software developed by a user community for closed platforms, was manually loaded onto the devices. A port of APT, Advanced Packaging Tool, was quickly developed for the iPhone allowing users to manage applications in the same way as on many other *nix based systems. A native GUI, Cydia, was released shortly after, providing much of the functionality of the App Store client, to be released by Apple a number of months later, with combined access to any number of repositories the user cared to subscribe to.

A ruling by the copyright office in the U.S.A. [34] has established the legality of jailbreaking devices in order to run legally obtained software users would otherwise be unable to use, removing any danger of legal action being taken against an end user. The warranty, however, is invalid while a device is in a jailbroken state.

The jailbreak process is easily reversible: a device can be restored to its default state with the click of a single button in iTunes – the media management application released by Apple to complement their iOS ecosystem.

However, as jailbreaking methods become more user-friendly and less the domain of highly technical users, the number of users with jailbroken devices who do not understand the consequences or the procedure to reverse it can be expected to rise. The existence of such users raises questions of the responsibility of researchers releasing software in this manner. Does the act of providing desirable software only to jailbroken devices constitute an encouragement to jailbreak? If that is the case, should researchers inform users of the consequences of, and the procedure to reverse, jailbreaking

within an app they will only be able to launch on a jailbroken device?

3.2 Repository Contents and Potency

In order to collate the contents of the 38 most popular repositories, the release list of each was downloaded onto the mobile device using the Cydia application, copied to the desktop and parsed into a database in December 2010.

The download statistics page for each of the packages in the largest 3 repositories were scraped and parsed. The smaller repositories did not provide public access to download counts, but this still resulted in download statistics for upwards of 80% of packages seen. This, plus information collected from the repository websites, was used in the calculation of statistics for comparison with the Apple App Store.

3.2.1 Applications by Genre

The graphs in Figure 3.1 and Figure 3.2 show the distribution of applications and packages available for download by category. Where possible the APT categories have been coded by the author to match those in the App Store in order to allow direct comparison. Due to the nature of the two different distribution methods, the APT repositories' two largest categories have no comparison on the Apple App Store. Ringtones are sold through a different outlet, the iTunes music store, and themes are not available without 3rd party software modifying restricted files.

As can be seen from Figure 3.1, the vast majority of available downloads center around Themes and Ringtones, neither of which are allowed in the Apple App Store and are only available via this channel. The majority of applications in the rest of the categories can also be seen to fall foul of the rules Apple have set for App Store submission. Modification of the iPhone operating system (e.g. adding folders to pre-iOS4 devices, enabling wifi-only applications to run over the cellular data network), breaking the sandboxing of applications in order to add features and interoperability with other applications (e.g. adding copy & paste support to pre-iOS3 devices), running in the background, duplicating Apple functionality (e.g. 3rd party SMS clients) or using APIs which Apple have deemed private (e.g. directly accessing the WiFi) all cause rejection from the App Store but are the bedrock of

this small development ecosystem. The number of available downloads that are neither themes or ringtones is only 2541 — less than 15% of the total available. In comparison, the App Store is dominated by books, games

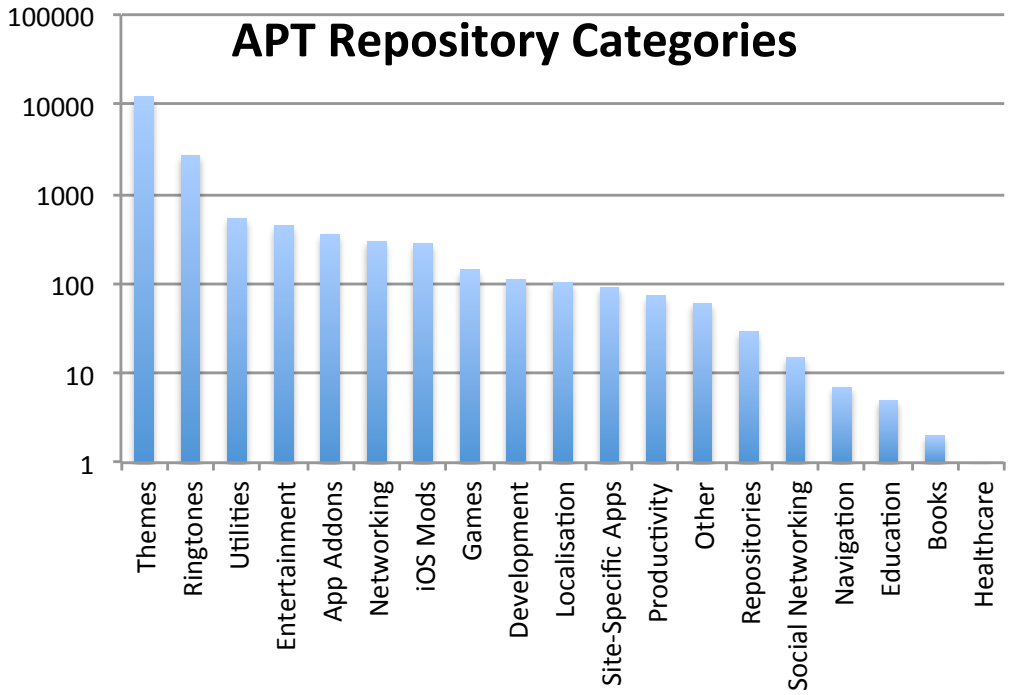

Figure 3.1: Distribution of packages across APT repositories on a logarithmic scale.

and entertainment applications — with these three of the twenty categories accounting for 46% of the total applications available for download [220]. The difference in scale between the two distribution methods can be seen between Figure 3.1 and Figure 3.2, with even the smallest of categories in the App Store providing a choice of more applications than are available across all the application categories in all the APT repositories combined.

3.2.2 Number of Downloads

The number of downloads per application or per genre is not publicly released by Apple. Most companies keep information about the number of application purchases they have gained private; only the total number of downloads for the App Store as a whole is directly available. Companies like Flurry[1] collect aggregate statistics by offering a logging framework to developers for free — giving them details on usage of their own applications

[1] www.flurry.com

Figure 3.2: Distribution of applications in the Apple App Store.

in return for the aggregate data that they can leverage in the marketplace. Admob[2] provide a large proportion of the ads seen in iPhone applications, by some estimations 61% [76], and also release some aggregate data.

Unfortunately most of the data publicly available from these sites is updated infrequently and focuses on paid-for applications, which make up 77% of the applications available for download, and has over the last year been pulled more and more behind their respective paywalls, but can still be useful in providing insight to the ecosystem.

The data for the repository packages is freely available on the three largest community repositories, although only on a package-by-package basis. This information was collated to give information in the same format to give data on 80% of packages across the APT repositories.

The same general shape of trend can be seen for each distribution method with each decile having much less impact than the one before. However, a high number of users downloading an application is no guarantee of a high number of users engaging in the application to an extent to which they can be seen to be a valid or valued trial participant.

In order to reach more than 10,000 users an application need only be in the top 50% of applications on the APT repositories as opposed to the top 20%

[2]www.admob.com

in the much larger set of applications in the Apple App Store. The exposure necessary to achieve this for each distribution method should be taken into account when deciding upon one.

3.2.3 Exposure

A common complaint made about releasing applications on the Apple App Store is being 'lost in the noise' generated by 550,000+ other applications all vying for attention within the 20 categories. A new app, depending on its release time, can have as little as 2 hours and 20 minutes on the first page of the 'New Games' section. As of the end of 2009, updates to an application no longer bump it back to the top of these lists, so paying for featured status or marketing out-with the store itself increasingly becomes necessary to achieve a reasonable amount of exposure in a short period of time.

The algorithm for computing an application's position on 'most popular' lists in each category is not made public. However, recently, the total number of downloads and the number of recent downloads seem to be major components in this calculation. Anecdotally, from commercial iOS application developers, each page of applications the user must click through to reach yours on the 'Most Popular' list for your category results in 10x fewer downloads. This can be seen from the distribution in Figure 3.3, with the average price of an application on the App Store being $1.87 [220] down from a reported $2.89 in 2010 [166], the vast majority of applications make very little money.

In contrast, the toplists for non-theme related APT packages are easier to appear on, as an application is shown on update as well as on launch, and stays there longer due to the lower number of releases as seen in Figure 3.4. This increases exposure within the community for new and regularly updated apps.

Another way to achieve exposure in both distribution mechanisms is to collect a number of positively rated reviews from users — although the exact formula used to calculate the overall grading and list position is opaque for both. Users have been found to be reticent in producing reviews in comparison to in-application feedback mechanisms.

In one of the applications released which included a direct feedback mechanism, the users made only 2 reviews in the store compared to comments

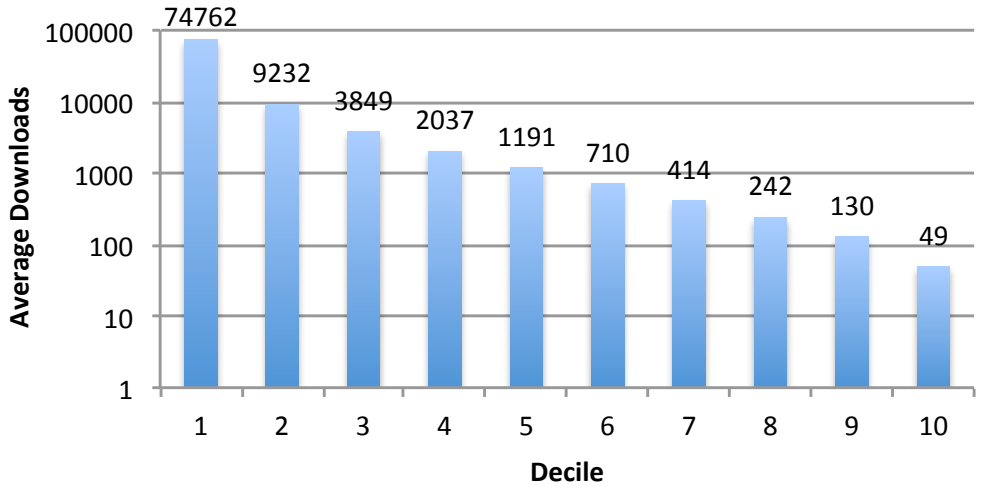

Figure 3.3: App Store Paid applications average downloads per decile on a logarithmic scale.

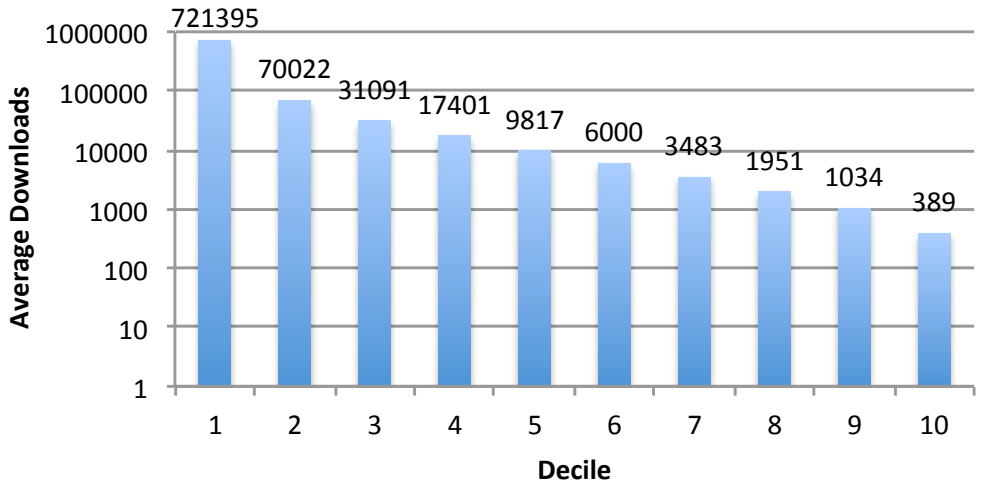

Figure 3.4: APT Packages average downloads per decile on a logarithmic scale.

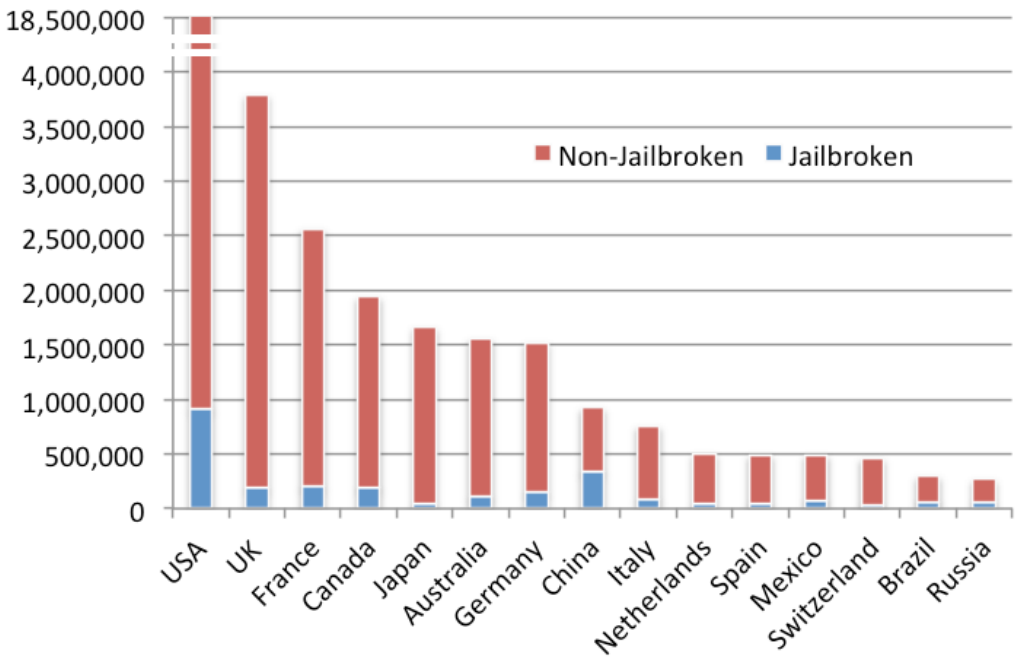

Figure 3.5: Total iOS devices per country including the number of jailbroken devices on a split Y-axis.

from 1,224 users in the application — many of which read as one might expect a review on a store to, for example "Everything is good about this app. Very useful."

The ease of exposure must be weighed against the number of users to whom the application is visible. This gives researchers interested in running a trial using one of these methods the option of lower exposure to around 315 million [106] devices or higher exposure to 39 million.

The breakdown of the location of jailbroken and non-jailbroken devices can be seen in Figure 3.5. These numbers come from combining the number of devices seen by Admob and the percentages of jailbroken devices reported by Pinch Media.

Figure 3.6 shows the average downloads per category of APT package. Unfortunately, no data is publicly available to examine the App Store in the same way.

The high number of average downloads in the relatively small category of Development points to the technical literacy of the user base. The large number of utility downloads in comparison to other categories can be explained by the need for utilities to enable the use of themes or application add-ons handled by APT dependency protocols.

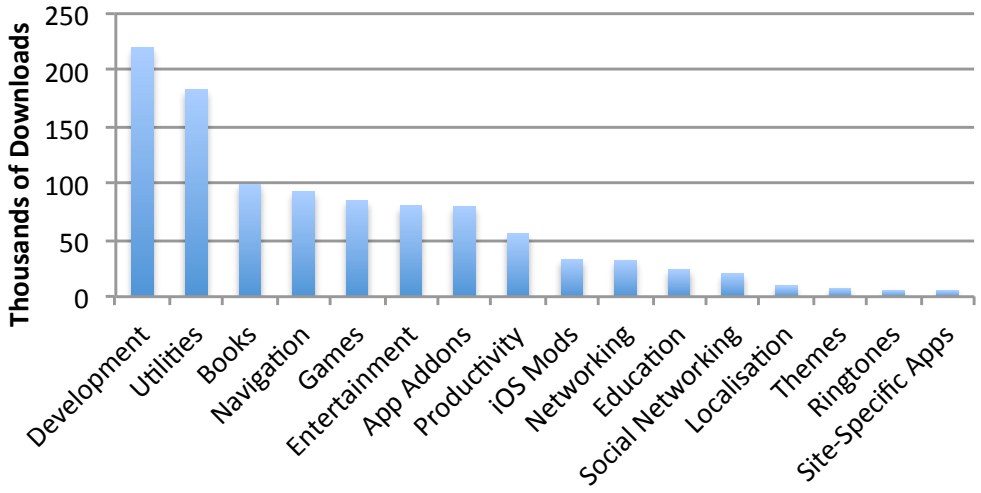

Figure 3.6: Average total downloads by repository category in thousands.

3.3 Trial release

In order to explore and document the differences in procedure, and results obtained from releasing an application in each method, a memory game application was created. Rather than the application's design being a demonstration of research concepts in itself, the application design was deliberately very simple, e.g. with no complex use of English so that it could be used worldwide, and a straightforward game design, so that it might easily be taken up and tried out by users.

The aim was to obtain significant numbers of users via each distribution method, and so inform methodology choices for later trials of applications that were more complex research prototypes.

Due to the terms and conditions of both distribution methods stating that any application must be exclusive, the application was 'skinned' to provide two very similar applications. This exclusivity was the source of the question, as any release using one distribution method raises the question of how the results would have been affected by using the other.

When users launch the application, they are confronted with a main screen giving them the option of playing a game, looking at the scoreboards, and reading the help information provided. The first screen also shows the highest score achieved so far on the device. On selecting the Play option the game board, Figure 3.7, is shown to the users.

This includes the time they have remaining on the top right, their current score on the top left, the item they must return to its correct place in the

Figure 3.7: Fruit version (APT) left and Animal version (App Store) right.

centre, and the four locations to which the item can be dragged located at the four points of the compass.

In early versions of the game, the timer would start immediately and the overlay of the items each location accepted faded out over 5 seconds. In internal testing this proved to be too difficult for users to understand initially, so the released versions of the game do not start the timer counting down until the user has placed the first item into the correct place. The overlay then slowly fades out until the user has placed 5 items. If an item is dropped onto the wrong place it is animated back to the centre of the screen. The user is unable to move the item during this animation which, as the timer is continuously counting down, is the penalty for a mistake. The game is over when the timer reaches zero.

However for each item placed correctly in under a second an additional second is added to the timer. In order to make the game more challenging the game board changes every time 5 items have been placed in one of two ways: either the items accepted by the 4 locations change to a new random four from the set of twelve, resulting in an overlay of the new items fading away after 3 seconds, or the locations are rotated one position clockwise or counter clockwise.

Initially, these two conditions were given an equal chance of occurring.

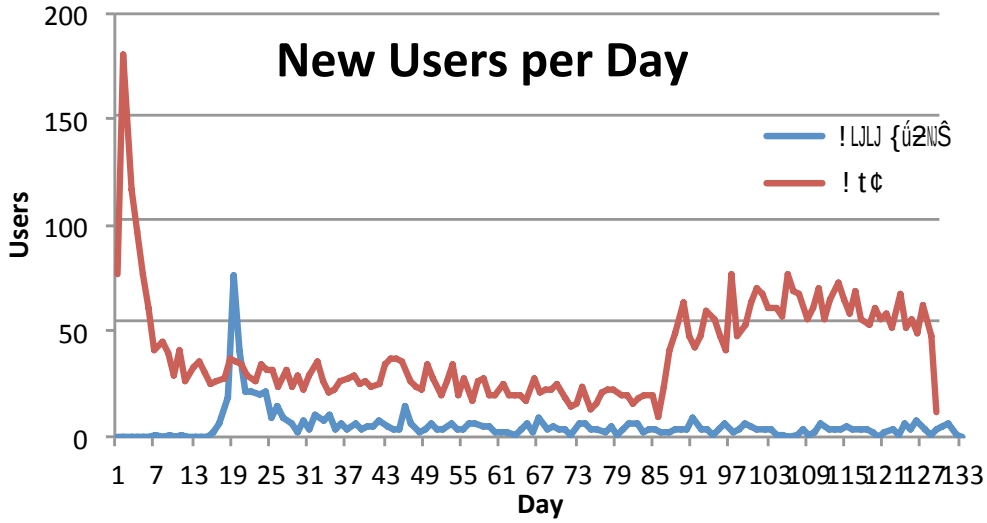

Figure 3.8: New users per day for each application.

However, the internal testing showed that the rotation condition was significantly harder for users to adjust to than the new items condition. This led to a change in the game whereby the chance of a rotation starts at 20% and increases as the user's score increases to an 80% chance — causing the game to get steadily harder as the user increases in skill.

One version based on animals was submitted to the Apple App Store, and the other based on fruit was submitted to the largest of the APT repositories. Each application was submitted to its respective distribution method on the same day. The App Store version was rejected twice in succession on submission, explaining the slow start to its user numbers. It was first rejected for the artwork of the large and small icons being too dissimilar and then for requesting the user's location without an obvious benefit to the end user.

Both problems were addressed, first with a change in the artwork and then with the addition of a country-based score board which translated the user's GPS location to the country using a reverse geocoding service running on the game server. The application was resubmitted within 24 hours each time. The new scoreboard was also added to the APT version of the game and released on the 7th day of the trial. However, the effect of this on the graph above was small as the application was still experiencing its initial high visibility.

These rejections resulted in a 17-day delay before the App Store version was available for download. The application was released on the store late at night on the 17th day of the trial with the large spike on the map being

seen on the 18th day. No more publicity was done for either version. As can be seen from Figure 3.8, the APT version of the game, available only to jailbroken devices, was more popular on all but a single day. In total, the APT version received a 10 times higher number of downloads and continues to gain on average 10 times the users each day.

3.3.1 Jailbreak Effects

The spike shown on the 91st day of the trial on Figure 3.8 represents a regular cycle seen across all applications released via the APT repositories that have been examined; the release of new iOS versions and the subsequent release of user level jailbreaking applications.

As Apple releases each new version of iOS, the community of developers who provide the jailbreaking applications must find new security flaws to exploit in order to alter the operating system to accept homebrew applications.

During this lead time a large number of jailbreak users will update to the latest version of the OS ahead of the release of a jailbreak – meaning they no longer have access to the APT repositories. So, when each new OS update is released, the number of jailbroken devices in the wild reduces for a period of time as a proportion of users who are not using jailbreak software for 'core functionality' – such as unlocking the device from its initial carrier – are likely to update the phone to the latest version of the operating system before an exploit has been identified and released. This causes peaks and troughs in the number of devices that have the ability to download and run software from such repositories, which directly affects user numbers. When a new jailbreak is released the publicity surrounding this drives large numbers of users to enter, or return to, the community at the same time.

3.3.2 Demographics

Each application also asked the user for simple demographic data on the first run and, when they accessed the location-based scoreboard, recorded the country in which they were playing the game.

As can be seen from Figure 3.9, while the number of users is much greater for the APT version of the game, the gender split is much more balanced than that seen in the App Store. Looking at the age demographics, it can be

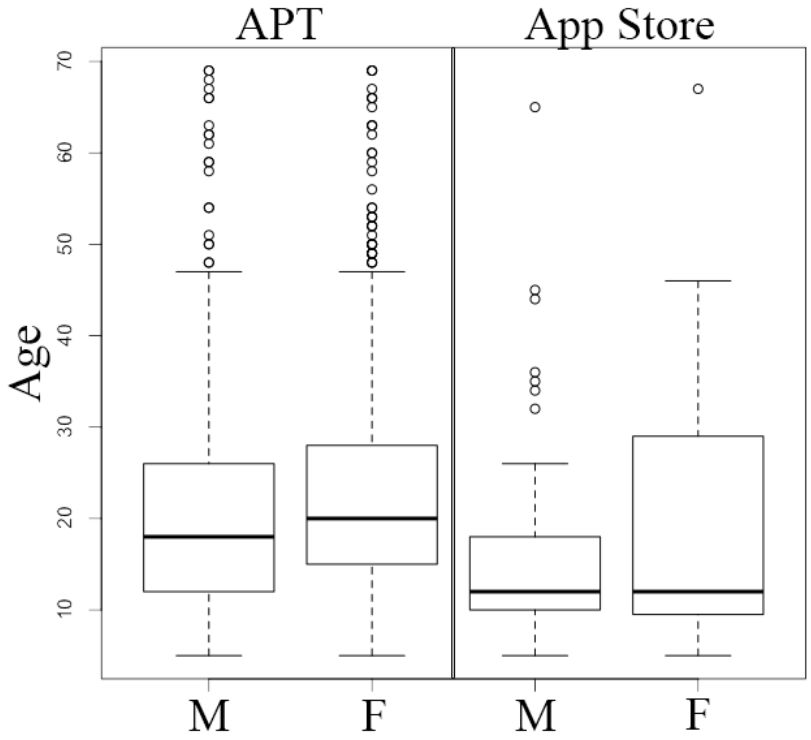

Figure 3.9: Gender spread of users and ages by gender for the Fruit (1432 users) and Animal (237 users) games.

seen that the median age is consistently higher for users recruited through the APT repository — possibly due to the higher technical barrier of entry.

The geographic spread of the users of each application where such information was available was compared to give the charts in Figure 3.10. 56% of users of the App Store version of the game agreed to share location data compared to 41% of the users of the APT version. As can be seen below, the

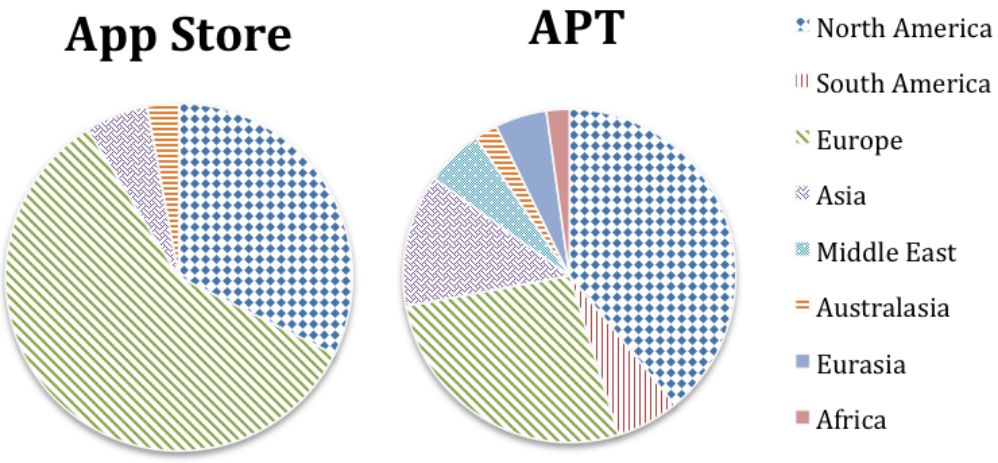

Figure 3.10: Geographic distribution of users by application.

larger user base of the APT version resulted in a larger spread of countries covered – the number of users in developing countries is higher than may be expected, this could in part be down to the necessity of users in a country without an official carrier for the device to jailbreak in order to unlock the device from its original, foreign, carrier. It is also interesting to note that while South American users account for the 4th largest in the APT version they are not represented at all in the App Store version, this is due to the fact that at the time of release there was no official app store for users in South America.

3.3.3 Usage and Engagement

An important consideration when determining how useful any particular set of users will prove over the course of a user trial is the level of engagement they have with the application and with the trial process itself. Any measurement of the activity of participants is dependent upon the application they are using, the methods through which they are participating in the user trial and the questions under examination by evaluators.

Figure 3.11 shows the percentage of total users, along the Y axis, who have used the application more than the number of times shown on the X axis. From this graph it can be seen that 60% of users of the APT version of the game used the application more than once, as opposed to 55% of users of the App Store version returning to the game a second time. The usage tails off with 11% of the users recruited via the APT repositories launching the application more than 10 times compared to 9% of those recruited via Apple's App Store.

The length of each session was also notably different between the two versions of the application, with the users of the App Store version averaging a session length of 62.8 seconds and those playing the APT version of the game averaging 78.3 seconds per session. The game dynamic being 'beat the clock' means that the more skilful a player becomes the longer their game play will last, suggesting that the population of users with jailbroken devices are more regular game players.

3.4 Discussion

There are a number of research groups that initially retooled for Apple's iOS platform on the release of the iPhone in the hope of taking advantage of the powerful hardware, low platform fragmentation and popularity of the devices. This was done at considerable cost in equipment and time.

As has been shown, there are methods to distribute applications to a wide audience on the iOS platform even if the application falls foul of Apple's App Store policies. However, there are trade-offs that must be made when using a distribution method without the support of the platform developer. Applications distributed in this way are unable to take advantage of services such as the push notification servers, allowing information to be sent to applications without the need for them to run continuously in the background, provided by Apple, Microsoft and RIM. Apple's Gamecenter, RIM's Games and Microsoft's Xbox Live services all provide a lightweight social network centred around apps and their usage as well as achievement badges users can earn by completing in-game tasks and are only available to apps which have gone through their approval processes.

In order to partake of these advantages and distribute through official channels, developers must take into account the delays to releases put in place by the review process, the need to justify in terms of end user experience any data harvesting, the restrictions put on how the hardware can be accessed and the restrictions on the content and look and feel of the software. Review processes are also in place for the Windows Phone Marketplace tak-

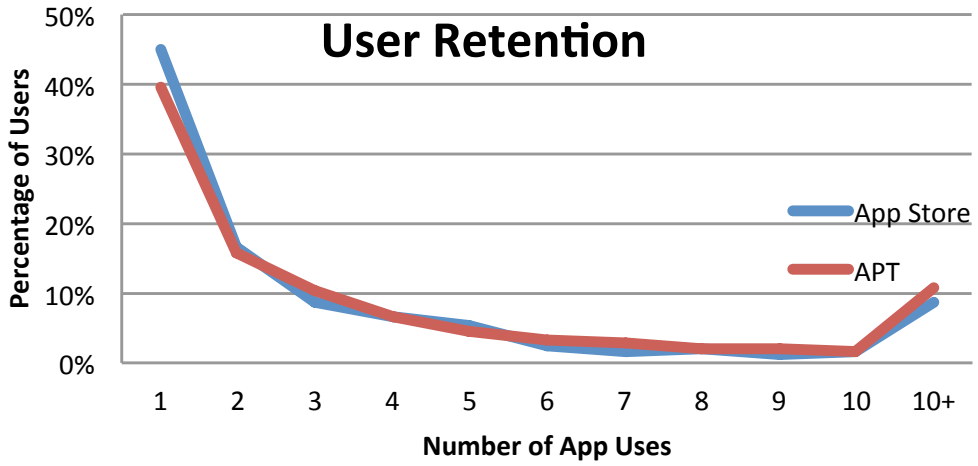

Figure 3.11: The percentage of users returning to use each application.

ing around 7 days and the Blackberry App World which can take up to 3 weeks.

The decision to take one route or the other can be dictated by the technology needed for the application to run, while the there are no restrictions on the software released through the Android Marketplace in terms of APIs used the structure of the operating system, with applications all running on a Java Virtual Machine, restricts direct access to hardware. Blackberry and Windows Phone 7 both have jailbreaks allowing unsigned code to be installed, with an APT repository system similar to that described above on iOS also available for Windows Phone 7.

Where it is possible to choose either path to distribution the pros and cons must be weighed. As will be discussed in Chapter 5, there are problems with user density within the jailbreak community both spatially and socially. There the users reported that they were unable to share the application with friends online, or to play co-located with family because potential users not only had to have the correct hardware and be interested in the software genre, they had also to have jailbroken their devices. As there is, as yet, no way for one user to send a link to another that prompts the direct install of an application on any of the jailbreak distribution options. Spreading APT repository packages between peers who have jailbroken their phones is more labour intensive than sharing applications released on the official App Stores. Users are also unable to share such applications with users who have chosen not to jailbreak their devices; the vast majority in all territories as seen from Figure 3.5.

This becomes a barrier for certain research questions or types of application. Those reliant on co-located use (e.g. Bluetooth P2P applications) or on a socially connected user base would be more suited to an App Store deployment.

There are also ethical considerations to be addressed when providing software via APT repositories. As mentioned above, the release of software on these repositories could be seen as encouraging users to jailbreak and although the act of jailbreaking has been proved legal, it is within the rights of both the manufacturer and the carrier to refuse to give customer support to customers with jailbroken devices.

As the only contact researchers would have with users would be necessarily after they had jailbroken their devices, addressing this seems to be a classic “catch 22” situation. The ethical responsibilities of researchers conducting

large-scale trials is discussed in greater detail in Chapter 6.

In this case, the question is whether it is the researchers' place to intervene by commenting on the participant's choice to jailbreak their device, presenting information on restoring their device in a context which they would not expect.

Deciding when and how to intervene when a participant is possibly in a situation where they may cause harm to themselves is dependent upon the magnitude of that harm and "the distinct rhetorical dynamics of online spaces. . . and the particularities of the contexts and communities." [165]

In the case of a jailbroken device, it can be argued that, as the procedure of jailbreaking is much more technically demanding than that of restoring a device, the potential harm is low.

In the same way that it is deemed ethically irresponsible to perform any action that can be seen as damaging an online community by encouraging its members to leave, so would distributing application to the jailbreak community which encouraged them to restore their devices to a pre-jailbroken state.

3.5 Conclusion

This chapter gives a comparison of different software distribution methods for Apple iOS devices. There is an initial outline of their characteristics, strengths and weaknesses, and a single application has been used as a comparative example.

The decision as to which distribution method to use for any application must take into account a number of factors.

The effectiveness of advertising, and the ability to target certain demographics through it, could be significant. While in a commercial setting the cost of advertising versus the benefit of increased revenue can be estimated, making such a calculation against the value to a research project of trial participants is more difficult. None of the published research in this area has mentioned using advertising to drive recruitment – indeed most explicitly state that they did not advertise – yet researchers regularly advertise locally to recruit participants for trials.

As the stores become larger and 'making a splash' becomes more difficult, the ability to either advertise for new users or build a relationship with users

of one application who can then be brought over to a new research project, or as in the case of Tap-It [113] new research questions within the existing application, will become increasingly sought after.

Finding ways to cultivate a relationship with user-participants and their willingness to engage with evaluators is necessary to determine which method of distribution is more suited to any specific area of research. More users using the application more often does not necessarily translate into more users willing to fill out in-application questionnaires or to be contacted for a more focused form of study, such as an interview.

- If the application needs access to hardware at a level Apple, RIM or Microsoft does not approve of, or needs to interact with other applications on the device in order to answer the research questions then the APT repositories are the only option.
- If the application requires the platform developer's network services, such as Gamecenter or Push Notifications, to operate then it must be distributed through an official App Store.
- If the application relies upon colocation of users then, due to the lower density of jailbroken devices, it would be advisable to distribute via an official App Store.
- If researchers are looking to explore spread or use across social networks, be they virtual like Facebook or traditional, the density of devices and the difficulty in sharing links to applications on the APT repositories would suggest that in this case an official App Store should also be used.
- If the application does not fall into any of the categories above there is no clear choice of distribution method. The researchers must weigh development freedom, faster releases and higher exposure of the APT repositories against the larger, denser, potential user base, network services and the relative ease with which applications can be shared within social groups seen in an official App Store.

The material in this chapter offers practical guidelines for enabling large scale experimentation, thus indirectly contributing towards a response to RQ1.

In the following two chapters, these guidelines are used to decide which distribution method is best suited to the application being released. In the next chapter, the decision was simple as the basis for the game, wifi scanning, can only be done by using APIs which would cause the application to be rejected from the App Store.

In Chapter 5, the decision was not as clear cut. While there were no technical reasons to restrict the distribution of this application, the time-sensitivity resulting from releasing an application that ran alongside a one-off live event — the FIFA World Cup weighed heavily in favour of releasing on the APT repositories. The risk of a critical bug resulting in a re-submisson of the application to the store that could take up to 7 days to be approved during a 30 day window in which data could be collected was deemed higher than the risk to the results of the lower density of users.

Chapter 4

Qualitative and Quantitative Data from Mass Participation Trials

This chapter describes an initial approach to answering RQ1. It focuses on a particular form of large-scale trial, which we call 'mass participation'. This term is explained in §4.6. This chapter describes the experience of making a free application available via the APT repositories described in the previous chapter.

The decision to release this application via the APT repositories as opposed to Apple's App Store was was simple to make based on the guidelines established in the previous chapter. The technology around which the game was designed, wifi scanning can only be done by using APIs which would cause the application to be rejected from the App Store.

This mobile multiplayer game, called Hungry Yoshi, was a new version of Feeding Yoshi – a seamful game ported to the Apple iPhone and updated. Feeding Yoshi's main trial was described in [29] as a "long-term, wide-area" trial "being played over a week between three different cities" in the UK.

A 'seamful' system [27] is one based on the design principle that instead of attempting to patch over or hide problems in services, such problems are fully exposed to the user and taken advantage of in the system design. In this example, the 'seams' in the free wifi coverage available in an urban area are made more visible to the user as they form the space in which the game is played.

The goal was to scale up deployments and trials as part of a project, Contextual Software, that explored system support for collaboration with communities of users in the design and adaptation of software to suit users' varied

and changing contexts [105]. Distribution in the App Store style, along with new tools and infrastructure, allowed for a trial that involved a much larger number of participants than before – participants who were far more geographically dispersed than could previously have been handled – and which lasted longer than any trial previously carried out by the group. At the time of writing, the current trial of the new Yoshi system has been running for 32 months and has involved more than 75,000 users from all around the world.

The following section describes 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 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 are described. Following this is a discussion of some methodological and practical issues, before conclusions are offered.

4.1 Hungry Yoshi

This section describes how Hungry Yoshi was evaluated and modified in the course of the trial. It discusses distribution, management, data gathering, analysis and redesign, and how they were coupled together in a form of iterative design suited to the large scale of the trial. It outlines how interaction with trial participants was conducted by researchers, how users interacted with each other, and how these interactions fed into a new version of the application so as to begin another design iteration.

As mentioned above, Feeding Yoshi [29] was a mobile multiplayer game for Windows Mobile PDAs. It was re-implemented for the Apple iPhone and given the name Hungry Yoshi.

This mobile, location based game uses wireless network infrastructure as the basis for a shared, global game board.

Each player's mobile device performs regular scans of the WiFi access points detectible in the current area, classifies each of these access points according to its security settings, and displays each 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' affording the growing of fruit over time.

Yoshis ask players for particular fruit, and players score by collecting the desired fruit from plantations and bringing them to the yoshis. Yoshis also provide seeds that enable players to grow new fruit in empty plantations, a shared resource for local players.

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 iPhone providing data connections over its cellular network, the system could 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 differing 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.

An added functionality in the iPhone version of Yoshi is the ability to exchange pieces of fruit for a small cost. Players are able to insert fruit into a fruit swapper, Figure 4.1 *(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 or in exchange for points.

§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 implemented as new players might join in at any time and could be months behind the early users. The score tables are divided into a section showing the top players, and a second section showing the players around the user's current position.

Figure 4.1: The list of nearby yoshis and plantations *(left)*, a yoshi screen *(centre)*, and the fruit swapper *(right)*.

4.1.1 Distribution

Hungry Yoshi was released in early September 2009. At the time of writing it has been publicly available for 32 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 the announcements of the trial, e.g. in interviewing one of Yoshi's users, it was 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, some of the enthusiasm of amateur and professional writers who regularly scour the store for new applications to try and discuss can be harnessed. The inverse is also true. In releasing research prototypes this way researchers are also competing for this enthusiasm, and the associated free publicity that comes with it, with the community of commercial and amateur app developers within the same space.

The publicity received from being included in such a blog could be seen

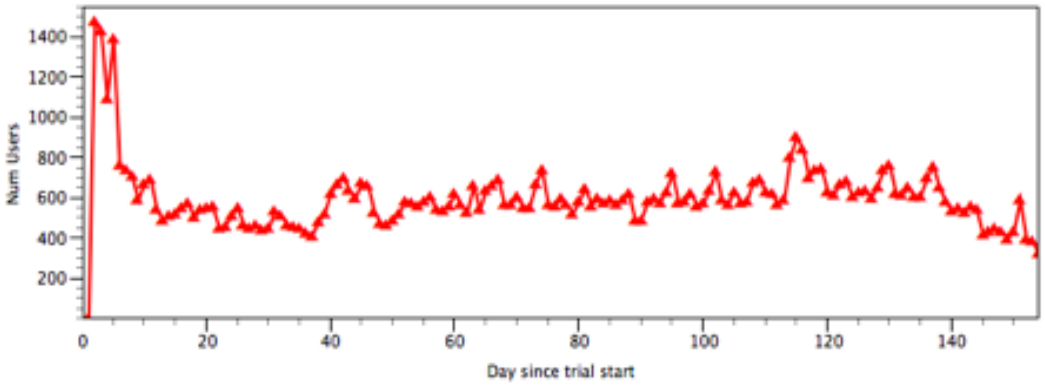

Figure 4.2: Number of downloads of Yoshi per day since release.

as an artefact from releasing an application in the so called 'Golden Age' of app store trials. Inclusion here was based upon Hungry Yoshi being a location based game, matching the style of application they wished to write about at that time. As these stores become larger, attracting more developers and software companies, it will become increasingly difficult for researchers to compete directly in terms of design and 'polish'. However, research applications taking advantage of new 'soft' sensors or interaction techniques will have the advantage of being novel even within the larger stores, allowing them to stand out from the crowd.

The goals of industry and academia are very different when releasing an application via an app store. The goal of profitability raises the minimum number of users that constitute a 'successful' release of an application much higher than would be necessary for answering most research questions explored in this way. It may become necessary to actively publicise the release of research applications, through press releases from the research institution or online advertisement campaigns for example, or to foster a community of users willing to use research applications on their own devices to provide the core users for a number of projects or research groups. Research on the best methods for attracting a suitable number of engaged users in any specific demographic is an area for future work, however it is still possible to recruit thousands of users with minimal publicity.

Figure 4.2 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. 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

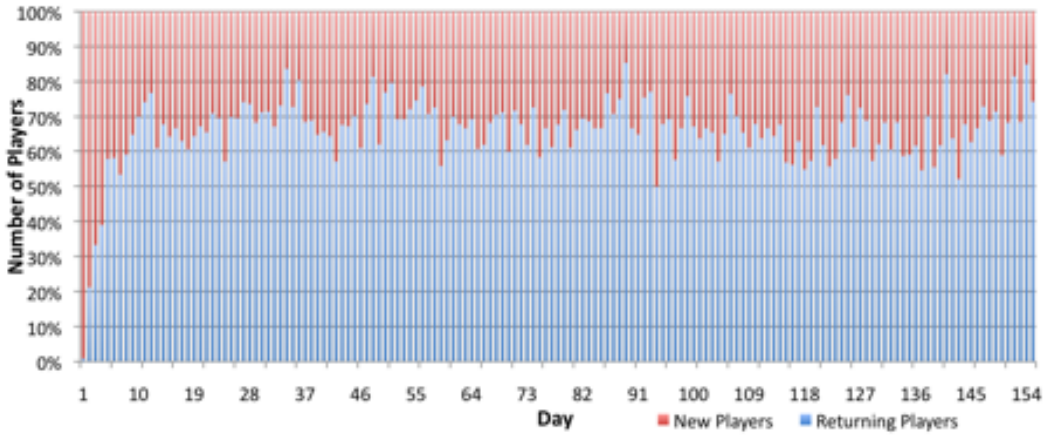

Figure 4.3: Proportion of new and returning players per day of the trial.

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 291,425 downloads in total. This figure includes people updating to new versions of the game; 196,617 unique downloaders were recorded. Figure 4.3 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. 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, it is stated that log data will be stored securely, that users can opt out at any time, that the data logged about them will be destroyed 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.

At the time of writing, 75,818 out of the 196,617 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 75 thousand, many only briefly interacted with the game, but 42,919 played for long enough to produce log data that could be studied. Although this represents only around 22% of the total number of downloaders, the number of players is still very large. Interestingly, this percentage has continued to rise from 9% of downloader in the initial 6 months of the release playing the game long enough to generate enough data for analysis.

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 in 2006. Results of the quantitative analyses are covered in the following section.

4.2 Quantitative Analysis

To aid the evaluation of Yoshi, system log data is generated from every trial participant's phone. The system makes use of the SGLog logging framework (described in detail in [105]), 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. Figure 4.4 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.

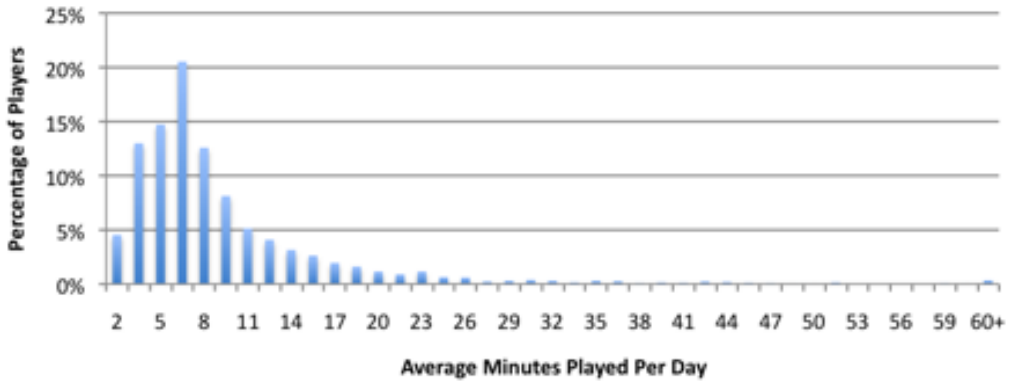

Figure 4.4: The distribution of players' average system use per day, with a mode of 6 minutes (20.5% of players).

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 [29], 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. Figure 4.5 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.

Daily patterns can also be seen, 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 §4.4.1

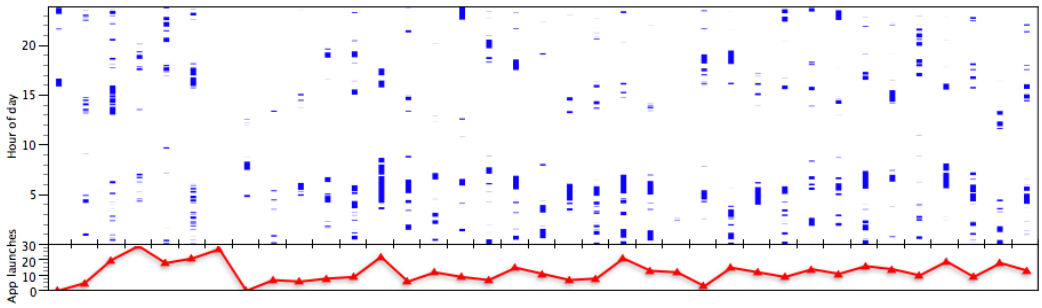

Figure 4.5: A visualisation of one of the most active players' use over the first 100 days of the trial. In the upper section, the x-axis shows days since the trial began and the y-axis shows the 24 hours of the day, with blue shading showing the periods at which the participant was playing the game. 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.

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 researchers 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 users. Two mechanisms were used: tools for communication within Yoshi, and communication via a social networking web site.

4.3.1 In-game communication with users

Rather than have communication with users happen in a way that clashed with the user experience, a bi-directional communication channel was built into the functionality of the Yoshi game. §4.1 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, researchers 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 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.

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 observed in their log data or during 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 there was no access to devices to update the software on them, questions could be altered at any point during the trial. Once edited, the new task set becomes live immediately, thus supporting adaptation of research interests.

The task and token-earning functionality proved popular with users, with 57,067 responses in total.

Before the trial, it was unclear 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 3,773 times, with all but 121 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 [237].

4.3.2 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.

The use of Facebook, a popular online social networking application, was selected as a means of supporting such interactions. Facebook has more than 900 million active users, 50% of whom log on to Facebook in any given day [10], 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 (called 'Lite mode').

Though it was still desirable that non-Facebook users be able to play the game, users were encouraged to log in through the Facebook Connect method to provide the benefits outlined above. As such, use of the fruit swapper described earlier was limited 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, users were able 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 the 42,919 users who agreed to the terms and played the game, 20,136 elected to use the Facebook login, including 44 of the top 50 scorers.

In addition to providing a login mechanism, content on the Facebook site itself was used both to provide features for the user and in contacting users to aid in the management of the trial.

A Facebook application was created — 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, messages could be sent to their Facebook accounts, and a forum was set up for players to communicate with each other and discuss potential new ideas.

4.4 Qualitative analysis

§4.2 described quantitative analysis performed on log data. With such data, gaining an in-depth understanding of individual player behaviour is challenging. While it was possible to visualise various aspects of play, this did not necessarily make a player's motives and reasoning comprehensible.

Here is described the allied forms of qualitative analysis, centred on interviews that allow the exploration and clarification of issues more adaptively than if, for example, an on-line questionnaire had been used 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.

4.4.1 Interview process

The process of interviewing when participants are spread 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, here there was no direct contact with users at the beginning of the qualitative analysis process.

Although all the users had agreed to a series of terms before playing the game, that explained that they might be contacted, 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 8,500 users at the time the interviews were conducted gave an opportunity to focus on interviewees that were deemed the most relevant to a design issue or potentially significant in terms of use and user experience. For example, the most active players could be chosen, 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. Selecting participants for interview began by using the information publicly available to application developers via the Facebook API to first filter the participant database to show only those over the age of 18. Thereafter, visualisations of log data to examine

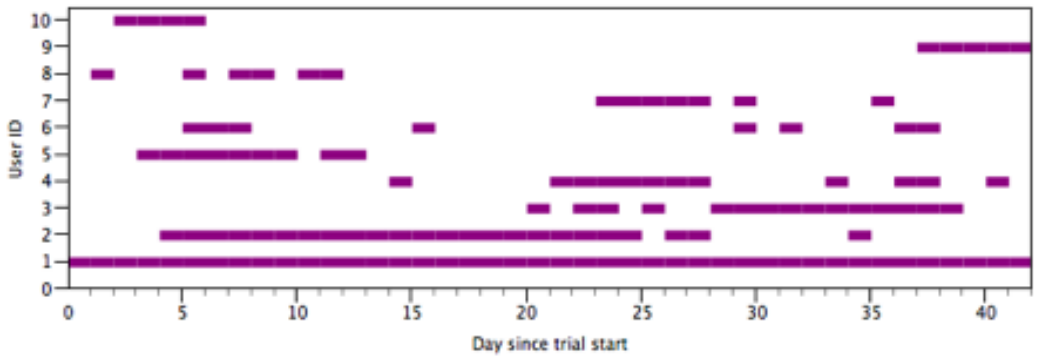

Figure 4.6: A snapshot from a tool used to select participants for interview, showing which days each participant used the application. From a chart showing all the trial participants, ten users have been selected who exhibit contrasting patterns of use. This figure shows the filtered set.

participant use over time were used. For example, Figure 4.6 has a separate row for each participant, and shades days, shown on the x-axis, if the user played the game on that day.

As there was interest in speaking to a set of users with a diverse amount of engagement, the visualisation was used to select rows with contrasting patterns.

Figure 4.6 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 using these methods of selection to interview a set of players which showed a broader range of activity in playing the game gives more breadth to the interview participants.

Having identified the interviewee set of choice, Facebook again proved a useful tool in making contact. Users were sent a message, enquiring whether they were interested in taking part in an interview over VoIP or telephone in exchange for $25 of iTunes or Amazon vouchers. 10 of these players were interviewed, 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, there was a noted 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 dropped as the trial continued.

It can be speculated 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 Figure 4.5, other tools were useful in interview preparation. The tool seen in Figure 4.4, 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 the interviewer was, and typically lasted between 15 and 45 minutes. All were transcribed afterwards.

4.4.2 Findings from analysis of interviews

In order to explore the research issue of running mass participation trials, as many questions about the trial mechanism itself as about the application were asked.

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 messages with each other, A discovered that B lived in a different continent. Their 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 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 the 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 contacted through 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. This is maybe because they had been provided with 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 this 'sampling' of users.

By targeting users with the task mechanism, Facebook messages and email it was possible to quiz those who declined interview requests on their reasons for doing so. The response rate was low but those received 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 experiences of earlier trials of other systems, it was observed that players had more good will towards 'giving something back' than had been 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 greater engagement with users that was achieved using the various methods of communication built into, and around the application the opportunity to open the design process was made available – allowing them to participate enmass.

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.

The tension she identified was in this case a secondary tension between the need for the plantations to generate fruit for the community of players and her desire to change the type of fruit being produced for her own gain. Alleviating this tension could be done by providing new tools to the users, allowing them to manipulate the plantation objects in a new way, or by

changing the rules around which the current tools allowed the planting of new types of fruit.

Following this interview, it was agreed that this was a valid tension that should be addressed if it reflected a common concern or problem among users. The task mechanism was used again to consult the user-base at large as a lightweight substitute for an Action Workshop.

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. Three options were presented:

A: Leaving the game unchanged.

B: Players could burn empty plantations to stop them re-growing (as suggested by the interviewee).

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 the design process.

On detecting that Helen had installed the new version, she was contacted again to gauge her reaction towards the new feature and she replied positively, agreeing that the version implemented, rather than the design she had suggested, was the better of the new options. The addition of a new tool, allowing users to burn down plantation, and rules as to the price in terms of points a user must pay to use it changed the relationship between the community and the plantation objects. In order to ensure that this change had not raised unforeseen tensions within the model that would be detrimental to the user experience another vote was included on the new feature, consulting the opinion of the user-base at large after they had had a chance to use this feature.

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 using the same mechanisms. 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 generation of hardware 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. In the first five months the software was live, seven versions were 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.

4.6 Discussion

The tools and techniques described in the previous sections let us carry out a participatory iterative design process but at a much larger scale than normal. Methodologically, when this novel mass participation approach is compared to large scale trials, there are 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 it was sometimes expedient to use common database query tools to work with the 'raw' log data, it was beneficial to also develop custom visualisations to better understand patterns and detail in the data and to choose where to focus requests for interviews. Compared to more ethnographic trials that involve local participants who are paid to use software in a trial, it can be argued that this process of 'recruitment' led to more realistic conditions in that users were using software that they themselves chose to use — without inducement or obligation on their part to keep using it if they did not want to.

However, this advantage has to be weighed against issues such as the inability to gather data from those who dislike the application, and the reduced knowledge of local context and culture.

In practical terms, this method incurred expenses in terms of development time and interviewer effort. The language skills of the research group were put to the test with the creation of French, German and Japanese internationalisations – giving greater access to those for whom English is not their first language. Initial worries about being able to interview users with limited English and a first language out with 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 a potential bias should be noted: 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 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.

Taking careful note of the sender's time of day when a message was created, and addressing his or her perception of the passage of time until the response was sent, 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 the trial, it was found that relative wealth scales also played a part, with the level of entry to the 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, as the average income of a country decreased so did the density of Yoshi users there. 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 §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 [47] 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.

4.7 Conclusion

In this chapter, a mass participation method for running large scale trials with the goal of combining the benefits of large-scale 'field surveys' and ethnographic or activity theory based methods was explored with the running of a worldwide trial of the game Hungry Yoshi. A central aim of this methodology was to push the upper limit on the number of participants as far as possible, while still combining quantitative and qualitative approaches in ways that usefully and efficiently fed into the redesign process.

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 were used. The benefits of such mechanisms include an increase in the numeric and geographic scale of the user-base combined as well as an increase in the engagement of these users with the trial process.

The worldwide nature of the trial meant that tools and methods to maintain awareness of participants had to be adapted. The use of quantitative and qualitative assessments to assess the activity and engagement of the user-base, and using this to perform targeted interaction with participants, how that interaction took place on a variety of scales, and how feedback mechanisms were embedded within the system and their use encouraged was described.

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 allow the orchestration of a trial involving a very large number of participants for a long period of time to have relatively quick redesign cycles set within that process.

This trial and its evaluation also fills a gap identified in the current literature on large scale mobile trials in §2.2, where detailed focus upon the user experience and how a mobile application's fit into the everyday life of a user can be examined at a distance.

Here follows a summary of guidelines based on the experience of releasing Hungry Yoshi 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. An order of magnitude less participation for every step 'out of the game' users were asked to take was observed.

In retrospect, the amount of researcher time and energy that was required to collect the rich interview data from remote participants was disproportionate to the quality and the utility of these transcripts. While it is not opposed to the methodology outlined here to analyse each textual response in detail, to arrange hundreds of remote interviews or to travel the globe meeting and observing the participants in the trial – these courses of action are not practical.

Indeed, the cost in researcher hours per interview conducted in the course of this trial was estimated to be at least a factor of 4 higher than when using local participants.

Even the qualitative data that proved relatively easy to collect, the free text answers to the tasks, was difficult to analyse — the time and effort that would have been involved in screening, coding and analysing the 57,067 responses that have been returned at the time of writing would far outweigh their value in an iterative design process.

There are also ethical considerations that must be taken onboard as this mass participation methodology is formed. As has been shown, many users do not even know that they are participants in a user trial, much less understand what this means. For trials run using this methodology to be successful, they must be both productive in terms of research and acceptable to the user communities they rely upon.

This method is therefore put forward as a partial answer to RQ1. While it has been shown that qualitative and quantitative data can be collected, analysed and fed back into the redesign process from a large scale user trial, the practical and ethical difficulties mean that further refinement is necessary to give a full answer to the research question.

In the following chapter another worldwide release is described, but in this case the trial is run alongside a more traditional trial involving local participants. The mitigation of the problems of difficulty of access to high quality qualitative data, the self-selection bias present in the participants who do agree to be contacted and the detailed analysis of an unaware participant by use of a local group of users will be discussed.

Chapter 5

Hybrid Mass Participation Trials

The rise in opportunities for researchers to be able to conduct large scale 'field trial' research by taking advantage of the increasingly powerful and sensor rich devices more and more people carry about with them every day, described in Chapter 1, has begun to draw researchers into this area.

However the vast majority of publications involving such deployments of software, as shown in §2.2, have forgone the rich qualitative detail available from a local user trial completely as they moved to this larger scale.

The methodological tensions between observational theories, such as ethnography and activity theory, and the normal practices in these large scale deployments, were discussed in §2.4.3. The challenge is to find a practical synthesis of the two, harnessing the strengths of each for the purpose of design.

In the previous chapter it was shown that there are methods by which detailed qualitative data can be collected from the participants in a mass participation user trial. However, in using a wholly remote set of participants there were a number of practical and ethical problems.

Without the traditional researcher-participant relationship, users were less willing to be contacted out with the application, and bias was inserted into the sample of users who agreed to interviews as they were necessarily only drawn from those who were highly engaged. The effort expended in arranging and conducting the interviews during the trial of Hungry Yoshi was much higher than when using local participants.

The participants were also unaware that they were part of a user trial, despite best effort being made to inform them in the manner standard across the trials outlined in §2.2 – the Terms and Conditions page when the ap-

plication is first launched. Traditionally discharging the ethical responsibilities researchers have towards their participants stems from the negotiated understanding of the trial procedure. Confronting people who had no knowledge that their data was being analysed with hypotheses about them and their lives in a situation where the circumstances and mental state of the participant cannot be checked before-hand, or mitigated after the fact in a debrief if they choose to stop using the application, makes ensuring ethical conduct extremely difficult.

This chapter describes an method to counter a number of these difficulties with a hybrid methodology combining a global software release with a concurrent local trial. A phone-based game, World Cup Predictor, was created to explore the uptake and use of ad hoc peer-to-peer networking, and to explore evaluation using this hybrid trial method, i.e. combining a small-scale local trial with a mass participation trial.

This hybrid method allows for locally observed findings to be verified, for patterns in globally collected data to be explained and addresses ethical issues raised by the mass participation approach. Based on this study and that discussed in Chapter 4, a set of guidelines to researchers working in this area is provided.

Particular aspects of the research being undertaken may be best suited to investigation using one or other of the trial groups. In some circumstances the two groups can be used together to achieve greater insights than could be gained from studying both groups separately.

The application, World Cup Predictor, was designed to run alongside the 2010 FIFA World Cup, and looked at users' real-world uptake of peer-to-peer data transfers. The local trial involved 11 users, and a further 10,806 registered users were recruited via the software's release on a mobile application repository.

In this chapter we reflect on this hybrid trial, to investigate whether these different methodologies can be effectively combined, explore productive ways of managing the tasks best suited to each, and discuss whether this hybrid approach is sufficient to address the shortcomings identified in Chapter 4. This chapter culminates with a set of guidelines to aid other researchers who are interested in performing such a study.

Taking a sample of data from a larger population and conducting interviews is not, in itself, a novel approach to research. In fields such as sociology and market research it is not uncommon for a survey to be followed by

targeted interviews, either individual or together as part of a focus group, to gain more in-depth qualitative data [66]. This approach has been applied in mass-released mobile applications [6], although without the methodological detail or discussion provided here.

The World Cup predictor application described in this chapter was designed to apply this large-scale deployment model to explore the viability of research applications based on mobile ad hoc networks (MANETs) using commodity mobile phones in uncontrolled environments. Research of MANETs has been extensive [141], yet few applications of MANETs have taken place outside of simulated environments [147] as a substantial density of users is needed to ensure successful transfer of data through a community via opportunistic, face-to-face encounters [141].

MANET research systems have used mobile clients as data carriers [99], exchanging information as they moved around or used ad hoc networking to provide awareness of others in the environment [191, 50]. Although some of these systems underwent user trials, they were small-scale in terms of participant numbers.

5.1 Difficulties Encountered In 'App Store' Trials

As outlined previously, the provision of 'app stores' on several mobile platforms in recent years led to several research projects using these methods of deployment. Although most report positively on the benefits gained using this methodology, several potential shortcomings have been identified.

One was identified by Henze et al. in their use of the application Tap-It in large scale deployment for an experiment that sought to isolate 'cause and effect' on a single task [116], the type of experiment commonly undertaken in a controlled lab setting. As Android runs on a great number of devices from many different handset manufacturers, it is no surprise that the authors report that their captured data came from 40 different types of device, with different processor speeds, OS versions, screen resolutions and physical dimensions. This lack of uniformity could easily be a problem in certain types of studies, if the goal is to compare behaviour among different conditions.

The recruited users might also not be as diverse a group as it might first appear. As noted in Chapter 4, relative wealth scales also played a part, with

the level of entry to the trial set at having an Apple iPhone, so although a large number of users might be attracted from a wide geographic spread, the platform is likely to attract a group of users that is relatively homogeneous in economic terms.

Additionally, although the recruited user base might be considered more representative of the audience such an application would attract, there is no way to gather data from those users who do not like the application. Users will judge the software by the same standards as they judge commercial software they download, meaning researchers can expect a large proportion of users to download and 'browse' the software but not continue to use it if they do not see a benefit. They are unlikely to feel a responsibility to keep using the software in order to complete an academic trial.

Several other types of difficulty might arise in using this mass participation approach to Ubicomp trials. As mentioned before there can be difficulties in both finding participants willing to consent to speak to researchers and in the practicalities of conducting these interviews. Whereas offering in-game rewards for answering quick questionnaire-style queries led to a reasonable return rate, the same strategy did not persuade people to consent to telephone calls.

The offering of financial incentives was similarly unsuccessful in encouraging interviewees. Once willing interviewees had been identified, another round of difficulties began as attempts were made to timetable a schedule for VoIP calls with users who lived in different time zones. These had to be arranged at times best suited to the interviewees, working round work, family and social commitments and in many cases necessitated researchers staying up until the early hours of the morning.

It is then especially unfortunate that remote participants were found to be very unreliable in attending arranged interviews, with many of them failing to answer calls made at pre-arranged times.

This might be due to the nature of the researcher-participant relationship in this style of trial; there might be less social pressure to answer a call from an app developer with whom you have had the most cursory contact than there would a researcher who you had met and whose hardware you were using.

In addition to the struggles of finding interviewees and arranging interviews, the data gained from such encounters is less rich than would be captured

from those conducted face to face, due to increased difficulties in establishing a rapport and the inability to capture nonverbal information.

While the quality of qualitative data collected between telephone and face-to-face interviews has been seen to be comparable [230], the gap between email and other asynchronous web technologies, and synchronous voice has been shown to be significantly detrimental to the credibility and trustworthiness of the researcher – and as a result the quality of the data gathered [126].

Synchronous video chat has been shown to provide higher trust ratings in certain circumstances [92], this is generally seen where the participants in the videoconference had previously met face-to-face [4]. Modern computer based video conferencing also includes collaboration tools, such as shared drawing areas, useful in communicating clearly when conducting a qualitative interview [109]. However, this puts extra constraints on interviews – shrinking an already small pool. Synchronous video chat is rarely used in a mobile setting [111], and on a mobile device these collaboration tools are less effective, where they are present at all, due to the limited screen space and the requirement to hold the device in the correct way to successfully send video. The effectiveness of the communication when video is compared to audio only channels has been shown to be negligible, and in some cases detrimental, to the conversation [199]. Using video also raises the technical literacy necessary to be an interviewee in the trial and could constrain their ability to take part to when they are at a computer with the necessary hardware and an internet connection. Most of the interviews conducted in Chapter 4 were conducted by calling the participant's mobile phone.

A lightweight asynchronous voice feedback module was developed by the author and included in the release of an application called World Paint, a simple location based game where users can turn on and off a coloured paint brush and leave a trail on the world map when they move. This feedback module allowed for users to leave either text or voicemail-like messages for the development team, who could reply with text. These messages were arranged as conversations with the user. This was not as well used as had been hoped. The majority of audio clips uploaded to the server seemed to be from users who were using the feature as a voice memo system – with an unexpected number of people recording and uploading versions of their favourite pop songs presumably with the expectation of receiving constructive criticism, it was decided best to reply only with praise in these

circumstances.

It was also of interest to us to perform repeated interviews at regular intervals throughout the trial, to follow up on points raised and learn about how usage of and feelings towards the system changed over time. However, in many cases this would have been impossible as users can stop using the system at any time, with no implied obligation to keep playing for the duration of a trial.

Although very large amounts of quantitative data were being gathered, the reduction in the amount and quality of qualitative data meant it was harder to explain the patterns observed or infer users' motives: what was happening in the trial could be seen, but there were barriers in the path of discovering why it was happening.

Running research trials using these app store repositories also raises a number of additional ethical questions not inherent in more standard deployments among local users.

Possibly chief among these is the notion of informed consent: do the users of these applications understand their role as participants in a trial and how data may be recorded on aspects such as their usage of the software, their unique identifiers and their location? As seen in §4.6, 70% of users did not know that their application was part of a research trial, despite the presentation of a terms and conditions page on first launch of the application, echoing previous findings on the number of users reading terms presented in desktop applications by FAST.

Even if users did want to read this information, language barriers may prevent it; although translations were provided in four major languages, this will not cover all the users likely to download a globally released application. Additional ethical concerns arise in researchers not being able to know their users, or to verify submitted demographic information. There is no way of knowing whether a user reporting to be above the legal age of consent to enter into a contract in any given country is being truthful, or whether a child has lied about his or her age in order to gain access to functionality restricted to adults.

The age of a subject is of particular importance when they are agreeing to interviews; in many institutions separate – and more rigorous – ethical approval of a project must be sought before engaging with minors.

The hybrid mass participation approach presented in this chapter attempts to address many of these technical, administrative and ethical difficulties

that can arise when evaluating a system using a mass participation method alone.

The following section describes the World Cup Predictor application and the hybrid trial methodology used in its evaluation.

5.2 World Cup Predictor

As previously stated, most trials of MANET systems have been based on simulations or small numbers of users. Using MANETs in ways that fit with the everyday interactions of users and the limitations imposed by current technology, so as to augment and support their sharing of information, has not been fully explored.

Here the techniques of the mass participation approach are applied to the study of peer-to-peer data transfers in real-world settings.

As one of the requirements of such a trial is to create a sufficient density of users, it is necessary to create an application compelling enough to gain a large number of downloads. Releasing a research application based solely on testing MANET communications was considered, an attempt at altruistic message passing through a number of intermediary nodes for example, but this would perhaps not attract huge numbers.

Rather an attempt was made to create an appealing application, which could be used by an individual player and in which it was possible to be confident in it being downloaded in large numbers.

Peer-to-peer data transfer functionality was then built into this single-player application as an optional mode, but incentivised with points and prizes, to examine the uptake and usage of these features.

Whether groups of socially-connected users would all download and use the application when it was released in this way, and whether similar or different uptake of MANET features among these users as compared to a local user group that was handpicked to have known social links would be observed was part of the research being undertaken.

5.2.1 System description

The World Cup Predictor is an iOS-based game created to run alongside the 2010 FIFA World Cup, a sporting event of global interest hosted in South

Figure 5.1: The initial app screen *(left)*, the news feed *(centre)*, and the prediction screen *(right)*.

Africa between 11th June and 11th July 2010.

The application allowed users to predict the results of World Cup football matches and awarded points for correct guesses: 3 points were awarded for getting a result exactly correct and 1 point for predicting the correct winner or correctly predicting a draw. Point tallies were accumulated for all users' predictions and collated in a global leader-board.

The game divided the World Cup into seven rounds, such that every team remaining in the tournament played once per round. The final two rounds consisted of only two matches each. Users were only able to predict results for the matches in the current round, this constraint being designed to encourage continued engagement as users had to interact with the application at least once per round to continue to earn points.

The deadline for submitting predictions for a given round was the kick-off time of the first match in that round. At the end of each match, the server allocated points to users with correct predictions and recalculated the leader-board.

The application also included peer-to-peer functionality, where a user could challenge another co-located player to a 'head-to-head' game.

In this mode, both players would predict the results of the same randomly selected subset of the current round's matches and the player getting the most correct would be awarded 5 points on the main leader-board. Upon challenging another player, an ad hoc connection would be formed between

the devices using Bluetooth.

This connection was used to transfer the prediction data locally between devices and, on the next occasion one of the users had a connection to the server, the details of the head-to-head were uploaded so that points could be allocated to the winner.

There were no restrictions on the number of head-to-head games users could play, except that each pair of users could only play each other once per round. The game rules were designed to incentivise usage of this feature, as users willing and able to play many head-to-heads would have a far higher chance of winning the prizes on offer.

5.2.2 Pilot Study

The Predictor application was initially trialled with a group of eight locally recruited users – a group of friends, who saw each other socially and were football fans.

The pilot ran alongside part of the local football season, from December 2009 until April 2010. Log data was captured, and participants were brought together for three focus groups during the pilot study to discuss their experiences and thoughts on the application.

The pilot was useful as a beta testing phase in which several user interface issues were identified and corrected. However, the main outcome was the lack of usage of the head-to-head functionality, with none of the seven users ever using it except when brought together for focus groups.

When queried on this issue, users suggested that they did not see each other regularly, and when they did meet face-to-face, they did not remember to do it.

Motivation was also suspected to be a problem, as the pilot study participants were paid a flat fee for their time, but no prizes were offered for winning the competition. Additionally, the head-to-head feature was not linked to the main game points total, so players winning a head-to-head would not receive any benefit to their status on the league table. The subsequent design sought to provide further motivation to users.

5.2.3 User trial

Although the World Cup Predictor study sought to investigate the uptake of ad hoc peer-to-peer networking via the application's 'head-to-head' feature, the primary aim was to test a hybrid mass participation methodology combining large and small scale trials.

As such, the World Cup Predictor application was distributed to both locally recruited participants and via an iOS APT software repository to users worldwide.

In order to encourage downloads, and particularly usage of the ad-hoc peer-to-peer functionality, the application offered a prize of £250 to the top player at the end of the World Cup. Smaller prizes of £40 were offered to the player winning each round, so that players coming to the game late or those who had not performed well in the first few rounds could still win a prize, and therefore would still be motivated to play the game.

5.2.3.1 Local Participant Group

As it was of particular interest how social bonds between users impacted on game participation and use, local users who had existing social ties were recruited.

In recruitment it was stressed that this group should have regular day-to-day contact, to maximise the opportunities for use of the peer-to-peer functions.

A smaller group of satellite players with no connection to the group of friends and colleagues was also sought.

This social topography of users was chosen to give direct access to the different types of users and use expected in relation to the peer-to-peer functionality; i.e. friends and co-workers using it together on a regular basis as well as serendipitous use between strangers. The recruitment of local participants was undertaken primarily by putting up posters in several locations around the city: public libraries, music shops and university buildings.

In order to find a group with day-to-day social connections volunteers were asked to recommend their friends or colleagues who would be likely to be interested in playing the game. In order to boost the group size, hardware in the form of Apple iPhones was provided for the duration of the study to members of the local group who did not own a compatible device.

Each participant was paid a nominal fee for their participation in the study, which included pre-trial familiarisation with the system, a brief visit from a researcher during the trial and an interview after the trial was complete.

Of the eleven participants recruited, five worked together in a shared office. Of the others, three were acquaintances of one another and the other three had no social ties to any of the other participants. Participants ranged in age from 18 to 37 years of age, with 10 males and 1 female.

All bar two participants owned the devices on which the trial software was run. As well as their payment, local participants were of course eligible to compete with the global users for the array of prizes on offer throughout the World Cup.

5.2.3.2 Global Participant Group

World Cup Predictor was released via an iOS APT repository on Sunday 6th of June 2010, with the World Cup commencing on Friday the 11th of June. An interesting aspect of running mass distribution trials is the difficulty in quantifying the number of users involved in a trial.

In the local trial, a user can be identified as someone who was given the software, paid for participation and interviewed about their experience, and the number of such participants can be reported with confidence.

With an application released through an online software repository, this becomes more complicated.

Statistics provided from the online repository state that the application has had 44,613 downloads. This figure includes software updates and reinstalls, so the same user might be included more than once.

By the start of the tournament there were 3,720 registered users, with this number increasing to 10,806 by the end of the World Cup.

Of these, 5,941 made at least one prediction, and 5,602 predicted in more than one round. The application is still available to download and even though the game has now ended, users have continued to sign up, with 23,671 registered at the time of writing.

On first launching the application, users were prompted to provide a username for use on the leader-board and an email address so they could be contacted in the event of winning a prize.

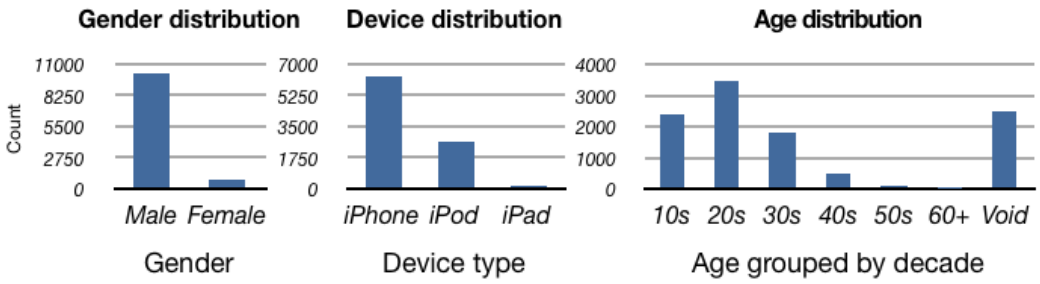

Figure 5.2: User Demographics.

Simple demographic information was also requested at this stage, with users being asked to input their age via a slider and their gender via one of two buttons.

There was no obligation to answer these questions and it must be taken into account that the reported answers cannot be verified. The results of this survey are shown in Figure 5.2, showing the spread of the ages of the 80% of users who provided an answer to the question and the distribution of the gender of the participants, which is heavily biased towards males.

Device information was logged automatically. Of the iOS devices on which it was possible to run the application, iPhones outnumber iPod Touches by around a factor of 2.5. Far fewer participants used iPads, which were only released shortly before the World Cup. Users must agree to the terms and conditions stating that their locations may be logged before using the application. A contact email address was also supplied for users to opt out of the trial at any time, and it was explained that all data stored about them would be destroyed on receipt of such a request. This information was presented in four different languages as in the previous trial described in chapter 4.

Before data is logged, the user must also agree via the standard iOS request for the application to use his/her location. The game would not be affected if this request was refused or if location services were turned off at any time.

The majority of users were based in Europe, North America and South America.

Fewer than 400 users played the game in Africa, the continent hosting the tournament, although there was activity recorded at five different World Cup stadia in South Africa. Both the local and global deployments of the application logged participant usage data.

The application feeds log data to local caches on the phone, which in turn

are opportunistically uploaded to a server using the same data connection required to run the game.

The data logged includes activities within the game, such as moving between application screens, and general contextual information, such as location. 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.

5.2.3.3 Qualitative Data Capture

As researchers did not physically meet the global participants, there were additional challenges in establishing a dialogue with them in order to gain data for qualitative analysis.

In addition to asking simple demographic information, users were presented with short questionnaires, allowing them to enter free text into a form within the application using the device soft keyboard.

Although not answered by the same high percentages of users as the more simple demographic questions, many participants usefully responded to this form of concise information gathering.

5.2.3.4 User Self-Selection Bias

Packaging the peer-to-peer functionality under scrutiny within a football predictor game may have introduced a bias in the sample of users.For example, the vast majority of users were male.

When conducting a mass participation trial, the body of trial participants is self-selected, in that users download the trial application from a public repository themselves. This is in contrast with more traditional techniques in which participants are directly recruited, or techniques used in industry, in which a recruitment agency ensures a demographic distribution in accord with what a company expects or desires.

The population of users resulting from a mass participation trial is likely to be more 'lifelike' in representing the types of users who would actively seek out and use such an application when compared to the inherent biases likely to occur when users are directly recruited.

However, these users, who elect to play a football predictor game, are not necessarily representative of the average population, or of the set of people

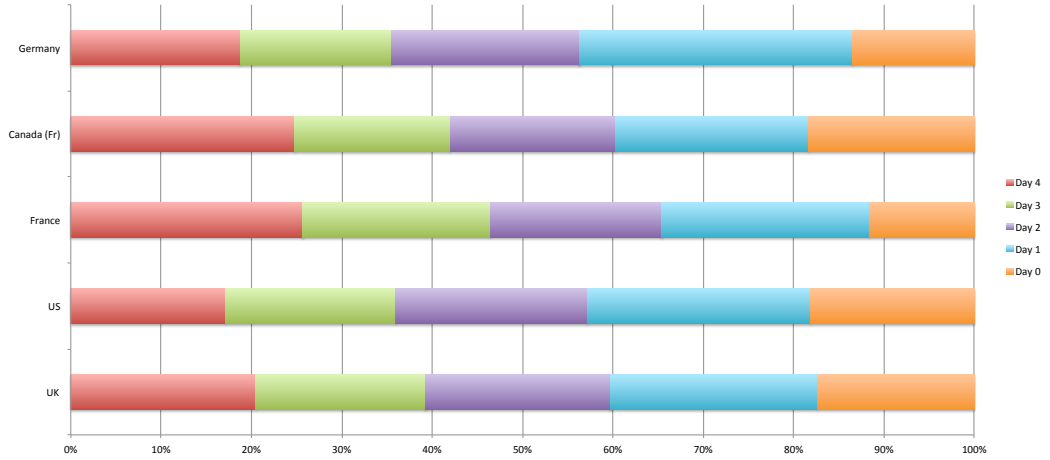

Figure 5.3: Graph showing, for each of a number of geo-cultural regions, how long before the deadline predictions were submitted.

who would be most likely to use peer-to-peer functions in applications in general.

5.2.3.5 Geo-Cultural Differences

Another interesting point noted in the data was a cultural difference in the behaviour of users with regards how long before the deadline they submitted their predictions.

Geo-cultural differences in user behaviour have, in the past, been measured by running multiple small scale trials in different locations [55, 110, 18], by purposefully recruiting from a representative minority group within the researchers' local area [5, 240, 241], or by using internet mediated surveys [137, 244, 164, 202].

Here the log data collected in the course of the trial was examined. The logged data on submission times was filtered and partitioned to create subsets in which the language setting of the device matched the national language in a given country (or region in a country) corresponding to the device's location. For example, figure 5.3 shows data from a number of countries as well as the French-speaking Canadians in Quebec. The data from the final match was omitted as there were only two days before this fixture in which players could submit predictions.

As can be seen from Figure 5.3, the UK and USA show slight a weight towards users predicting closer to the deadline, but the predictions were fairly evenly spread over the available time. This is in contrast to France

and Quebec, where a tendency to pick early is displayed, and Germany where the time picking was closer to the deadline.

This difference in the time of interaction with the application based on the culture of the participant could have an affect on how researchers interact with them during a trial. If the timing of use researchers expect does not generalise across cultures then cultural groups could be unevenly represented in responses to feedback questions, depending on when they are released and when the answers are analysed.

5.3 Usage of Ad Hoc Networks

The game was completely playable without using ad hoc networks, but had this functionality presented as an option for 'head-to-head' play.

The head-to-head mode was encouraged both through the scoring mechanism and as a means to challenge friends, providing a more overt social element to the game.

As the tournament progressed, the game mechanism encouraged head-to-head play more as the number of matches per round became fewer.

For the final two rounds, where only two matches were played, the maximum possible score through the main predictions game was 6 points, whereas each successful head-to-head game would gain a user 5 points.

It was hoped that this game mechanic would provide motivation to use this feature, but there were several obstacles that would need to be overcome before a player could use the head-to-head mode. A user would need to know somebody else with an iOS device, who was sufficiently interested in the World Cup to install the application and who could physically meet the user in order to establish a Bluetooth connection. These are significant demands to place upon a feature compared to the main game, where a single user only needs an Internet connection to play.

These results show very different usage levels between the local and global user groups. Of the global users, only 45 played head-to-head. This is 0.8% of the registered 5,602 users who played in more than one round. Of those 45 users, 23 completed more than one head-to-head game.

The most head-to-heads undertaken in total by a single user was 4: this player engaged 2 other users twice each. The greatest diversity of head-to-head partners achieved by a single user was 3. A single pair of users could

have performed up to 7 head-to-heads (1 in each round), yet no player performed a head-to-head in more than 3 different rounds, suggesting that even those users who were using the feature and presumably seeing the points benefits felt that the barrier of being collocated with another participant was too great.

These results seem to indicate that head-to-head play had significant hurdles for users. Responses gained through the questionnaire section of the application support this. Comments on this issue focussed on two main areas, with several users stating they lacked the opportunity to perform the feature, for example "I would have used head-to-head more if more people amongst my friend using the software/feature too".

Many users also suggested alternative means to engage in head-to-heads that were not reliant on co-location, requesting for example "i think the head to head should be just random ppl going against each other and not bluethooth", or requesting "The ability to challenge global users over wifi for a head to head".

Turning attention to data gathered from the local deployment of the application reveals a very different pattern of use. All 5 of the officemates and all 3 of the friends performed at least one head-to-head, yet none of the 3 singletons used the feature. The average number of head-to-head plays performed by each local participant was 5.2. Users reported enjoying this feature in terms of adding an extra social dimension to the application. For example one user talked positively of it adding "more friendly rivalry when watching games".

5.4 Conducting Hybrid Mass Participation Trials

The primary goal in conducting this study was to examine the different opportunities for research that arise when a mass participation trial is run concurrently with a more traditional local deployment, and to weigh gains against the additional expense, in time and money, for the researchers involved.

The value of a mixed methods approach to research has been well established in fields such as sociology. Denzin notes that "By combining multiple observers, theories, methods and data sources, sociologists can hope to overcome the intrinsic bias that comes from single methods, single observer, and single theory studies." [67]

This section, drawing from experiences of several app store style deployments and this hybrid trial, presents a set of practical benefits researchers can expect from this approach. These are followed by a set of recommendations for other researchers who wish to run a mass participation ubicomp trial in a way in which these biases can be overcome.

5.4.1 Practical Considerations for Hybrid Trials

Running a hybrid trial necessarily involves more work on the part of the researcher than running either a local trial or a global trial on its own.

However some work can be done once and the benefits utilised with both groups, and some tasks are much easier to complete with one group or the other. This goes some way to making a hybrid trial economically viable in research terms.

5.4.2 Commodity Hardware and Release Software in Local Trials

The use of participants' own devices for a majority of those recruited, combined with the release of the research software via a channel users are familiar with, greatly reduces the work necessary in managing hardware and software deployment.

Even locally-recruited participants might be using their own devices, so an email with the appropriate link or a few minutes with the participant and an Internet connection are enough to kit out a participant for the trial.

In general, software released in this fashion must also be more polished and stable than more standard research prototypes [61], which obviously incurs greater expense in terms of implementation and testing. However, this, in turn, greatly reduces the amount of technical support necessary during the trial – producing less down time and more valid results. And where problems do occur the local participants are available to perform quick and accessible testing.

5.4.2.1 Numbers with ease; Interviews with ease

With the public distribution to provide validation of the generalisability of the results, there is less pressure to reach the 'magic number' of local par-

ticipants deemed necessary in the research community at that time.

Moreover this access to local participants greatly reduces the effort required to glean useful and detailed qualitative data in comparison to using remote participants.

5.4.2.2 More Interactive Design Cycle

One benefit to be gained from a hybrid trial method is in utilising the benefits to be gained from each group of users as part of the software design cycle.

As shown in Chapter 4, there is benefit in leveraging the large numbers of users in a mass participation user trial to identify the tensions within the system using a mechanism to allow all users to submit suggestions for new features for the application.

In this trial, this generated a large volume of responses, though each suggestion was only a short piece of text. These ideas could then be presented to a locally based group of users to discuss in greater depth at an Action Workshop, allowing those with a stake in the system to explore the subtle effects these modifications would cause.

Conversely, the redesign process could begin with the Action Workshop. Developers may wish to evaluate current opinion, present new ideas, or probe the users for suggestions for desired features in a more interactive and rich way than through a task system. In beginning with speaking to local users to hear current thoughts and future directions in which they would like to see the software develop a richer understanding of how these changes would affect the system could be achieved before polling the large body of global users, allowing the descriptions of the options of the poll to more accurately describe the consequences of each choice.

In this way, the benefits of both groups of users are being exploited – the opportunity for in-depth discussions, and putting possibilities out to large number of people for more certainty as to what will be popular decisions.

The concurrent use of the two user groups results in greater benefit than would have been possible with either group alone. Here the two groups are used together, but utilise the strengths of each to maximise the benefits offered by this hybrid methodology.

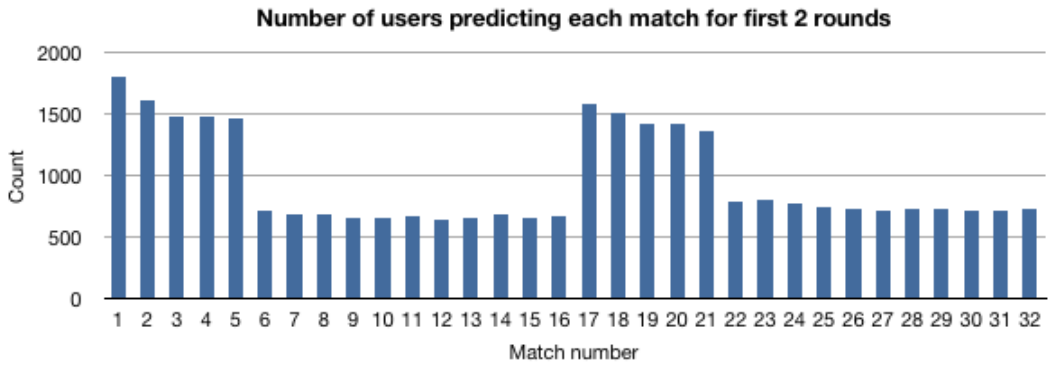

Figure 5.4: The number of users predicting each match for the first 2 rounds of the world cup.

5.4.3 Recommendations for Hybrid Trial Research

Here is discussed not only the observed differences between the results available from each style of trial, but also the areas in which they are complementary – with one providing detail or contrast to points exposed by the other.

The different ethical responsibilities researchers have towards participants in each style of user trial, and the restrictions they present, are also discussed. This discussion is formed around the following four recommendations, and examples are drawn from the hybrid trial to illustrate these points.

5.4.3.1 Use the Small to Explain the Large

As identified earlier, a fundamental difficulty in conducting mass participation trials is the lack of rich qualitative data and consequently the reduced ability to explain the reasons behind patterns observed in the vast amount of data being generated.

In running a hybrid trial, the local users also afforded us the ability to ask questions relating to patterns of use observed in the mass participation users' aggregate data – patterns for which the motivations were unclear.

For example, during the early rounds, it was observed that not all users were predicting the results of every match. The number of users predicting each match for the first 2 rounds of the world cup is shown in Figure 5.4. It appeared that a significant number of users were only predicting results of the first five matches in each round.

It was speculated that this might be because many players were just trying out the game without fully committing to predicting every match, or that they misunderstood the deadline system and did not realise that predictions for every match in the round had to be submitted before the first match commenced.

When a local user was observed to have exhibited this same behaviour, he was questioned about it. He stated that initially he had not been aware that the match prediction screen (Figure 5.1, *right*) could be scrolled down.

This solved the mystery; iOS applications hide the scroll bar by default, only displaying it when a user drags to scroll, so a significant number of users thought that there were only 5 matches to predict each round.

To fix this problem, a new version of the application was released that displayed a popup when a user uploaded predictions, informing them on how many of the available matches they could still predict.

Following this update, the pattern of the first 5 matches receiving more predictions was no longer observed. Another unexpected behaviour displayed by a large number of mass participation users was the use of the application after the World Cup had ended.

While activity dropped significantly after the event was concluded, it did not drop to zero as may be expected. The users could no longer make predictions and there was no new data being added to the application so the motivation for repeated launches in the weeks after the World Cup was not obvious.

The question was presented to the local participants during the post-trial interviews and, while most had stopped using the application, one reported that the application presented the full results of the World Cup in one place, in a way easier to access than a web page.

5.4.3.2 Use the Large to Verify the Small

One of most obvious advantages of running a mass participation trial is the much larger number of users that can be expected, as compared to a more traditional local deployment, and therefore the increased confidence with which claims can be made based on consistently observed behaviour.

As an inverse to the previous section, the hybrid mass participation trial also allowed the use of the global users to verify the generality of behaviour

observed during detailed analysis of observations and interview transcripts from local participants.

One behavioural pattern reported by several local users was the use of the application during the matches to check the predictions of the top players and compare these to the current state of play on the field.

It had been expected that users would not welcome distractions during the matches themselves, so the application had been designed to fit around the matches – users were only allowed to enter predictions up to the start of the first match in a round, and the scores within the application were not updated until each match had finished.

In order to verify that this appropriation occurred across the user population, graphs showing launches per day were generated which confirmed that the local users' behaviour did generalise to the user population at large, with usage during the periods in which matches were being played more than double the baseline.

As well as being able to use the body of mass participation users to verify small-scale findings, the same procedures can also be used to detect where results from the local group are not observed among the global group.

There is always a risk in running only a standard, local trial that the data will be skewed by the inclusion of outliers and that their behaviour becomes erroneously considered as being representative of a large proportion of the population. A further risk of local trials is the participants' susceptibility to 'experimenter effects' [225]: subtle conscious or subconscious cues a researcher might give users that affect performance.

Such an effect is less likely among globally-recruited users, where the users' contact with researchers is generally far lower. As an example of this in the trial of the hybrid method, it was found that 8 of 11 local participants used the head-to-head function of the game.

This represents an uptake of almost 73% – but the percentage of mass participation users who made use of this feature was 0.8%. Even ignoring the head-to-head aspects, the number of matches predicted in one-player mode was also far higher: 71% for local participants compared to 15% for the mass participation user group.

It is possible that both of these differences are due to experimenter effects: local users were very aware that their participation was being measured, they were paid to use the application during the trial and therefore felt

compelled to put in more effort.

Regardless of the reasons behind the discrepancy, without the mass participation element of the hybrid trial, a very different impression of the features' popularity would have been gained and therefore, it could be argued, informed very misleading results.

5.4.3.3 Maintain an ethical approach through a framework of levels of engagement

Studying the differences between mass participation trials and normal (i.e. small) scale trials highlights certain ethical concerns, and the methods used for the trial of the World Cup Predictor application were specifically designed to address these.

As mentioned previously, an important consideration when conducting this form of trial is the issue of informed consent.

When conducting traditional trials, evaluators are generally able to interact directly with participants, thus gaining the opportunity to assure themselves that truly informed consent has been obtained with regard to the trial procedure. As reported in the previous chapter 70% of users of Hungry Yoshi had not understood that they were part of a trial. If users are unwilling, or due to extraneous circumstances such as cognitive capacity or language skill, cannot read presented terms it is infeasible to expect them to understand the possible consequences of logging or academic publication.

Additional concerns centre on the minimum age requirements for giving informed consent, and the extent to which this can be verified by researchers.

Participants in any form of experiment can deceive evaluators, by intent or by misunderstanding, but such deception is easier for mass participation users and is compounded by difficulties in validation of reported facts.

These techniques for gaining consent within mass participation trials follow the standard practices used within commercial settings.

However, this raises a research question as to methodology: how can the ethical responsibility of evaluators be satisfied when there are no feasible ways to make sure that users are of an acceptable age, and capable of giving informed consent?

Furthermore, standard ethical practice involves a debriefing following a trial. Conducting such a debriefing is more difficult in a mass participa-

tion trial. Such trials are frequently without a defined end date. It cannot be reliably predicted when a user will play for the last time, and given the primary means of communication is through the application itself, it can be difficult to have any significant contact with the user after that time.

Given these ethical concerns, the mass participation users were approached differently to the local users. The mass participation user group was engaged with in a much lighter manner than would have been desirable had the goal been to utilise them as a resource to the fullest extent possible: direct interaction was limited to non-compulsory survey questions, as opposed to requesting interviews, and aggregate logged data was examined as opposed to examining in detail the data for any single user exhibiting an interesting pattern of behaviour.

This was seen as a compromise that protected users whose consent was not or could not have been 'informed.' A single user, even when actively engaging in the trial by answering survey questions or providing log data on a specific issue, would not have his/her privacy, or expectations of such, compromised.

The types of questions asked of remote users were also limited: potentially invasive or sensitive questions were avoided not only due to the problem of verification of consent but because it is much harder to converse sensitively at a distance, i.e. it is harder to read a participant's reaction to a subject matter and stop if necessary.

While researchers should be encouraged to continue to explore novel ways to meet their ethical responsibilities, using the full range of engagement possibilities with participants in this manner, researchers can feel confident that they are pursuing an ethically sound research path.

5.4.3.4 Do not rely on the emergence of specific social structures in your participant base

Certain research questions are predicated upon social use of an application by users with a certain topology of relationships. If it is important that the system is used among groups of users with this social structure pre-dating the use of the application the study should include a local trial, where users can be selected to match the required social graph.

While a global trial could certainly be interesting in seeing how often or how rarely such social features are used, the results suggest that social groups

are not guaranteed to all adopt an application, even when strong incentives are given for using social features.

As shown, the results from the World Cup Predictor trial indicate that the ad hoc network functionality was used to a reasonable degree within the handpicked social group of people who had regular contact with each other: all of the locally recruited participants who had existing social ties using the head-to-head feature.

This contrasts with 0.8% of the global user base. As suggested above, experimental bias could be a factor here, with local participants feeling more of a duty to use the application. However the 3 local users who were not part of a social group also did not participate in head-to-head play and qualitative data gathered during the experiment confirms that many users could not find suitable partners to play with.

More research is needed to verify whether, when in the right social context, users are more likely to take advantage of the head-to-head mode, yet it seems that handpicking participants matching a desired social topography is far more likely to lead to usage of features designed for such a group than relying on users acquired 'by chance' in a global release.

5.5 Conclusions

In this chapter, an iOS application to test the hybrid mass participation trial methodology of combining the use of a large-scale deployment with a local trial, was studied with the release of the World Cup Predictor application.

Based on these experiences of the hybrid mass participation method, it can be seen that it offers a useful means to alleviate the weaknesses of both local and global trials, as identified in the release of Hungry Yoshi discussed in chapter 4.

This study has also shown that regular usage of ad hoc networking in the World Cup Predictor application only took place among those with pre-existing social ties and regular co-location.

Despite offering generous prizes and designing a scoring mechanism to encourage peer-to-peer usage, there was very little uptake of this feature among the global user base.

The serendipitous use of such networking technologies should not be expected within current applications. This also reiterates the recommenda-

tion from Chapter 3 that research questions relying upon any pre-existing social topology among participants should be carried out with a selected local group.

Although this kind of hybrid methodology may have applicability in other application areas, it is particularly worth exploring in future ubicomp research. Local context, which is vital to ubicomp yet clearly variable as one looks worldwide, can be studied with greater assurance as to what can be generalised, and what is specifically local, thus helping to address a key design issue for the field.

The benefits for researchers who running a mass participation-style trial when using the hybrid approach can be summarised as follows:

- Reduce the difficulty of gathering qualitative data while improving its quality.
- Allow for more solid ethical practice to be maintained.
- Allow for the explanation of patterns emerging from analysis of data through interviews and local observation.

When compared with conducting trial with only local participants, using this approach will:

- Allow findings based on consistently observed behaviour to be reported with great confidence.
- Mitigate problems resulting from including outliers in a small number of participants.
- Mitigate bias resulting from experimenter effects.

In relation to RQ1 it can be seen from the summary above that the complimentary effects of simultaneous remote and local trials provide benefits in both directions, and that these benefits are available to researchers with little additional cost in time or effort.

An important component of the hybrid method is a means for managing the ethical responsibility of researchers in conducting large-scale trials, providing a balance between utility and ethical practice: a focus on keeping interactions with the remote participants lightweight and giving individuals more privacy than a local trial participant could reasonably expect.

This begins to answer RQ2, however there are many more questions that remain to be answered, for example assessing whether information available to participants after the trial has ended, e.g. on a web site, may serve to acceptably 'debrief.' These and many other aspects of the ethical problems faced when running mass participation trials are discussed in detail in the next chapter.

Chapter 6

Ethical Responsibility

As has been shown, large scale trials not only offer huge opportunities for the community, but also world-scale challenges of validity and ethics.

In the previous chapter there was documentation of a first step towards a more grounded ethical practice in this area, but there is a need for a more detailed and explicit investigation of the ethical responsibilities of researchers. This chapter offers such an investigation.

Being involved in the organisation of the Ethics, Lies and Videotape workshop [51] at ACM CHI 2011 gave an interesting and in depth insight into the concerns and practices of other researchers in this and adjacent fields – including the appropriation of user generated content for research analysis and online ethnographies where the full process is archived and public on a web forum.

The practices of the researchers who participated in the workshop were heavily influenced by the institutions for which they worked, with different levels of ethical review placed upon them, from a complete lack of guidance or oversight to HCI user trials being subjected to the same checks and balances as clinical trials.

However, even those researchers working in institutions with the lightest touch of ethical guidelines saw the need for a community-wide discussion of best practice, citing a concern that if the community did not rise to the challenge and provide guidelines for researchers they may, as a backlash to some real or perceived transgression, be imposed from above by a body with much less understanding of the domain.

At the same time, new concerns are arising among the general public. There has been a recent backlash against mobile applications' logging of data ir-

relevant to the functionality of the application, with applications such as TaintDroid [78] displaying the information transmitted by other Android applications. There have also been negative reactions to the Facebook iPhone application update giving the impression that it was sharing phone numbers[1], and researchers at the University of Bath covertly tracking the Bluetooth devices of thousands of people—and then publicly releasing the software so that it has been deployed in more than 1,000 locations worldwide[2]. Researchers have a responsibility to the community not to 'poison the well' by fuelling such mistrust.

HCI is by no means the only field of research which uses human trials as a method of evaluating hypotheses and exploring ideas. Notably, the fields of Psychology and Medicine have well established guidelines compiled and upheld by professional bodies. While it may be argued that the potential harm of an ill-run medical trial [108] or revisiting the Stanford Prison Experiment [23] is much greater than that posed by a piece of mobile or online research, the increasing value of personal data, the difficulties in anonymisation and the volume of data available pose a problem.

In 2000, Sweeney showed that 87 percent of all Americans could be uniquely identified using only three pieces of information: their postal code, birthdate, and sex. In doing so she was able to take the 'anonymised' data released by the Massachusetts Group Insurance Commission on all their state employees and, when combined with the 'anonymised' voter rolls from the city of Cambridge, Massachusetts which were purchased for $20, identify the current state governor's health records including his diagnoses and prescriptions [233]. She then sent them to the governor's office.

This work, and others like it, show that almost all information can be defined as 'personal' when combined with enough other, relevant data.

Researchers at AOL and Netflix made large amounts of data available and despite removing any obvious personally identifiable aspects of the data, such as IP addresses and user names, others were quickly able to reassociate large proportions of the data with end users by combining these data sets with other, publicly available, data.

The New York Times [24] identified, and interviewed, user number 4417749 simply from the information contained within her historical search terms. In the case of Netflix, researchers at The University of Texas, Austin were

[1] www.guardian.co.uk/technology/blog/2010/oct/06/facebook-privacy-phone-numbers-upload

[2] www.guardian.co.uk/uk/2008/jul/21/civilliberties.privacy

able to cross reference the anonymised information with reviews on the public Internet Movie Database website[3] and re-identify that data set [177]. The same researchers were also able to de-anonymise the Twitter accounts of a third of users who also had a Flickr account, with only a 12% error rate [178].

Not all re-identification experiments have been carried out on such seemingly inconsequential data. Students at MIT cross-referenced the Chicago Homicide Database with the Social Security Death Index to re-identify the victims of homicides, and identify their families [182]. The homicide database included data that could cause real, identifiable harm to the surviving families of the deceased including flags indicating whether the murder involved drugs, child abuse, gang violence, or domestic abuse as well as previous criminal history of the victim.

With the continual advancement of re-identification techniques, what information can, and can't, be said to be anonymous or insignificant continually changes.

> For almost every person on earth, there is at least one fact about them stored in a computer database that an adversary could use to blackmail, discriminate against, harass, or steal the identity of him or her. I mean more than mere embarrassment or inconvenience; I mean legally cognizable harm. Perhaps it is a fact about past conduct, health, or family shame. For almost every one of us, then, we can assume a hypothetical 'database of ruin,' the one containing this fact but until now splintered across dozens of databases on computers around the world, and thus disconnected from our identity. Reidentification has formed the database of ruin and given access to it to our worst enemies.
>
> (*Ohm, P.*, 2010 [185])

Ohm paints a bleak picture of the state of current data privacy provisions, and points out that attempts to anonymise data to the point where re-identification would be impossible would render the data useless: "data can either be useful or perfectly anonymous but never both." [185]

Against this backdrop the aggregation of vast stores of usage data by researchers, including such things as participants' location (via. GPS), the times they use the software, the people around them (via. Bluetooth) and

[3] www.imdb.com

even simply if they have cellular connection or not at a particular time could all, under certain circumstances and if re-identified, cause 'legal cognizable harm' to participants. The timestamped locations could, for example, lead to a user being charged with a motoring offence. The use data could identify that a worker was using a mobile device when prohibited, such as a worker in a casino or a member of a jury. Having a Bluetooth scan indicating that one was in the proximity of a known criminal at a certain time may result in being seen to be guilty by association. Having the mistruth of 'I couldn't get a signal' exposed could cause embarrassment and distress to both the participant and the other party.

Researchers in HCI have the same possibility of causing harm, and therefore the same responsibilities, as any other researchers for whom humans are the subjects of their investigation.

Researchers are rewarded socially and economically for producing valid and valued research. These rewards can create apparent conflicts of interest between the needs of the researchers and the research participants. A careful balance of the scientific, legal and ethical considerations must be sought.

In this chapter, the current guidelines to researchers in the community for human trials are examined, compared against other fields which conduct human trials, and their application in remote trials discussed.

6.1 Current Guidelines for Researchers

Perhaps the best-known guidelines specific to mobile and ubiquitous computing are those in Greenfield's Everyware book [103]. High-level guidelines such as 'do no harm' and 'default to harmlessness' were discussed, and are still generally applicable, but have yet to be contextualised to suit new ubicomp research practices. New technologies support not only new research practices that challenge the old, but also new user practices. The widespread use of web sites such as YouTube and Facebook, and the near-ubiquity of cameras on phones make some established guidelines, e.g. in MacKay's CHI '95 Ethics, Lies and Videotape paper [157] seem rather quaint. People are increasingly accustomed to the dissolution of traditional social barriers of privacy driven by the traditionally poor privacy controls provided by such online social networking sites [229].

However this shift in user attitude cannot be expected to be consistent across, or even within, cultures and demographics and the possible harm

to the reputation of the researcher, or research as a whole, by overestimating this movement in opinion outweighs the benefits of taking a too relaxed attitude to the ethical diligence required by researchers.

Both the British Psychological Society (BPS[4]) [251] and the American Psycological Association (APA[5]) [215] provide guidelines for researchers conducting human trials, with the BPS providing a supplementary publication dealing with the specifics of conducting research over the Internet [250].

Increasingly, human research crosses institutional, professional and national boundaries, bringing further complication into the application of appropriate ethics protocols and review processes.For these reasons, the development of detailed and specific regulations on the handling of ethics issues in human research by researchers, with the aim of covering all eventualities, is seen by many ethicists as an ultimately flawed direction of travel. As soon as one new set of regulations is finalised, a new method or topic of research is likely to emerge that is not covered. The existence of lengthy, detailed and prescriptive professional or institutional regulations raises the risk of researchers following the letter, but not the spirit, of the regulations and may in consequence lead to research being carried out that is ethically flawed.

Both the BPS and the APA give principles based guidance to researchers.

6.2 Overview of BPS and APA Guidelines

The BPS gives four guiding principles specifically for researchers using human subjects. The APA builds its ethical guidance for research directly from the six general principles given to cover all aspects of their professional conduct. The titles of these are given in the following table, ordered to show their similarity. While items in one column do not necessarily map directly to their counterpart in the other, there is a marked overlap which is discussed below. The principles in the first row of table 6.1 are described by the BPS and the APA as dealing with the 'dignity and worth of all' [251, 215]. Beyond that they both discuss the fundamental rights that any participant has to self-determination & autonomy, personal liberty, privacy and confidentiality. Ensuring the autonomy of the participants by making sure that they understand the consequences of participation, and that it is optional,

[4]www.bps.org.uk
[5]www.apa.org

BPS	APA
Respect for the Autonomy and Dignity of Persons.	Respect for People's Rights and Dignity.
Social Responsibility.	Social Responsibility.
Maximising Benefit and Minimising Harm.	Concern for Others' Welfare.
Scientific Value	Professional and Scientific Responsibility.
	Integrity.
	Competence.

Table 6.1: The Principles of Ethical Research from the BPS and APA.

is covered in §6.4. As is ensuring that each participant's liberty is protected by giving them the right to withdraw from a trial at any time, and the knowledge and tools to do so.

The challenges of ensuring that the privacy of the research subjects is respected and the confidentiality of data and results are discussed in §6.3.

The second row shows that both the BPS and the APA feel that Social Responsibility is central in deciding if a course of action is ethical. The researcher must be mindful of, and responsible to, the societies in which they live and work. The BPS principle in this regard points out that the responsibility of the researcher goes beyond any direct harm that may be caused and that 'alert to the possible consequences of unexpected as well as predicted outcomes of their work' – such as the dangers of re-identification highlighted above and covered in §6.3.

The principles of 'Maximising Benefit and Minimising Harm' & 'Concern for Others' Welfare' both deal with the theme of risk. It is the researchers' responsibility when ensuring ethical conduct to, as a general rule, not expose the research subject to any risks greater than they would encounter in their everyday lives. That should be aware of the real and perceived power differences in their relationship with the participants and careful not exploit their research subjects. In research which poses risks to the participants' psychological well-being, mental health, personal values, or dignity, these risks should be assessed to determine their probability and severity, and measures put in place to minimise the exposure of the participants and recover should the worst-case scenario be realised.

The principle of Scientific Value put forward by the BPS incorporates many aspects of the principles of Professional and Scientific Responsibility, Integrity and Competence put forward by the APA. The scientific value of the research must be clear and appropriate (Professional and Scientific Responsibility), the research must be well designed and conducted in a way that ensures its quality (Competence) and integrity (Integrity). The APA princi-

ples add, in regards to research, that the practitioner has a responsibility to intervene with colleagues to prevent or avoid unethical conduct.

The BPS guidelines for Internet Mediated Research [250] identifies two axes along which the ethical problems faced when using remote participants; the identifiability of participants and if they were directly recruited to, and understand they are part of, the research. The identified problematic areas regarding this are in ensuring the autonomy of the participants when they do not read the briefing presented to them, matching the participants' understanding of how public or private their online activities are, and all aspects of the ethical responsibilities placed on a researcher that happen after the participant has completed or withdrawn from the research.

The mediating methods discussed are, in most cases, specific to the research cases that were examined of observation of online forums and other communication methods and of running online surveys. In this way many of the recommendations are not directly applicable to the methods of research discussed here, such as using Javascript to detect a participant closing a window during a survey and presenting them with a modal debriefing dialog, but the problems identified and the perceived severity of their ethical risk are used to inform the following sections.

6.3 Data Collection & Control

Researchers, by storing any personal data related to their participants, are already legally bound in many countries by legislation such as the UK's 1998 Data Protection Act [13] and the EU directive on the Protection of Personal Information [69] which is soon to be augmented by the EU regulation on the processing and free movement of personal data [16]. As shown on Forrester's Privacy and Data Protection by Country Heatmap[6], the majority of the world's population is covered by some form of data protection laws and any researcher must ensure that they comply with the laws where they live and work, here the ethical obligations as derived from the experience of releasing multiple applications and knowledge of the available guidelines are discussed.

As outlined above, any collection of identifiable or re-identifiable data could cause potential harm and therefore must be stored and transferred securely. While it cannot be expected that each research group, or individual

[6] heatmap.forrestertools.com

researcher, be at the forefront of data security and encryption, they have an ethical duty to keep abreast of the current industry standards in this area. They must also take steps to mitigate risks that are identified in the planning and in the course of their research, including the risks resulting from them not being security professionals, by instigating suitable procedures with regards to their handling of the data.

All trial data on the participant's device should be encrypted and deleted after being successfully transferred back to the researcher or when the participant withdraws from the trial, either by deleting the software or withdrawing consent from within the application. This mitigates the risk of personal historical data being made available to persons of ill intent via physical access to the device or rogue processes running on said device.

The externally visible server which receives the uploaded data from the client software provides a possible point of failure which could affect a large number, if not all, of the trial participants. Beyond ensuring that the operating system and web-server software is regularly updated, ensuring that passwords are of an appropriate complexity, and ensuring that access permissions are correctly set the responsible researcher should recognise that the risk of a security breach is still there and minimise the damage it would cause. One way to do this would be to avoid keeping the database of historical log data on such an externally visible server – by regularly moving the incoming user data to another, fully fire-walled or offline machine the potential amount of data compromised by a breach is greatly reduced.

Sharing raw or anonymised data between researchers and institutions is standard practice in many fields – ensuring that results are reproducible and that the greatest amount of knowledge can be extracted from the effort expended in gathering such corpora. Yet as outlined at the beginning of this chapter successful anonymity of data is increasingly difficult to obtain.

The recently passed EU directive on the movement of personal data [16] introduces a principle that could result in more ethical practice when data is shared between researchers. Here an entire chapter is devoted to the 'Rights of the Data Subject' obliging *data controllers*, those collecting and processing the data, to provide transparent, easily accessible and understandable information on what has been collected and that they provide procedures and defined deadlines for requests for access, and deletion, of personal data. Of interest with regards to the sharing of data between researchers is the chapter that deals with the transfer, or onward transfer, of

data outside of the EU or to an international company. It states that the collector of the data is responsible for ensuring that it is not transferred to another party with less stringent security protocols – adapting this as a guideline for researchers would mean that the risk of data exposure through a security breach should be *at least* as low in the second institution as in the first, and that the original researcher is responsible for ensuring that this is the case before transferring the data.

Before any data is transferred out-with the control of the data collector, it should be subjected to a Privacy-Preserving Data Publishing technique, a survey of the state of the art of such methods can be read in [91], which transforms the data by replacing any explicit or quasi-identifiers in the original with new identifiers that hide some detailed information so that several records become indistinguishable in this respect. This is, necessarily, reducing the fidelity of the data transferred so the researcher must take into account their trust in both the integrity and the security practices of the receiving researcher when deciding to what extent to employ these techniques. They should also discuss with the researcher requesting data its expected use and remove fields which are not directly relevant.

If users declare they no longer wish to be part of the study, standard practice dictates that researchers would delete all data collected on them. However, information that has been used within an application or community, configurations or forum posts, or information that has been combined into the products of other users, such as mash-ups or derived configurations raise significant problems. Beyond the purely practical challenges in deleting this data, the seemed ethical commitment to purge all data from one participant could be seen to cause harm to another.

A related dilemma relates to trial data that may well be copied, commented on and published by participants, without researchers' knowledge, e.g. on YouTube, on their own blogs, and on Facebook. The level of responsibility researchers have for such self-published information, and the validity of collecting it for analysis must be examined.

In both these cases, the ethical principles involved are those concerned with the participants' expectations of privacy and risk. Where the trial and the application have been correctly designed, there should be little ambiguity as to what, and with whom, the participant is explicitly sharing content. Where it can be reasonably expected that the sharing would be subject to few limits, uploading a video to YouTube or sharing a configuration in an

in-app repository for example, then the researcher should not feel under obligation to take action that would harm derivative content if a request for withdrawal is received and can use such postings in their research. Where the participant has an expectation of privacy within the space that the content is shared, be that a Facebook post assumed to be limited to friends or a post in an online forum expected to be read by community members only, then the researcher must respect that expectation and work within its limits.

All this assumes that the participant has been informed, and understood, that their data is being collected for analysis. The problems with ensuring this are covered in the next section.

6.4 Informed Consent & Terms and Conditions

An important part of maintaining ethical standards in the running of a human trial is the nature of the agreement, the mutual understanding that researchers are able to make between them and the users of a downloaded application or online service. The standard procedure of presenting a briefing page of terms and conditions (T&Cs), and asking for confirmation of understanding and acceptance before allowing use, has been seen to fail to produce truly informed consent. The percentage of people who read T&C pages on installation of desktop software was reported by FAST as being only 28%. Only 30% of respondents to a survey in Hungry Yoshi indicated that they had understood it was a university trial; of those interviewed directly none had read the T&Cs and the log data shows that of the 75,818 people who agreed to the T&Cs only 1,645 opened the full document, and of them not one spent longer than 60 seconds reading the 842 word document.

Determining that a participant has the legal capacity to enter into the contract set out by the researcher, at least in the UK, means determining that they are “An adult of sound mind” [144] and although entering into contracts with minors, those with mental health problems or the intoxicated is not illegal in itself there are important distinctions between such contracts and those entered into in the traditional manner. A contract made with a minor is, barring a few special circumstances, binding only to the adult or corporation – the minor may not face sanctions for failing to uphold their responsibilities according to the contract. For those with court recognised mental incapacity resulting in the state taking control of their

assets, all contracts are deemed ineffective. In the case of any lesser mental incapacity, such as intoxication or a cognitive disability less severe than those requiring state intervention, then the contract *is* binding unless they can establish that they didn't understand what they were entering into *and* that the other party knew this to be the case. While this gives some comfort to researchers about the legality of the T&Cs of their applications while their participants are in the UK, or in the European Union provided that the contract states that any disputes would be adjudicated by a UK court [11], when extended to an international distribution this becomes more complex.

Not only is the question of the legality of the contract raised, but also the question of whether the state in which the participant resides feels that it has jurisdiction over those releasing the software. In the USA there is the notion that the software developer must 'purposefully avail' the individual state in which the plaintiff resides [249], meaning that the software or website must *actively interact* with people in that state, precluding the prosecution of a website owner simply presenting information but opening the door to prosecute an e-commerce site shipping to that state for example, or *target* that state by providing a toll-free number accessible by residents. Australian courts, however, have taken action against foreign companies for providing passive sites accessible from Australia [248].

Given this complex, and rapidly evolving, topology of laws, regulations and precedents, simply following the lead of online businesses by providing opt-in contracts with end users that state the jurisdiction in which it is to be considered may not be enough to ensure that the researcher is acting legally, and it is certainly not enough to ensure that any research is being carried out in an ethical manner.

6.4.1 Capacity

Simply determining that a participant meets the age requirements of a given jurisdiction or, for a given university ethics board, when they are physically present, is by no means foolproof – a participant may make a false statement regarding his or her age for a variety of reasons, the most obvious of these being the reward offered for participation by the researcher – but replacing the researchers' judgements on physical appearance, personality and validity of identification with the typical online solutions of 'confirm your age' check boxes, requesting the user to enter a date of birth or providing the details of a valid credit card leads to a marked degradation in the

confidence that can be put in the understanding that any given participant in a remote trial is of age, and of what the age of majority is where the participant resides.

Determining the capacity of a prospective participant in mental health terms is even more difficult in a remote setting.

Researchers must then take care not to intentionally or unintentionally target vulnerable groups when advertising their trial. The recruitment of participants for remote mobile trials can take many forms, from physical fliers to demographically targeted online advertisements, but the most basic form is the icon, description, keywords and chosen categories used in the online repository. Where the store allows, the researcher should restrict the application to those over the age their institution has deemed acceptable for participation and, if the trial would not be adversely affected, it would be advisable to raise this limit to an internationally recognised age of majority such as 16, 18 or 21.

In combination with the age restrictions, which cannot be expected to be 100% effective, researchers should study the icon sets and the language used in other entries in the store which they have reason to believe to be more popular with an older audience and model their submissions upon them so as to not unintentionally target children.

6.4.2 Understanding

In the context of a traditional trial, researchers can determine the participant's level of understanding and dynamically adjust the amount of information they present verbally to ensure that the most important points are clearly understood. If they are unable to bring the participant to a suitable level of understanding of the consequences of participating in the trial, they have a number of options available to them, from letting the participant complete the trial but deleting the data to halting the process altogether and finding an alternative subject.

Briefings at a distance over the Internet remove this interpersonal communication, the subtle clues and queues giving the researcher extra information on the participants level of understanding and the chance to reiterate and reword as needed, exacerbating a problem that does exist in traditional trials: that it may be impossible to verify any given user *understands* the T&Cs of participation to a sufficient extent to give informed consent. If a

user can not be seen to be giving informed consent the extent to which the researcher can ethically collect data and publish the results of its analysis must be examined.

After any human trial, the debriefing stage is needed to ensure that the understanding negotiated before the trial stands up to the reality of participation. This debriefing stage is critical in ensuring ethical practice. It allows the participants the opportunity to ask for clarification on aspects of the trial or data collected they were unable to fully grasp before participation and it allows the researcher to gauge the effect that participation has, and spot any areas in the pre-trial briefing that needs to be clarified. Without an effective debriefing stage, it is also effectively impossible to ethically perform a human trial involving deception of the participants. While any human trial with remote participants causes problems with debriefing, this is even more difficult when the communications with the participant are primarily through an application under study.

The suggested solution in the BPS's guidelines for internet mediated research is applicable only to browser based studies, namely of using browser scripting to detect the user navigating from the survey page before the debrief has occurred and presenting it in a modal dialog. However, detecting the end of participation beyond this is an incredibly difficult task. On the four most popular smartphone operating systems, iOS, Android, Blackberry and WindowsPhone, the developer is not given the opportunity to interact with the user when an application is removed from the device in the same way they would on the desktop by providing a custom uninstall program. They may have effectively ceased participation long before they choose to remove the application from the device, and after the analysis or publication of results based on their data has occurred, rendering any debriefing in such an uninstall program ineffective.

Without an effective way to perform a credible debrief of the participants even more weight must be put on the initial T&C briefing, as this must engender full understanding of the trial process, or an alternative method must be pursued.

6.4.3 Proportional Engagement

> The principle of proportionality should apply, such that the procedures for consent are proportional to the nature of participation and the risks involved. (BPS Ethical Guidelines [251, p. 15])

This principle of proportionality, combined with the the researcher's understanding of their participant group's expectation of data privacy in the area of study, allows a gradient of interaction to be drawn.

At the highest level, the researcher should limit their data gathering activities to that which the user could reasonably expect the developer of an application to collect: Anonymous use data and items publicly shared from within the application. For each additional piece of information requested from the participant, such as demographics or location data, a separate request for access should be presented to the user.

These should not take the form of lengthy T&Cs documents, as the purpose of these it not to ensure legal indemnity but to ensure the understanding of the participant. Such requests should be short, covering only what will be collected at this stage, e.g. GPS location, and include how it will be used, e.g. to show use differences between rural and urban areas. They should provide a link to further information, to the full details of the trial that should be available within the application, and indicate that this includes information on how to have their data removed.

Beyond the initial request for access, researchers should also show a summary of the information they have collected on an individual participant after they have provided a useful amount of data. When this should be shown is determined by the application or service being scrutinised and the use average across the participating users. When this point has been reached it can be seen that this participant's understanding should be checked, the summary should take a form relevant to the research being conducted.

Continuing the previous example, it could state the number of different locations of use logged, and the proportion of them rural versus urban. In cases where the research question is not as defined the researcher could show a clustered map of recorded locations or, using reverse geo-coding, a list of the most frequented street addresses. In all cases, the summary should be meaningful to the participant, precluding showing them a list of GPS coordinates, and they should be again given the option to withdraw.

Such a procedure can be seen to take on the role of the debrief in a traditional human trial, giving them a greater understanding of what data has been collected contextualised by their participation and allowing them to withdraw.

In keeping with the principle of rewarding research participants for their cooperation, researchers can restrict some functionality to those who have agreed to participate, restricting location based services to those willing to share their location or restricting social networking functionality to those willing to share demographic information for example.

If this approach is taken then the socio-cultural pressure applied should be monitored and, in the event that the pressure of a peer group using a service or application is suspected of impacting a participant's agency, access should be made available without collecting and analysing the data.

6.5 Applied Ethical Framework

In order to provide a concrete example of the application of these ethical principles with regards to the systems presented in this thesis, each one will be examined and, with the benefit hindsight, the application of these ethical principles described.

Common to all the applications described in this thesis is the lack of an effective method for an individual user to review the data that has been collected on them and opt out of future logging, pause logging for a certain period of time or delete their data from within the application. In hindsight this is an important component of any system that logs user data, and as such should be part of the SGLog framework or any equivalent logging system for research.

Another question raised for all but one of the applications discussed regards the applications being seen as an encouragement to jailbreak devices. As discussed in Chapter 3 taking into account the consequences and methods of jailbreaking a device, the relatively simple procedure to restore the device to its default state, and the options available to alleviate this perceived problem it is unnecessary to take action in this regard.

For Zoo Escape and Packer discussed in Chapter 3 the purpose of these releases was to gather aggregate demographic and location data. No individual was examined in detail and the location data was used at a country-

by-country fidelity. The demographic information was voluntarily submitted and access to location data for any application must be approved by the user at the OS level the first time it is requested. The information collected was apparent to the users in question and analysed in aggregate to preserve the privacy of individuals. Beyond the identified need for a method to allow users to review the data collected about them, the location data should have been reported back to the server at the fidelity that was needed. Instead, to ease implementation, the service that translates a single GPS coordinate to its corresponding country runs on the game server. This exposes the users to the unnecessary risk of this detailed location data, stored with reference to the users' unique device identifiers, being intercepted or stolen.

The release of Hungry Yoshi described in Chapter 4 in comparison to the release of Zoo Escape and Packer raised many more ethical questions. First is the nature of the game itself, as opposed to a memory game designed for quick play Hungry Yoshi aggregated the points achieved each session. This 'grinding' style of gameplay, where time spent interacting with the game very closely reflects the rewards the player receives, raises the risk that a user will fixate upon the application and play more than is healthy [142]. While games such as World of Warcraft remind users to take regular breaks as a result of identifying this risk, the relatively few rewards offered in Hungry Yoshi – points and a score table – combined with the enforced mobility aspect – a single location will gradually become saturated forcing the user to find new access points – means that while the risk is moderately sever to the individual involved, the probability of a user becoming fixated upon such a simple game is very low.

During early release of Hungry Yoshi there was one user who seemed to be playing the game much more than was healthy, with game sessions at all hours of the day. Even on reflection the probability of this happening was so low that building automatic safeguards into the system, such as counting the number of minutes played per day and warning a user if they exceeded a certain limit, would have been unnecessary.

The game board of Hungry Yoshi was built from examining the security settings of the wifi access points the device was able to detect. The more balanced a location, the easier it is to play in that location. This gives rise to the risk of users disabling the security on access points they control in order to gain points more easily. While the fact that each digital location mapped to a wifi access point was not explicit within the game, it was explained in the help and terms and conditions screens. The risk to a user of changing

their home or office wireless access point to have no security is relatively high. While it could be argued that the game could encourage users to add security to their network for the same reasons, reducing their level of risk, the number of secured access points seen was 1,253,734 compared to 307,081 open access points. One user reported changing his network security from being password based to being based upon checking the network address of the device connecting – effectively fooling the game into seeing the access point as one without security – less technically proficient users may be tempted to simply turn off security.

On reflection this is a foreseeable and relatively severe risk to the user, with a probability which rises as the user becomes more engaged with the application. Recording the initial state of an access point and not reflecting changes to security settings in the game would be one solution, another would be to allow a user confirm their ownership of a access points and give them the option within the game to change how their access points are reflected on the game board without changing their security settings. In adding this functionality users could be actively encouraged to add security to their access points, lowering their overall level of risk.

With regards to the levels of engagement with the trial there can be seen to have been three levels that users could have moved through. The first being used for aggregate data, the second being those responding to tasks and the last those who became interview subjects. For those who engaged in the task mechanism and provided textual data to be analysed there was an extra level of information about the trial provided – namely a task which asked the user if they were aware that in using the application he or she was part of a university user trial. In hindsight this task should have been the only one available to any user until they had completed it, and it should have provided more information on how and where any responses they provided would be used. As each task was analysed, a summary of this analysis should have been made available to users in place of that task, possibly giving them the option to withdraw their information on a case-by-case basis and giving them a greater understanding of how the data would be used.

In selecting users to be interviewed interesting users were identified, their their data examined closely to determine if they were a valid interview subject, then the user was asked if they would be willing to take part in the interview process. In hindsight the more ethical course of action would be to poll the whole user base to determine those who would be willing, in principle, to take part in an interview and only examine the data of these

users in any detail.

The release of the World Cup Predictor application described in Chapter 5 raised far fewer ethical questions than Hungry Yoshi. All the data from the participants recruited via the app store was intended to be used in an aggregate manner, reducing the risks associated with storing large volumes of identifiable data and the privacy concerns when looking at a user in detail.

The time limited nature of the application and the limited number of interactions possible per round meant that the risk of a user fixating upon the game, identified as a possible risk in Hungry Yoshi, was negligible in this case.

In hand picking a social group of local participants to encourage the P2P aspect of the game, and further encouraging it for all players with bonus points, the local participant group were given a better chance at winning at winning the cash prizes. During the trial none of the local players scored well enough to win a prize, however in order not to miss represent the chance that a participant recruited via the app store as to win the prizes the local participants should not have been competing for the same prizes as those in the global group.

In conclusion, there would have been a number of changes to the systems presented in this thesis had the applications been developed and the trials been run after the compilation of the ethical guidelines presented in this chapter. Most fundamental would be the inclusion of tools allowing users to review the data collected about them within the logging framework, potentially changing their attitude towards the application and causing them to revoke access to data or stop using the application all together. The greatest risk posed to users was from the possible encouragement to disable wifi security, however the solution to this problem could increase users' engagement with the game as the ability to 'own' a Yoshi or Plantation was a feature requested by users.

6.6 Conclusion

The time is ripe for reconsideration of established research norms and practices, and researchers' understanding of public practices and sensitivities, so as to strike a new balance between invasiveness and utility. There are

many ethical challenges being faced by researchers in many fields involving human trials as a result of the fast pace of technological advancement and incorporation into our everyday lives. With these challenges comes a number of exciting opportunities to use these new technologies to inform not only the design of the novel, but the understanding of the mundane. Understanding how we researchers can use this technology in ways which allow us to answer new and old questions with new levels of validity without harming the moral integrity of the community can help inform, direct and reassure research for years to come.

New guidelines must be found to discharge our ethical responsibilities as researchers in this regard. The variation in ethical clearance procedures is also noteworthy; for example, in various European countries/institutes there are no formal approvals for HCI research studies, while other countries often have quite strong constraints and official procedures to follow for any trial with participants. A comprehensive survey of these differing guidelines across the HCI community would give those defining a community wide set of guidelines an understanding of the change in practice they could cause.

The discussion of the issues of informing participants and the control of collected data directly answer the first part of RQ2, identifying the ethical issues that arise in the mass participation approach. The following summary of the ethical guidelines discussed above provides an answer to the second part of the research question posed in §1.1.

Use Proportional Engagement

While a T&Cs screen may be legally necessary, they are ineffective ethically. Use bite-sized, contextualised snippets to increase the chance of comprehension.

Gradually increase the amount of data collected on each participant and provide an inline briefing after sustained engagement.

Take Responsibility in Collection

- Don't store excess data on the user device.
- Don't store or transfer unencrypted user data.
- Don't archive data for analysis on a web server.

Take Responsibility in Data Sharing

- Ensure that any research group or institution receiving data you have collected has security measures at least as good as your own.
- Discuss research collaborators' needs before transfer. Don't include unnecessary fields and anonymise as much as is appropriate for the proposed research.

The discussion of the ethical responsibilities of researchers conducting human trials with remote participants should continue and involve research practitioners from across HCI and beyond. However, if each of these guidelines is considered by researchers working in this area in the planning stages of a trial then the guidelines could provide a valuable stop-gap solution while a community sourced code of practice is under discussion. With a baseline to start from the overall confidence in the ethical practice of the community should be seen to increase.

Chapter 7

Conclusion

The rapid adoption of mobile technology has brought with it transformations in social and cultural practice; the expectations of what computers are, what they can do, and the role of digital objects in everyday life have evolved considerably [72], leading many researchers in ubiquitous computing to voice the opinion that the mobile phone should be the primary platform for deploying new information services [217, 190].

Using the new App Store style software distribution methods to reduce the cost, in terms of hardware, of recruiting a large group of participants for a trial 'in the wild' while increasing the potential diversity of users is becoming an attractive option for researchers pursuing this vision.

This thesis examines the procedures for running large scale trials in such a way as to keep the qualitative detail necessary to inform design while gaining the diversity of users for claims of generalisability. More generally, it discusses the results that can be expected from this 'mass participation' approach, and the ethical responsibilities they place upon researchers.

In order to do this, the releases of three different systems with a total user base of over 135,000 users were described, the related literature examined and the result of submissions to and discussions at the ACM CHI 2011 workshop the ethics of large scale trials, co-organised by the author, fed into an analysis of ethical considerations.

A summary of the thesis is given in the next section, followed by a reiteration of the research questions and a summary of the contributions made. This chapter ends with a summary of possible future work that would complement the results already presented.

7.1 Thesis Summary

In Chapter 2, the trend of user trials in HCI spreading into more and more real-world situations of use, for the length of such trials to continue to increase and for the understood acceptable techniques for the evaluation of the effects of mobile technology on the everyday lives of end users to expand to include large scale, mass participation user trials was examined.

The chapter continued by discussing the suitability of using games as a vehicle for research, pointing out the increased engagement and the in-built reward systems available to influence participant behaviour but also the dangers of higher expectations of 'fun' and quality than developers of research applications may be used to providing.

The focus was then turned towards the methodological challenges of conducting research in this way. Starting with the history of ethnography and its move into HCI, the practicality of using an ethnographic approach in this setting was discussed as was the use of ethnomethodology and activity theory to guide the analysis of the ethnographic data collected.

Chapter 2 ended by pointing out the complementary nature of such methodologies and large scale user trials, noting the potential benefits of finding some form of synthesis or combination of these approaches.

Chapter 3 compared the different software distribution methods available for Apple iOS devices. After an initial investigation of their characteristics, strengths and weaknesses, a single application was deployed as a comparative example. The resulting analysis of this deployment and the details of each of the distribution methods allowed the formulation of guidelines to help researchers choose which approach best suited their planned deployment.

These guidelines were then applied and refined in the course of the world-wide trial of Hungry Yoshi, as described in Chapter 4. Here, the aim was to push the upper limit on the number of participants as far as possible while still combining quantitative and qualitative approaches in ways that usefully and efficiently fed into the redesign process. Comprehensive system logging, a means of interacting with users that was integrated with the user experience of Yoshi, and interaction via Facebook were some of the methods used.

The use of quantitative and qualitative assessments to assess the activity and engagement of the user-base, and using this to perform targeted in-

teraction with participants, how that interaction took place on a variety of scales, and how feedback mechanisms were embedded within the system and their use encouraged was described.

Chapter 5 addressed some of the limitations of the nascent mass participation approach outlined in Chapter 4. It was shown that a hybrid trial methodology combining a large-scale deployment with a local trial can be a powerful tool in addressing shortcomings of trials that are either solely local or solely global.

The deployment of an iOS application to test this new hybrid mass participation methodology was carried out, involving a large-scale global trial with a simultaneous small-scale local one.

An important component of the hybrid method is a means for managing the ethical responsibility of researchers in conducting large-scale trials, providing a balance between utility and ethical practice: a focus on keeping interactions with the remote participants lightweight and giving individuals more privacy than a local trial participant could reasonably expect.

The ethical challenges presented by this new hybrid approach, and related approaches, were discussed in detail in Chapter 6. Such reconsideration of established research norms and practices is needed in order to strike a new balance between invasiveness and utility.

The guidelines given in Chapter 6 aim to form a basis for discussion of the ethical responsibilities of researchers conducting mass participation trials, and if each of these guidelines are at the very least discussed and considered by researchers working in this area when planning a trial the overall confidence in the ethical practice of the community should be seen to increase.

7.2 Contributions

This thesis outlines a methodology for large scale trials that allows the research to keep the qualitative detail necessary to inform design while gaining the diversity of users for claims of generalisability while discharging their ethical responsibilities to their participants.

This was framed by two research questions, introduced at the outset of the thesis:

RQ1: Can qualitative and quantitative data be effectively collected, analysed and fed back into the redesign process from a large scale user trial?

RQ2: What ethical issues arise in the approach to answer RQ1, and what ethical guidelines should such research follow?

The approach described in Chapter 4 is a partial answer to RQ1. It showed that, in large-scale trials, relatively rich qualitative data could be collected along with substantial quantitative data. The cost of collecting and analysing all of that data was excessive, however, and there remained concerns as to the generalisability and accuracy of resultant findings.

A refined answer to RQ1 was therefore developed, in the form of the hybrid approach described in Chapter 5.

The hybrid mass participation method gives researchers access to subjects for interviews and observational studies, it provides a cohort of users with close ties to the researchers for participatory design techniques and allows for more ethical practice by moving detailed analysis of use from remote participants to locals — with whom informed consent can be obtained and verified.

Limitations of the new approach were discussed in §5.4.3, and further research is needed to refine and deepen this approach, but nevertheless we proposed that the approach offers a practical and efficient answer to RQ1.

In order for such approaches to be not only practical and efficient, but acceptable, ethical concerns have to be addressed. This issue was framed as RQ2, and we propose that the discussion of Chapter 6 — summarised in the guidelines of §6.6 — form an answer to RQ2. Again, we emphasise that these contributions should be seen as a starting point and, while following these guidelines would result in more ethical practice in the majority of cases, a wider consultation of the community must be entered into before any such set of guidelines could be proposed as universal.

7.3 Future Work

In taking forward the work presented in this thesis, there are at least three areas in which research could be carried out. These centre around the

problems of recruitment & engagement, data collection, and ethical responsibilities.

The challenges of recruiting users and keeping them engaged long enough, and engaged to a sufficient extent to be useful participants in research have been felt by almost all researchers in this area. There are two directions, by no means mutually exclusive, that could reap benefits in this regard.

The first is the use of advertising to recruit users. The effects of different methods of advertising on the rate of recruitment and the retention of those users would give important information to those planning any large scale or mass participation trial. While there is a large body of research on the effectiveness of internet advertising [21], the questions that need to be asked of advertising here go beyond how to get a certain demographic to click a banner. For example, would adverts that highlight the trial aspect of participation recruit as well as those that didn't, and what would be the effect on engagement with the application and the trial process?

Another method which could help in the recruitment and retention of users would be the fostering of a community of users willing to give their time for research, much in the same way that online communities have formed around the citizen science projects like GalaxyZoo [136], where users find and classify celestial objects, FoldIt [77], where users play a game to fold proteins, and Cerberus [239], an online game in which users survey the surface of Mars. By harnessing a large community, such as the one that Ocarina [245] encouraged, to engage in a number of mobile experiments, in the manner of Henze et al. with Tap-It [118] researchers could find quicker uptake of their applications within this user community and a more engaged and informed participant base. Increasing the engagement of the participant has been shown to produce better results in online research [218].

In order to explore this, a deliberate effort to build a community of participants by the researchers involved, by no means a trivial task, would have to be undertaken. The challenges of creating and maintaining an online community have been discussed in detail [39], yet despite these challenges they have been successfully nurtured for online learning [232], showing that the creation of a community for large scale and mass participation mobile research is within the reach of researchers.

The challenge of data collection is connected with the stated advantage of large scale trials of gaining examples of use in a wide range of contexts. The

methods for collecting qualitative data from users shown in Chapters 4 and 5 were primarily passive with respect to the researcher. In Hungry Yoshi there was the option to increase the rewards for completion of a particular task, but tasks could not be targeted by context.

Experience sampling work, in which a questionnaire appears on-screen when the mobile device detects that it is in a context of interest [90], could complement the collection of contextualised qualitative data. Incorporating this into applications released using the hybrid mass participation methodology, as a method to explore in more detail the generalisability of findings from local participants in contextual aware software would be an interesting next step.

Exploring the ethical responsibilities of researchers in large-scale and mass participation trials by consultation with researchers in this, and adjacent fields has already been shown to be worthy of further work.

More detailed examination of what is acceptable to participants in such trials is also an area in which work is needed. The acceptability of interruptions [48], as well as the extensive investigation of people's 'privacy preferences' in regards to sensor-based technologies [26] have been explored. However, as discussed by Barkhuus [26], the notion of privacy is contextually grounded — explicitly exploring what participants in a mass participation user study would find acceptable in relation to how much they were made to understand the trial process would allow researchers to balance the interaction they have with participants and avoid alienating large numbers of users by seeming to take advantage of their position as developers.

An area of future work which is now being actively pursued within the research group is how to refine the hybrid mass participation approach for iterative design. This methodology is being refined as an integral part of the project A Population Approach to Ubicomp System Design[1] [49].

The approach centres on an analogy with the biological understanding of a population; where each member of the population has a strong resemblance to the others, but also has small differences that make it unique. These differences allow the gradual evolution, and divergence, of species through natural selection.

In the population approach to software, each instance of a program supports end-user adaptation via component-based approaches. This affords variation within the 'population' of instances, driven by users' individual

[1] EPSRC Programme Grant EP/J007617/1, Dec 2011 – Nov 2016, £4M

choices and social interaction. A key aspect of the approach is offering users appropriate models of patterns of variation and dynamism within the population, so as to support awareness as well as further adaptation.

It is intended that the mass participation method presented in this thesis will be used, and extended, in this new project so as to provide insight into experimental deployments based on the population approach. For example, the project will carry out traditional user trials involving locally recruited groups of participants, paid to use applications and tools in controlled contexts and ways. It will also make selected (and suitably anonymised) experimental findings available to the wider community of users, and canvass them for opinions on recent findings as well as potential experiments — the latter being useful in itself but also provoking users' reflection on system design and use. Similar canvassing and reflection will be done for programming work. Overall, the project will bring 'lead users' more into the design process than in prior work, exploring new interactions such as keeping them in close touch with our analyses and plans for local experiments.

This example is intended to show that the hybrid mass participation approach affords new avenues of research into tools and design processes, and corresponding new work on good ethical practice. It is hoped that this thesis will have an indirect contribution, beyond those listed in the previous section, in helping others develop further and better ways to run large scale trials in such a way as to keep the qualitative detail necessary to inform design while gaining the diversity of users needed for claims of generalisability.

Bibliography

[1] G. D. Abowd and E. Mynatt, "Charting past, present, and future research in ubiquitous computing," *ACM Transactions on Computer-Human Interaction*, vol. 7, no. 1, Mar. 2000.

[2] J. Accot and S. Zhai, "Beyond Fitts' law: models for trajectory-based HCI tasks," *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 295–302, 1997. [Online]. Available: http://dl.acm.org/citation.cfm?id=258760

[3] A. Ahmed, S. Benford, and A. Crabtree, "Digging in the crates: an ethnographic study of DJS' work," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, May 2012.

[4] L. A. Albertson, "Telecommunications as a travel substitute: Some psychological, organizational, and social aspects," *Journal of Communication*, vol. 27, no. 2, pp. 32–43, 1977.

[5] T. Alsheikh, J. Rode, and S. Lindley, "(Whose) value-sensitive design: a study of long-distance relationships in an Arabic cultural context," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 2011, pp. 75–84.

[6] M. Ames and M. Naaman, "Why we tag: motivations for annotation in mobile and online media," in *Proceedings of the 25th international conference on Human factors in computing systems*. ACM, Apr. 2007, pp. 971–980.

[7] Anon. About HTC. [Online]. Available: http://www.htc.com/www/about/?id=66338&lang=1033

[8] ——. App Store Review Guidelines. [Online]. Available: https://developer.apple.com/appstore/guidelines.html

[9] ——, "Global Mobile Phone Shipments Top 1 Billion Units in 2006," Strategy Analytics. [Online]. Available: http://www.strategyanalytics.com/default.aspx?mod=pressreleaseviewer&a0=3260

[10] ——. Key Facts - Facebook Newsroom. [Online]. Available: http://newsroom.fb.com/content/default.aspx?NewsAreaId=22

[11] ——. Overview of Jurisdiction – European Perspective. [Online]. Available: http://www.kentlaw.edu/cyberlaw/docs/rfc/euview.html

[12] ——. Request an Expedited App Review. [Online]. Available: https://developer.apple.com/appstore/contact/appreviewteam/index.html

[13] ——, "Data Protection Act 1998," 1998. [Online]. Available: http://www.legislation.gov.uk/ukpga/1998/29/contents

[14] ——, "64 million smart phones shipped worldwide in 2006 ," Canalys, Feb. 2007. [Online]. Available: http://www.canalys.com/newsroom/64-million-smart-phones-shipped-worldwide-2006

[15] ——. (2011, Jul.) 1 Billion Smartphones To Be Sold Each Year. [Online]. Available: http://news.accuracast.com/mobile-7471/1-billion-smartphones-to-be-sold-each-year/

[16] ——, "COM (2012).11 Regulation Of The European Parliament And Of The Council On The Protection Of Individuals With Regard To The Processing Of Personal Data And On The Free Movement Of Such Data (General Data Protection Regulation)," pp. 1–119, Jan. 2012. [Online]. Available: http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm

[17] ——, "Smart phones overtake client PCs in 2011," Canalys, Feb. 2012. [Online]. Available: http://www.canalys.com/newsroom/smart-phones-overtake-client-pcs-2011

[18] W. Ark, B. Shaw, A. Lelescu, and S. Stucky, "Data mining the cross-cultural communication gap," in *Proceeding of the 2009 international workshop on Intercultural collaboration*. ACM, Feb. 2009, pp. 199–202.

[19] K. B. Bærentsen and J. Trettvik, "An activity theory approach to affordance," in *Proceedings of the second Nordic conference on Human-computer interaction*. ACM Press, 2002, pp. 51–60.

[20] R. Ballagas, S. Kratz, J. Borchers, E. Yu, S. Walz, C. Fuhr, L. Hovestadt, and M. Tann, "REXplorer: a mobile, pervasive spell-casting game for tourists," *CHI'07 extended abstracts on Human factors in computing systems*, pp. 1929–1934, 2007.

[21] G. Baltas, "Determinants of internet advertising effectiveness: an empirical study," *International Journal of Market Research*, vol. 45, no. 4, pp. 505–513, 2003.

[22] P. Banister, *Qualitative methods in psychology: a research guide*. Open University Press, 1994.

[23] C. Banks and P. Zimbardo, "Interpersonal dynamics in a simulated prison." *International Journal of Criminology and Penology*, no. 1, pp. 69–97, 1973.

[24] M. Barbaro and T. Zeller, "A Face Is Exposed for AOL Searcher No. 4417749," *The New York Times*, Aug. 2006. [Online]. Available: http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000&en=996f61c946da4d34&ei=5088&partner=rssnyt&emc=rss

[25] E. Barker, *The making of a Moonie: choice or brainwashing?*, ser. Modern revivals in sociology. Gregg Revivals, 1993.

[26] L. Barkhuus, "The Mismeasurement of Privacy: Using Contextual Integrity to Reconsider Privacy in HCI," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 2012, pp. 367–376.

[27] L. Barkhuus, M. Chalmers, P. Tennent, and M. Hall, "Picking pockets on the lawn: the development of tactics and strategies in a mobile game," in *Proceedings of the 7th international conference on Ubiquitous Computing*, 2005, pp. 358–374.

[28] D. Bass. (2011, Jan.) Microsoft Shipped 2 Million Windows Phone 7 Units. [Online]. Available: http://www.bloomberg.com/news/2011-01-26/microsoft-says-it-shipped-2-million-windows-phones-last-quarter.html

[29] M. Bell, M. Chalmers, L. Barkhuus, M. Hall, S. Sherwood, P. Tennent, B. Brown, D. Rowland, and S. Benford, "Interweaving mobile games

with everyday life," *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 426, 2006.

[30] M. Bell, S. Reeves, B. Brown, S. Sherwood, and M. Chalmers, "Eye-Spy: supporting navigation through play," in *Proceedings of the 27th international conference on Human factors in computing systems*. ACM, Apr. 2009, pp. 123–132.

[31] R. Benbunan-Fich, S. Hiltz, and M. Turoff, "A comparative content analysis of face-to-face vs. asynchronous group decision making," *Decis Support Systems*, vol. 34, no. 4, pp. 457–469, 2003.

[32] S. Benford, A. Crabtree, M. Flintham, A. Drozd, R. Anastasi, M. Paxton, N. Tandavanitj, M. Adams, and J. Row-Farr, "Can you see me now?" *ACM Transactions on Computer-Human Interaction*, vol. 13, no. 1, pp. 100–133, 2002.

[33] S. Benford, M. Flintham, A. Drozd, and R. Anastasi, "Uncle Roy All Around You: Implicating the City in a Location-Based Performance," in *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*, 2004, pp. 357–358.

[34] J. H. Billington, "Statement of the Librarian of Congress Relating to Section 1201 Rulemaking," Jul. 2010. [Online]. Available: http://www.copyright.gov/1201/2010/Librarian-of-Congress-1201-Statement.html

[35] L. Blatt, M. Jacobson, and S. Miller, "Designing and equipping a usability laboratory," *Behaviour & Information Technology*, vol. 13, no. 1-2, pp. 81–93, Jan. 1994.

[36] J. Bloomberg, J. Giacomi, A. Mosher, and P. Swenton-Wall, "Ethnographic field methods and their relation to design," D. Schuler and A. Namioka, Eds. Participatory design: Principles and practices, 1993.

[37] S. Bødker, "Applying Activity Theory to Video Analysis: How to Make Sense of Video Data in Human- Computer Interaction," in *Context and Consciousness: Activity Theory and Human-Computer Interaction*, B. A. Nardi, Ed. The MIT Press, 1996, pp. 147–174.

[38] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer, "Falling asleep with Angry Birds, Facebook and Kindle: a large scale study

on mobile application usage," in *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM Press, 2011, pp. 47–56.

[39] B. Brian Butler, L. Sproull, S. Kiesler, and R. Kraut, "Community Effort in Online Groups: Who Does the Work and Why," *Research in Technologically Supported Work*, vol. 11, pp. 171–194, 2002.

[40] A. Bridgwater. (2011, Nov.) The many ages of mobile application development. [Online]. Available: http://www.computerweekly.com/blogs/cwdn/2011/11/a-jurassic-history-of-mobile-application-development.html

[41] B. Brown, M. Chalmers, M. Bell, M. Hall, I. MacColl, and P. Rudman, "Sharing the square: collaborative leisure in the city streets," *Proceedings of the Tenth European Conference on Computer Supported Colaborative Work*, pp. 427–447, 2005.

[42] B. Brown, S. Reeves, and S. Sherwood, "Into the wild: Challenges and opportunities for field trial methods," *Proceedings of the 2011 annual conference on Human factors in computing systems*, pp. 1657–1666, 2011.

[43] A. Budde and F. Michahelles, "Product Empire—Serious play with barcodes," in *Internet of Things (IOT)*. IEEE, 2010, pp. 1–7.

[44] P. Buddharaju, Y. Fujiki, I. Pavlidis, and E. Akleman, "A novel way to conduct human studies and do some good," *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pp. 4699–4702, 2010.

[45] E. Calvillo Gámez, P. Cairns, J. Gow, J. Back, and E. Capstick, "Video games as research instruments," in *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*. ACM, 2010, pp. 4493–4496.

[46] A. Campbell, S. Eisenman, K. Fodor, N. Lane, H. Lu, E. Miluzzo, M. Musolesi, R. Peterson, and X. Zheng, "Transforming the social networking experience with sensing presence from mobile phones," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*. ACM, 2008, pp. 367–368.

[47] R. Campbell-Burt, "Federation Asks: Do you know what you're agreeing to?" Federation Against Software Theft, Mar. 2006. [Online]. Available: http://www.fastiis.org/resources/press_details/id/304/federation-asks-do-you-know-what-yoursquore-agreeing-to/

[48] S. Carter, J. Mankoff, and J. Heer, "Momento: support for situated ubicomp experimentation," in *Proceedings of the 25th international conference on Human factors in computing systems*, 2007, pp. 134–144.

[49] M. Chalmers, "A population approach to ubicomp system design," in *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference*. British Computer Society, 2010.

[50] M. Chalmers, M. Bell, B. Brown, M. Hall, S. Sherwood, and P. Tennent, "Gaming on the edge: using seams in ubicomp games," in *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*. ACM, Jun. 2005.

[51] M. Chalmers, D. McMillan, A. Morrison, H. Cramer, M. Rost, and W. Mackay, "Ethics, logs and videotape: ethics in large scale user trials and user generated content," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*. ACM, 2011, pp. 2421–2424.

[52] F. Chehimi and P. Coulton, "Motion controlled mobile 3D multiplayer gaming," in *ACE '08*. ACM, Dec. 2008.

[53] A. D. Cheok, K. H. Goh, W. Liu, F. Farbiz, S. L. Teo, H. S. Teo, S. P. Lee, Y. Li, S. W. Fong, and X. Yang, "Human Pacman: a mobile wide-area entertainment system based on physical, social, and ubiquitous computing," in *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*. ACM, Sep. 2004.

[54] A. Chin, "Finding Cohesive Subgroups and Relevant Members in the Nokia Friend View Mobile Social Network," in *The 12th IEEE International Conference on Computational Science and Engineering*, 2009, pp. 278–283.

[55] B. Choi, I. Lee, J. Kim, and Y. Jeon, "A qualitative cross-national study of cultural influences on mobile data service design," in *Pro-*

ceedings of the 23rd international conference on Human factors in computing systems. ACM, 2005, pp. 661–670.

[56] G. Cockton, "Value-centred HCI," in *Proceedings of the third Nordic conference on Human-computer interaction*. ACM, Oct. 2004.

[57] P. Collins, S. Shukla, and D. Redmiles, "Activity theory and system design: A view from the trenches," *Computer Supported Cooperative Work*, vol. 11, no. 1, pp. 55–80, 2002.

[58] W. A. Corsaro, "Something Old and Something New: The Importance of Prior Ethnography in the Collection and Analysis of Audiovisual Data," *Sociological Methods & Research*, vol. 11, no. 2, pp. 145–166, Nov. 1982.

[59] A. Crabtree and T. Rodden, "Domestic routines and design for the home," *Computer Supported Cooperative Work*, vol. 13, no. 2, pp. 191–220, 2004.

[60] A. Crabtree, T. Rodden, P. Tolmie, and G. Button, "Ethnography considered harmful," in *Proceedings of the 27th international conference on Human factors in computing systems*, Apr. 2009, pp. 879–888.

[61] H. Cramer, M. Rost, and F. Bentley, "An Introduction to research in the Large," *International Journal of Mobile Human Computer Interaction*, vol. 3, no. 4, 2011.

[62] P. I. Crawford, D. Turton, and G. C. f. V. Anthropology, *Film as ethnography*. Manchester University Press, Nov. 1992.

[63] C. R. B. de Souza and D. F. Redmiles, "Opportunities for Extending Activity Theory for Studying Collaborative Software Development," in *Workshop in Applying Activity Theory to CSCW research and practice*, 2003, pp. 14–18.

[64] B. DeKoven, "Changing the Game," *The Game Design Reader: A Rules of Play Anthology*, p. 548, 1978.

[65] S. Denef, L. Ramirez, T. Dyrks, and G. Stevens, "Handy navigation in ever-changing spaces: an ethnographic study of firefighting practices," in *Proceedings of the 7th ACM conference on Designing interactive systems*. ACM, Feb. 2008.

[66] M. Denscombe, *The Good Research Guide for Small-scale Social Research Projects*. Open University Press, Aug. 2010.

[67] N. K. Denzin, *The Research Act : A Theoretical Introduction to Sociological Methods*. Aldine de Gruyter, Mar. 2009.

[68] H. D'Hooge, L. Dalton, H. Shwe, and D. Lieberman, "Smart toys: brave new world?" in *Proceedings of the 18th international conference extended abstracts on Human factors in computing systems*, 2000, pp. 247–248.

[69] E. Directive, "95/46/EC-The Data Protection Directive," *Official Journal of the European Communities*, 1995. [Online]. Available: http://europa.eu/legislation_summaries/information_society/data_protection/l14012_en.htm

[70] P. Dourish, "Implications for Design," in *Proceedings of the 24th international conference on Human factors in computing systems*. ACM Press, 2006, pp. 541–550.

[71] ——, "Responsibilities and implications: further thoughts on ethnography and design," in *Proceedings of the 2007 conference on Designing for User eXperiences*. ACM, Nov. 2007.

[72] P. Dourish and G. Bell, *Divining a Digital Future*, ser. Mess and Mythology in Ubiquitous Computing. The MIT Press, Apr. 2011.

[73] P. Dourish and G. Button, "On "Technomethodology"," *Human-Computer Interaction*, vol. 13, no. 4, pp. 395–432, 1998.

[74] S. Dow, J. Lee, C. Oezbek, B. MacIntyre, J. Bolter, and M. Gandy, "Exploring spatial narratives and mixed reality experiences in Oakland Cemetery," *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pp. 51–60, 2005.

[75] M. DuFon, "Video recording in ethnographic SLA research: Some issues of validity in data collection," *Language Learning & Technology*, vol. 6, no. 1, pp. 40–59, 2002.

[76] T. Duryee. (2010, May) Beyond AdMob: There's Plenty More Mobile Ad Networks To Go Around. [Online]. Available: http://paidcontent.org/2010/05/11/419-beyond-admob-theres-plenty-more-mobile-ad-networks-to-go-around/

[77] C. B. Eiben, J. B. Siegel, J. B. Bale, S. Cooper, F. Khatib, B. W. Shen, F. Players, B. L. Stoddard, Z. Popovic, and D. Baker, "Increased Diels-Alderase activity through backbone remodeling guided by Foldit players," *Nature Biotechnology*, vol. 30, no. 2, pp. 190–192, Jan. 2012.

[78] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, "TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones," in *Proceedings of the 9th USENIX conference on Operating systems design and implementation*. USENIX Association, Oct. 2010.

[79] Y. Engeström, R. Miettinen, and R. Punamäki-Gitai, *Perspectives on Activity Theory*. Cambridge University Press, 1999.

[80] F. Erickson, "Audiovisual Records as a Primary Data Source," *Sociological Methods & Research*, vol. 11, no. 2, pp. 213–232, 1982.

[81] ——, "Ethnographic microanalysis," in *Sociolinguistics and Language Teaching*, S. McKay and N. H. Hornberger, Eds., 1992.

[82] F. Erickson and J. J. Shultz, *The counselor as gatekeeper: social interaction in interviews*. Academic Press, 1982.

[83] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin, "Diversity in Smartphone Usage," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*. New York, New York, USA: ACM Press, 2010, pp. 179–194.

[84] H. Faste and H. Lin, "The untapped promise of digital mind maps," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, May 2012, pp. 1017–1026.

[85] D. Ferreira, A. Dey, and V. Kostakos, "Understanding human-smartphone concerns: a study of battery life," in *Proceedings of the 9th international conference on Pervasive computing*. Springer, 2011, pp. 19–33.

[86] D. M. Fetterman, *Ethnography step-by-step*. Sage Publications, Inc, 2010.

[87] S. Fiksdal, "Verbal and nonverbal strategies of rapport in cross-cultural interviews," *Linguistics and Education*, vol. 1, no. 1, pp. 3–17, 1988.

[88] K. Fisher, S. Counts, and A. Kittur, "Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, May 2012, pp. 247–256.

[89] J. Frederking, M. Cruz, M. Baskinger, and K. Overbeeke, "Beyond the screen: designing immersive collaborative experiences for children through digital and physical interaction," in *Proceedings of the 7th ACM conference on Designing interactive systems*, Feb. 2008, pp. 156–163.

[90] J. Froehlich, M. Y. Chen, I. Smith, and F. Potter, "Voting with your feet: An investigative study of the relationship between place visit behavior and preference," in *Proceedings of the 8th international conference on Ubiquitous Computing*, 2006, pp. 333–350.

[91] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, pp. 1–53, 2010.

[92] D. Gammon, T. Sorlie, S. Bergvik, and T. S. Hoifodt, "Psychotherapy supervision conducted by videoconferencing: a qualitative study of users' experiences," *Journal of Telemedicine and Telecare*, vol. 4, no. suppl 1, pp. 33–35, 1998.

[93] S. M. Gass and N. Houck, *Interlanguage Refusals: A Cross-cultural Study of Japanese-English* . Walter de Gruyter, 1999.

[94] C. Geertz, *Deep play: notes on the Balinese cockfight*. American Academy of Arts and Sciences, 1972.

[95] P. Gilbertson, P. Coulton, F. Chehimi, and T. Vajk, "Using "tilt" as an interface to control "no-button" 3-D mobile games," *ACM Computers in Entertainment*, vol. 6, no. 3, pp. 1–13, Oct. 2008.

[96] A. Girardello, "AppAware: which mobile applications are hot?" in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, 2010, pp. 431–434.

[97] A. Girardello and F. Michahelles, "Explicit and implicit ratings for mobile applications," in *Service Science - Neue Perspektiven für die Informatik*, 2010, pp. 606–612.

[98] ——, "Twiphone: Sharing communication behavior on twitter," in *What can the Internet of Things do for the Citizen Workshop at The Eighth International Conference on Pervasive Computing*, 2010.

[99] N. Glance and D. Snowdon, "Pollen: using people as a communication medium," *Computer Networks*, vol. 35, no. 4, pp. 429–442, 2001.

[100] B. Glaser and A. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*, ser. Observations (Chicago, Ill.). Aldine de Gruyter, 1967.

[101] R. Goldman-Segall, "Interpreting video data: introducing a Significance Measure to layer description," *Journal of Educational Multimedia and Hypermedia*, vol. 2, no. 3, pp. 261–281, 1993.

[102] ——, "Configurational validity: a proposal for analyzing ethnographic multimedia narratives," *Journal of Educational Multimedia and Hypermedia*, vol. 4, no. 2-3, pp. 163–182, Jan. 1995.

[103] A. Greenfield, *Everyware: The Dawning Age of Ubiquitous Computing*. Peachpit Press, Mar. 2006.

[104] A. D. Grimshaw, "Sound-Image Data Records for Research on Social Interaction," *Sociological Methods & Research*, vol. 11, no. 2, pp. 121–144, 1982.

[105] M. Hall, M. Bell, A. Morrison, S. Reeves, S. Sherwood, and M. Chalmers, "Adapting ubicomp software and its evaluation," in *Proceedings of the 1st ACM SIGCHI symposium on Engineering interactive computing systems*. ACM, Jul. 2009, pp. 143–148.

[106] S. Harsh, "AdMob Mobile Metrics Highlights," AdMob, May 2010. [Online]. Available: http://www.scribd.com/doc/33741060/May-2010-AdMob-Mobile-Metrics-Highlights

[107] K. Hastrup, "Anthropological visions: Some notes on visual and textual authority," *Film as ethnography*, 1992.

[108] S. Hattenstone, "Simon Hattenstone meets the 'Elephant Men' to hear about their fight for compensation," *The Guardian*, p. 20, Feb. 2007. [Online]. Available: http://www.guardian.co.uk/society/2007/feb/17/health.lifeandhealth

[109] N. V. Hay-Gibson, "Interviews via VoIP: Benefits and Disadvantages within a PhD study of SMEs," *Library and Information Research*, vol. 33, 2009.

[110] Y. He, C. Zhao, and P. Hinds, "Understanding information sharing from a cross-cultural perspective," in *Proceedings of the 28th international conference on Human factors in computing systems*. ACM, 2010, pp. 3823–3828.

[111] J. Heary. (2011, Jan.) Cisco Security Expert: Apple iOS Facetime: Is anyone really using it? [Online]. Available: http://www.networkworld.com/community/blog/apple-ios-facetime-anyone-really-using-it

[112] N. Henze, "Push the study to the app store: Evaluating off-screen visualizations for maps in the android market," in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, 2010, pp. 373–374.

[113] ——, "Hit it!: an apparatus for upscaling mobile HCI studies," in *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems*. ACM, May 2012.

[114] N. Henze and S. Boll, "It does not Fitts my data! analysing large amounts of mobile touch data," in *Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction*. Springer, 2011, pp. 564–567.

[115] ——, "Release your app on Sunday eve: finding the best time to deploy apps," in *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, 2011, pp. 581–586.

[116] N. Henze, B. Poppinga, and S. Boll, "Experiments in the wild: public evaluation of off-screen visualizations in the Android market," *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, Oct. 2010.

[117] N. Henze, E. Rukzio, and S. Boll, "100,000,000 taps: analysis and improvement of touch performance in the large," in *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, Aug. 2011.

[118] ——, "Observational and experimental investigation of typing behaviour using virtual keyboards for mobile devices," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, May 2012.

[119] G. Homans, "Anxiety And Ritual: The Theories Of Malinowski And RadcliffeBrown," *American Anthropologist*, vol. 43, no. 2, pp. 164–172, 1941.

[120] J. Hughes, V. King, T. Rodden, and H. Andersen, "The role of ethnography in interactive systems design," *interactions*, vol. 2, no. 2, pp. 56–65, 1995.

[121] L. Hughes, "Beyond the Rules of the Game: Why Are Rooie Rules Nice?" *The Game Design Reader: A Rules of Play Anthology*, pp. 504–518, 1983.

[122] S. Hughes, "SHAKE–Sensor Hardware Accessory for Kinesthetic Expression," in *The Third International Conference on Enactive Interfaces*, 2006, p. 155.

[123] T. Igarashi, T. Motoyoshi, J. Takai, and T. Yoshida, "No mobile, no life: Self-perception and text-message dependency among Japanese high school students," *Computers in Human Behavior*, vol. 24, no. 5, pp. 2311–2324, 2008.

[124] M. Iino, "The Trap of Generalization: A Case of Encountering a New Culture." *Working Papers in Educational Linguistics*, vol. 9, no. 1, p. 27, 1993.

[125] ——, "Issues of Video Recording in Ethnographic Studies." in *Annual Meeting of the American Association for Applied Linguistics*, 1998.

[126] N. James, "Credibility, authenticity and voice: dilemmas in online interviewing," *Qualitative Research*, vol. 6, no. 3, pp. 403–420, Aug. 2006.

[127] K. L. Jensen, "Evaluating Complex Mobile Interaction in the Wild," *Research in the Wild (RITW 2012) as part of ACM DIS 2012*, pp. 1–5, Apr. 2012. [Online]. Available: https://www.dropbox.com/s/jvrb957jxayd70m/L%C3%B8vborgJensen_DIS12_RITW.pdf

[128] E. E. Jones, F. Rhodewalt, S. Berglas, and J. A. Skelton, "Effects of strategic self-presentation on subsequent self-esteem." *Journal of Personality and Social Psychology*, vol. 41, no. 3, pp. 407–421, 1981.

[129] E. Kanjo and P. Astheimer, "Interactive environment by narrative playmates toys," *ACM SIGGROUP Bulletin*, vol. 23, no. 2, pp. 6–7, 2002.

[130] V. Kaptelinin, "Activity Theory: Implications for Human-Computer Interaction," in *Context and Consciousness: Activity Theory and Human-Computer Interaction*, B. A. Nardi, Ed. The MIT Press, 1996, pp. 103–116.

[131] ——, "Computer-Mediated Activity: Functional Organs in Social and Developmental Contexts," in *Context and Consciousness: Activity Theory and Human-Computer Interaction*, B. A. Nardi, Ed. The MIT Press, 1996, pp. 45–68.

[132] ——, "Activity Theory," in *Encyclopedia of Human-Computer Interaction*. Aarhus, Denmark: The Interaction-Design.org Foundation, 2012.

[133] S. Karpischek, G. Geron, and F. Michahelles, "Towards a Better Understanding of Mobile Shopping Assistants–A Large Scale Usage Analysis of a Mobile Bargain Finder Application," *Proceedings of the 2nd international workshop on Research in the large*, pp. 17–20, 2011.

[134] S. Karpischek and F. Michahelles, "my2cents - Digitizing consumer opinions and comments about retail products ," in *Internet of Things (IOT)*. IEEE, 2010, pp. 1–7.

[135] T. Karrer, M. Wittenhagen, and J. Borchers, "DragLocks: handling temporal ambiguities in direct manipulation video navigation," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, May 2012, pp. 623–626.

[136] S. Kaviraj, Y. Ting, M. Bureau, S. Shabala, R. Crockett, J. Silk, C. Lintott, A. Smith, W. Keel, and K. Masters, "Galaxy Zoo: Dust and molecular gas in early-type galaxies with prominent dust lanes," *Monthly Notices of the Royal Astronomical Society*, vol. 423, no. 1, pp. 49–58, 2011.

[137] S. Kayan, S. Fussell, and L. Setlock, "Cultural differences in the use of instant messaging in Asia and North America," in *Proceedings of the*

2006 20th anniversary conference on Computer supported cooperative work. ACM, 2006, pp. 525–528.

[138] G. Kelly. (2009, Oct.) Symbian App Store Goes Live. [Online]. Available: http://www.trustedreviews.com/news/Symbian-App-Store-Goes-Live

[139] T. Kennedy, "Research In Motion Reports Year-End And Fourth Quarter Results For Fiscal 2012," Research In Motion, Mar. 2012. [Online]. Available: http://press.rim.com/release.jsp?id=5638

[140] F. Kensing and J. Simonsen, "MUST: A method for participatory design," *Human-Computer Interaction*, vol. 13, no. 2, pp. 167–198, 1998.

[141] A. Khelil, C. Becker, J. Tian, and K. Rothermel, "An epidemic model for information diffusion in MANETs," in *Proceedings of the 5th ACM international workshop on Modeling analysis and simulation of wireless and mobile systems*. ACM, Sep. 2002, pp. 54–60.

[142] D. King and P. H. Delfabbro, "Understanding and assisting excessive players of video games: A community psychology perspective," *Australian Community Psychologist*, vol. 21, no. 1, 2009.

[143] J. Kjeldskov, M. Skov, B. S. Als, and R. T. Høegh, "Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field," *Mobile Human-Computer Interaction - Mobilehci 2004, Proceedings*, vol. 3160, pp. 61–73, 2004.

[144] L. Koffman and E. Macdonald, *The law of contract*. Oxford University Press, Dec. 2007.

[145] R. Koster and W. Wright, "A theory of fun for game design," *Paraglyph Press*, 2004.

[146] M. Kuhn, R. Wattenhofer, and S. Welten, "Social audio features for advanced music retrieval interfaces," *Proceedings of the international conference on Multimedia*, pp. 411–420, 2010.

[147] S. Kurkowski, T. Camp, and M. Colagrosso, "MANET simulation studies: the incredibles ," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 9, no. 4, pp. 50–61, Oct. 2005.

[148] K. Kuutti, "Activity Theory as a Potential Framework for Human-Computer Interaction Research," in *Context and Consciousness: Activity Theory and Human-Computer Interaction*, B. A. Nardi, Ed. The MIT Press, 1996, pp. 17–44.

[149] S. Lee and K. Mase, "Activity and location recognition using wearable sensors," *Pervasive Computing, IEEE*, vol. 1, no. 3, pp. 24–32, 2002.

[150] W. Lee, W. Woo, and J. Lee, "Tarboard: Tangible augmented reality system for table-top game environment," in *Proceedings Second International Workshop on Pervasive Gaming Applications*, 2005.

[151] A. N. Leontyev and M. Cole, "The Development of Mind," Erythrós Press and Media, Mar. 2009.

[152] K. Li and S. Counts, "Exploring social interactions and attributes of casual multiplayer mobile gaming," *Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology*, pp. 696–703, 2007.

[153] C. Licoppe and Y. Inada, "Emergent Uses of a Multiplayer Location-aware Mobile Game: the Interactional Consequences of Mediated Encounters," *Mobilities*, vol. 1, no. 1, pp. 39–61, Mar. 2006.

[154] ——, "'Timid encounters': a case study in the use of proximity-based mobile technologies," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, May 2012.

[155] S. Lindtner, G. R. Hayes, and P. Dourish, "Towards a framework of publics: Re-encountering Media Sharing and Its User," *ACM Transactions on Computer-Human Interaction*, vol. 18, no. 2, p. 23, Jun. 2011.

[156] R. Macefield, "Usability studies and the Hawthorne Effect," *Journal of Usability Studies*, vol. 2, no. 3, pp. 145–154, 2007.

[157] W. Mackay, "Ethics, lies and videotape. . . ," in *Proceedings of the 1995 SIGCHI conference on Human Computer Interaction*, 1995, pp. 138–145.

[158] C. Magerkurth, A. D. Cheok, R. Mandryk, and T. Nilsen, "Pervasive games: bringing computer entertainment back to the real world," *Computers in Entertainment*, vol. 3, no. 3, pp. 4–24, Jul. 2005.

[159] R. Magielse and P. Markopoulos, "HeartBeat: an outdoor pervasive game for children," in *Proceedings of the 27th international conference on Human factors in computing systems*, Apr. 2009, pp. 2181–2184.

[160] B. Malinowski, *Argonauts of the Western Pacific: An account of native enterprise and adventure in the archipelagoes of Melanesian New Guinea*. Routlege, 2002.

[161] R. Mandryk and D. Maranan, "False prophets: exploring hybrid board/video games," in *Proceedings of the 20th international conference extended abstracts on Human factors in computing systems*, Apr. 2002, pp. 640–641.

[162] G. Mark and S. Voida, ""A Pace Not Dictated by Electrons": An Empirical Study of Work Without Email," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 2012, pp. 555–564.

[163] P. Marsh, E. Rosser, and R. Harré, *The Rules of Disorder*, ser. Social worlds of childhood. Routledge, 1980.

[164] D. Mauney, J. Howarth, A. Wirtanen, and M. Capra, "Cultural similarities and differences in user-defined gestures for touchscreen user interfaces," in *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*. ACM, 2010, pp. 4015–4020.

[165] H. A. McKee and J. E. Porter, *The Ethics of Internet Research: A Rhetorical, Case-Based Process*. Peter Lang Publishing, Aug. 2009.

[166] D. McMillan, A. Morrison, and M. Chalmers, "A Comparison of Distribution Channels for Large-Scale Deployments of iOS Applications," *International Journal of Mobile Human Computer Interaction*, vol. 3, no. 4, pp. 1–17, 2011.

[167] E. McMullin, "Values in science," in *Proceedings Of The Biennial Meeting Of The Philosophy Of Science Association*. JSTOR, 1982, pp. 3–28.

[168] E. C.-G. mez, J. Gow, and P. Cairns, "Introduction to special issue: Video games as research instruments," *Entertainment Computing*, vol. 2, no. 1, pp. 1–2, Apr. 2011.

[169] D. Millen, "Rapid ethnography: time deepening strategies for HCI field research," in *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, Aug. 2000, pp. 280–286.

[170] T. Miller and C. Monaghan, "Apple's App Store Downloads Top 25 Billion," Apple Inc., Mar. 2012. [Online]. Available: http://www.apple.com/pr/library/2012/03/05Apples-App-Store-Downloads-Top-25-Billion.html

[171] E. Miluzzo, N. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. Eisenman, X. Zheng, and A. Campbell, "Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*. ACM, 2008, pp. 337–350.

[172] E. Miluzzo, N. Lane, H. Lu, and A. Campbell, "Research in the app store era: Experiences from the cenceme app deployment on the iphone," in *Proceedings of The First International Workshop Research in the Large: Using App Stores, Markets, and other wide distribution channels*, 2010.

[173] A. Morrison, M. Bell, and M. Chalmers, "Visualisation of spectator activity at stadium events," *Information Visualisation, 2009 13th International Conference*, pp. 219–226, 2009.

[174] A. Morrison, P. Tennent, and M. Chalmers, "Coordinated visualisation of video and system log data," in *Proceedings of the Fourth International Conference on Coordinated & Multiple Views in Exploratory Visualization*, 2006, pp. 91–102.

[175] J. Murchison, *Ethnography Essentials*, ser. Designing, Conducting, and Presenting Your Research. Jossey-Bass, Jan. 2010.

[176] D. Mwanza, "Where Theory meets Practice: A Case for an Activity Theory based Methodology to guide Computer System Design," in *Proceedings of the 3rd IFIP TC 3 international conference on Human-computer interaction*, 2001, pp. 342–349.

[177] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pp. 111–125, 2008.

[178] ——, "De-anonymizing Social Networks," in *The 30th IEEE Symposium on Security and Privacy*, 2009, pp. 173–187.

[179] B. A. Nardi, "Activity Theory and Human-Computer Interaction," in *Context and Consciousness: Activity Theory and Human-Computer Interaction*, B. A. Nardi, Ed. The MIT Press, 1996, pp. 7–16.

[180] C. Nguyen, Y. Niu, and F. Liu, "Video Summagator: An Interface for Video Summarization and Navigation," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM Press, 2012, pp. 647–650.

[181] J. Norris, H. Schnädelbach, and G. Qiu, "CamBlend: an object focused collaboration tool," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, May 2012, pp. 627–636.

[182] S. Ochoa, J. Rasmussen, C. Robson, and M. Salib, "Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study," *Unpublished Student Paper*, 2001.

[183] R. J. Ocker and G. J. Yaverbaum, "Asynchronous Computer-mediated Communication versus Face-to-face Collaboration: Results on Student Learning, Quality and Satisfaction ," *Group Decision and Negotiation*, vol. 8, no. 5, pp. 427–440, 1999.

[184] J. Oh, J. Herrera, N. Bryan, L. Dahl, and G. Wang, "Evolving the mobile phone orchestra," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2010, pp. 82–87.

[185] P. Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review*, vol. 57, p. 1701, 2010.

[186] E. Oliver, "A survey of platforms for mobile networks research," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 12, no. 4, pp. 56–63, 2009.

[187] ——, "The challenges in large-scale smartphone user studies," in *Proceedings of the 2nd ACM International Workshop on Hot Topics in Planet-scale Measurement*. ACM, 2010, p. 5.

[188] E. Oliver and S. Keshav, "An empirical approach to smartphone energy level prediction," in *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 2011, pp. 345–354.

[189] E. O'Neill, V. Kostakos, T. Kindberg, and A. Schiek, "Instrumenting the city: Developing methods for observing and understanding the digital cityscape," in *Proceedings of the 8th international conference on Ubiquitous Computing*, 2006, pp. 315–332.

[190] S. N. Patel, J. A. Kientz, G. R. Hayes, S. Bhat, and G. D. Abowd, "Farther than you may think: an empirical investigation of the proximity of users to their mobile phones," in *Proceedings of the 8th international conference on Ubiquitous Computing*. Springer-Verlag, Sep. 2006.

[191] E. Paulos and K. Anderson, "UbiComp in the urban frontier," in *Proceedings of the UbiComp in the Urban Frontier Workshop*, 2004, pp. 5–8.

[192] L. Pelusi, A. Passarella, and M. Conti, "Opportunistic networking: data forwarding in disconnected mobile ad hoc networks," *IEEE Communications Magazine*, vol. 44, no. 11, pp. 134–141, 2006.

[193] R. Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 55–64, 2003.

[194] W. Piekarski and B. Thomas, "ARQuake: the outdoor augmented reality gaming system," *Communications of the ACM*, vol. 45, no. 1, Jan. 2002.

[195] M. Pielot, N. Henze, and S. Boll, "Experiments in app stores-How to ask users for their consent," in *Proceedings of the workshop on Ethics, Logs & Videotape*, 2011.

[196] L. Plowman, Y. Rogers, and M. Ramage, "What are workplace studies for?" *Proceedings of the fourth conference on European Conference on Computer-Supported Cooperative Work*, pp. 309–324, 1995.

[197] B. Poppinga, M. Pielot, N. Henze, and S. Boll, "Unsupervised User Observation in the App Store: Experiences with the Sensor-based Evaluation of a Mobile Pedestrian Navigation Application," in *Proceedings of the 1st International Workshop on Observing the Mobile User Experience*, 2010, pp. 41–43.

[198] J. Preece, T. Carey, Y. Rogers, S. Holland, H. Sharp, and D. Benyon, *Human-Computer Interaction*. Addison Wesley Publishing Company, 1994.

[199] R. Pye and E. Williams, "Teleconferencing: is video valuable or is audio adequate?" *Telecommunications Policy*, vol. 1, no. 3, pp. 230–241, 1977.

[200] S. Radwanick and C. Aquino, "Key Insights from 2011 and What They Mean for the Coming Year," Tech. Rep., Mar. 2012. [Online]. Available: http://www2.comscore.com/l/1552/2012-02-22/2ymhbd

[201] D. Randall, M. Mu, N. Race, and M. Rouncefield, "Research in the Wild: the Living Lab," *Research in the Wild (RITW 2012) as part of ACM DIS 2012*, pp. 1–5, Apr. 2012. [Online]. Available: https://www.dropbox.com/s/robkio01vq2pj8s/Randal-Mu-Race-Rouncefield_DIS12_RITW.pdf

[202] P. Rau, J. Liu, S. Verhasselt, T. Kato, and C. Schlick, "Different time management behaviors of Germans, Chinese and Japanese," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 2011, pp. 701–704.

[203] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 30–38, 2007.

[204] P. Reason and J. Rowan, *Human inquiry: a sourcebook of new paradigm research*. J. Wiley, 1981.

[205] S. Reeves, "Studying social organisation with video," in *Research Squared Conference*, Jul. 2011, pp. 1–10.

[206] S. Reinharz and L. Davidman, *Feminist methods in social research*. Oxford University Press, 1992.

[207] R. Riccamboni, A. Mereu, and C. Boscarol, "Keys to Nature: A test on the iPhone market," in *Tools for Identifying Biodiversity: Progress and Problems*. EUT Edizioni Università di Trieste, 2010, pp. 445–450.

[208] B. Rice, "The Hawthorne effect: Persistence of a flawed theory," *Psychology Today*, vol. 16, no. 2, pp. 70–74, 1982.

[209] T. Ricker. (2007, Jul.) iPhone Hackers: "we have owned the filesystem". [Online]. Available: http://www.engadget.com/2007/07/10/iphone-hackers-we-have-owned-the-filesystem/

[210] M. Risjord, "The politics of explanation and the origins of ethnography," *Perspectives on Science*, vol. 8, no. 1, pp. 29–52, 2000.

[211] Y. Rogers, K. Connelly, and L. Tedesco, "Why it's worth the hassle: The value of in-situ studies when designing UbiComp," *Proceedings of the 9th international conference on Ubiquitous computing*, 2007.

[212] D. Rudchenko, T. Paek, and E. Badger, "Text text revolution: a game that improves text entry on mobile touchscreen keyboards," *Proceedings of the 9th international conference on Pervasive computing*, pp. 206–213, 2011.

[213] A. Sahami Shirazi, M. Rohs, R. Schleicher, S. Kratz, A. Müller, and A. Schmidt, "Real-time nonverbal opinion sharing through mobile phones during sports events," in *Proceedings of the 2011 annual conference on Human factors in computing systems*. ACM, 2011, pp. 307–310.

[214] K. Salen and E. Zimmerman, "Rules of Play - Game Design Fundamentals," 2004.

[215] B. Sales and S. Folkman, *Ethics In Research With Human Participants*. American Psychological Association, 2000.

[216] W. Salmon, "Carl G. Hempel on the rationality of science," *The Journal of Philosophy*, vol. 80, no. 10, pp. 555–562, 1983.

[217] M. Satyanarayanan, "Swiss Army Knife or Wallet?" *IEEE Pervasive Computing*, vol. 4, no. 2, pp. 2–3, Apr. 2005.

[218] N. Schillewaert, "The future of online research: engaging and activating stakeholders." *International Journal of Market Research*, vol. 53, no. 4, pp. 560–562, Jul. 2011.

[219] K. Schmidt, "The critical role of workplace studies in CSCW," in *Workplace Studies: Recovering Work Practice and Informing Design,*, P. Luff, J. Hindmarsh, and C. Heath, Eds. Cambridge University Press, 2000.

[220] J. Scott. App Store Metrics. [Online]. Available: http://148apps.biz/app-store-metrics/

[221] G. Seland, I. D. Sørby, and I. Dybdahl, "Usability Laboratory as the Last Outpost before Implementation–Lessons Learnt from Testing New Patient Record Functionality," *Studies in health technology and informatics*, vol. 150, p. 404, 2009.

[222] S. Shahid, E. Krahmer, and M. Swerts, "GamE in action: Using the GamE paradigm as a tool for investigating human emotions," in *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, 2010.

[223] D. Shapiro, "The Limits of Ethnography: Combining Social Sciences for CSCW ," in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM Press, 1994, pp. 417–428.

[224] W. Sharrock and J. A. Hughes, "Ethnography in the workplace: Remarks on its theoretical bases," *Team Ethno Online*, vol. 1, no. 1, pp. 1–21, 2001.

[225] R. Sheldrake, "Experimenter effects in scientific research: How widely are they neglected," *Journal of Scientific Exploration*, vol. 12, no. 1, pp. 73–78, 1998.

[226] M. Slater, A. Antley, A. Davison, D. Swapp, C. Guger, C. Barker, N. Pistrang, and M. V. Sanchez-Vives, "A Virtual Reprise of the Stanley Milgram Obedience Experiments," *PLoS ONE*, vol. 1, no. 1, p. e39, Dec. 2006.

[227] T. Starner, B. Leibe, B. Singletary, K. Lyons, M. Gandy, and J. Pair, "Towards augmented reality gaming," in *Proceedings of IMAGINA*, 2000.

[228] G. Stocking Jr, "Special review: Empathy and antipathy in the Heart of Darkness: An essay review of Malinowski's field diaries," *Journal of the History of the Behavioral Sciences*, vol. 4, no. 2, pp. 189–194, 1968.

[229] K. Strater and H. R. Lipford, "Strategies and struggles with privacy in an online social networking community," in *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*. British Computer Society, Sep. 2008, pp. 111–119.

[230] J. E. Sturges and K. J. Hanrahan, "Comparing Telephone and Face-to-Face Qualitative Interviewing: a Research Note," *Qualitative Research*, vol. 4, no. 1, pp. 107–118, Apr. 2004.

[231] L. A. Suchman, *Plans and situated actions: the problem of human-machine communication*. Cambridge University Press, 1987.

[232] K. Swan, "Building Learning Communities in Online Courses: the importance of interaction," *Education, Communication & Information*, vol. 2, no. 1, pp. 23–49, May 2002.

[233] L. Sweeney, "Simple Demographics Often Identify People Uniquely," *Carnegie Mellon University, Data Privacy Working Paper*, vol. 3, 2000.

[234] B. Tedlock, "From participant observation to the observation of participation: The emergence of narrative ethnography," *Journal of Anthropological Research*, vol. 47, no. 1, pp. 69–94, 1991.

[235] E. Thoa. (2012, Jan.) Money and the App Store: a few figures that might help an indie developer. [Online]. Available: http://thegamebakers.com/money-and-the-app-store-a-few-figures-that-might-help-an-indie-developer.html

[236] P. Tolmie, S. Benford, and M. Flintham, "Act natural: instructions, compliance and accountability in ambulatory experiences," in *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems*, 2012, pp. 1519–1528.

[237] J. L. M. Trim and S. Bailly, "Common European framework of reference for languages," L'europe Division Des Politiques Linguistiques, 2001.

[238] M. Vakulenko, S. Schuermans, A. Constantinou, and M. Kapetanakis, "Mobile Platforms: The Clash of Ecosystems ," Tech. Rep., Nov. 2011. [Online]. Available: http://www.visionmobile.com/research.php

[239] J. S. S. van 't Woud, J. A. C. Sandberg, and B. J. Wielinga, "The Mars Crowdsourcing Experiment," in *16th International Conference on Computer Games*, Aug. 2011, pp. 201–208.

[240] C. Vance, T. Andersen, and V. Vaiman, "Intercultural encounters in off shoring: liaison roles of host country nationals in a danish MNC

in Malaysia," in *Proceedings of the 3rd international conference on Intercultural collaboration*. ACM, 2010, pp. 19–28.

[241] R. Vatrapu and D. Suthers, "Technological intersubjectivity in computer supported intercultural collaboration," in *Proceedings of the 2009 international workshop on Intercultural collaboration*. ACM, 2009, pp. 155–164.

[242] S. von Watzdorf and F. Michahelles, "Accuracy of positioning data on smartphones," in *Proceedings of the 3rd International Workshop on Location and the Web*. ACM, Nov. 2010.

[243] L. Vygotsky, "Interaction between learning and development," in *Mind in society*, M. Cole and M. Gauvain, Eds. Harvard University Press, 1978, pp. 79–91.

[244] T. Walsh, P. Nurkka, and R. Walsh, "Cultural differences in smartphone user experience evaluation," in *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 2010, p. 24.

[245] G. Wang, "Designing smule's iphone ocarina," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2009, pp. 303–307.

[246] M. Weiser, "The computer for the 21 stcentury," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 3, no. 3, pp. 3–11, Jul. 1999.

[247] M. Welch, "Majority of smart phones now have touch screens," Canalys, Feb. 2010. [Online]. Available: http://www.canalys.com/newsroom/majority-smart-phones-now-have-touch-screens

[248] R. Whitehead and P. Spikes. Determining Internet Jurisdiction. [Online]. Available: http://www.nysscpa.org/cpajournal/2003/0703/features/f072403.htm

[249] C. Wolf. Standards for Internet Jurisdiction. [Online]. Available: http://library.findlaw.com/1999/Jan/1/241482.html

[250] Working Party on Conducting Research on the internet, "Guidelines for ethical practice in psychological research online ," The British Psychological Society, Jul. 2007.

[Online]. Available: http://www.bps.org.uk/publications/policy-guidelines/research-guidelines-policy-documents/research-guidelines-policy-docum

[251] Working Party on Ethical Guidelines for Psychological Research, "Code of Human Research Ethics," The British Psychological Society, 2011. [Online]. Available: http://www.bps.org.uk/sites/default/files/documents/code_of_human_research_ethics.pdf

[252] K. Wray, "A defense of Longino's social epistemology," *Philosophy of Science*, pp. 538–552, 1999.

[253] B. Yan and G. Chen, "AppJoy: personalized mobile application discovery," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*. ACM, Jun. 2011, pp. 113–126.

[254] S. Zhai, P. Kristensson, and P. Gong, "Shapewriter on the iPhone: from the laboratory to the real world," in *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, 2009, pp. 2667–2670.