A Hybrid Mass Participation Approach to Mobile Software Trials

Alistair Morrison[1], Donald McMillan[1], Stuart Reeves[2], Scott Sherwood[1], Matthew Chalmers[1]
[1]School of Computing Science,
University of Glasgow, UK
{alistair.morrison,donny.mcmillan,scott.sherwood,
matthew.chalmers}@glasgow.ac.uk

[2]Horizon Digital Economy Research,
University of Nottingham, UK
stuart@tropic.org.uk

ABSTRACT
User trials of mobile applications have followed a steady march out of the lab, and progressively further 'into the wild', recently involving 'app store'-style releases of software to the general public. Yet from our experiences on these mass participation systems and a survey of the literature, we identify a number of reported difficulties. We propose a hybrid methodology that aims to address these, by combining a global software release with a concurrent local trial. A phone–based game, created to explore the uptake and use of ad hoc peer-to-peer networking, was evaluated using this new hybrid trial method, combining a small-scale local trial (11 users) with a 'mass participation' trial (over 10,000 users). Our hybrid method offers many benefits, allowing locally observed findings to be verified, patterns in globally collected data to be explained and addresses ethical issues raised by the mass participation approach. We note trends in the local trial that did not appear in the larger scale deployment, and which would therefore have led to misleading results were the application trialled using 'traditional' methods alone. Based on this study and previous experience, we provide a set of guidelines to researchers working in this area.

Author Keywords
User trial methodology, mobile multiplayer games, mass participation, ad hoc peer-to-peer networking, MANETs

ACM Classification Keywords
H.5.2 [Information Interfaces And Presentation]: User Interfaces – Evaluation/methodology;

General Terms
Experimentation, Human Factors.

INTRODUCTION
Evaluations of the use (rather than the usability) of mobile and ubicomp systems have moved away from very controlled conditions in order to more closely align with the systems' context, as recommended by Abowd and Mynatt [1]. The earliest ubicomp systems [24,30] were largely confined to controlled laboratory conditions but there has been a general progress towards studies that take place in contexts more representative of the technologies' eventual intended use, with systems trials taking evaluation out of the lab [8], and moving towards studying participants' appropriation of technology in their everyday lives [3].

A recent step on this journey is the use of 'app store'-style repositories to allow researchers to release applications that participants can install directly onto their own handsets, thereby extending the reach of the research to potentially very large numbers of users. The benefits of such 'mass participation' deployments have been discussed [7,20], such as a reduced cost to researchers and the ability to reach users from vastly diverse geographical backgrounds. Although mainly positive in tone, several of these studies have mentioned shortcomings of this approach to research, such as the inability to meet users [20], a lack of standardisation of trial hardware [14] and the raising of additional ethical concerns [5].

In this paper, we survey these reported difficulties and expand upon them based on our own recent experiences of distributing mobile research applications in this manner. To counter some of these we propose a hybrid trial methodology, where the research application is released to the general public in tandem with a more traditional local deployment to recruited participants. 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 either group alone. We also argue that conducting app store-style trials using this hybrid methodology allows for more solid ethical practice.

To investigate our hybrid method, we created an application on the iOS platform and trialled it concurrently at both a local and global scale. The application, World Cup Predictor (WCP), 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. We study the findings of this hybrid trial, investigate whether these different methodologies can be effectively combined, explore productive ways of managing

CHI 2012, May 5-10, 2012, Austin, TX, USA.

the tasks best suited to each, and discuss whether this hybrid approach is sufficient to address the identified shortcomings of the 'mass participation' approach. We culminate with a set of guidelines to aid other researchers who are interested in performing such a study.

RELATED WORK

There has been a move towards studying the use of mobile software 'in the wild' [6]. Arguments have been made in favour of this change in methodology [25], with others further noting that field studies are better suited for studying broad issues surrounding use of technologies rather than merely the uncovering of usability issues [18]. The methods for selecting participants for trials have also been a topic of research. For example, the recruitment of a cohort of participants for long term use in trialling a number of systems in their everyday lives has been explored by Jay et al. [16], who state that "the main advantage of maintaining the cohort has been that we are able to reduce problems of generalisability associated with laboratory studies." The methodology outlined in this paper aims to go further with regard to reducing problems of generalisability, with a trial larger in scale and with much greater variation in participant age and location.

The mass participation approach of releasing mobile software via public 'app stores' has become more and more popular with researchers in recent years, as it offers a comparatively easy mechanism for recruiting large numbers of users while reducing hardware costs for researchers. Several research applications have had public releases on the Android [13,14,27], Blackberry [23] and iOS platforms [4,20]. The analysis performed in the evaluation of these systems has tended to be quantitative, taking advantage of the large numbers of participants to offer more solid numerical or statistical findings.

One of the most long-term deployments of mobile research applications released in this way is Cenceme [4], 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. Our work distinguishes itself from the Cenceme study by running concurrent large- and small-scale trials of the same application, and exploring both qualitative and quantitative data.

Hungry Yoshi was a game designed to expose the wireless infrastructure of a city, using this as a resource in the game. In earlier work, we reported that the application had over 130,000 downloads from around the world and describe how it is still possible to conduct qualitative analysis among such a widespread global user base [20]. We take a different approach here, combining those mass participation techniques with a small-scale local user trial so as to assess or enrich preliminary findings.

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 [9]. This approach has been applied in mass-released mobile applications [2], although without the methodological detail or discussion provided here.

The WCP application described in this paper 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 [17], yet few applications of MANETs have taken place outside of simulated environments [19], as a substantial density of users is needed to ensure successful transfer of data through a community via opportunistic, face-to-face encounters [17]. Previous trials have tended to be small-scale in terms of participant numbers.

DIFFICULTIES ENCOUNTERED IN 'APP STORE' TRIALS

As outlined in the previous section, the provision of 'app stores' on several mobile platforms in recent years has seen several research projects use these methods of deployment. Although most report positively on the benefits gained using this methodology, several potential shortcomings have been identified. For example Henze et al. conducted a trial using Android Market investigating the use of the mass participation approach in an experiment that sought to isolate 'cause and effect' on a single task [14]. 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.

Several other types of difficulty might arise in using this mass participation approach to Ubicomp trials. It was noted in the Hungry Yoshi study that although it was possible to gain qualitative data from remote users, the "*methods incurred expenses in terms of development time and interviewer effort"* [20]. In our own more recent trials, this seemed to get more challenging, and we had great difficulty in both finding participants willing to consent to speak to us 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.

Even 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 conducted at times best suited to the interviewees, arranged around 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 we found remote participants 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 whom 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 [27], 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 [15].

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 we were gathering very large amounts of quantitative data, we found that this reduction in the amount and quality of qualitative data meant it was harder to explain the patterns observed or infer users' motives: we could see *what* was happening in our trial, but had less success in discovering *why* it was happening.

Releasing via an app store is also so there is no simple means of only targeting a desired demographic [7]. This could be an issue if the research questions of interest concern, for example, only users of a particular profession, or group of users with a specific social structure.

The recruited users might also not be as diverse a group as it would first appear. For example, relative wealth scales play a part in the sampling process, with the level of entry to our trial set at having an Apple iPhone. It has been shown that cultural values change within a population with income level [12] and the homogenizing effect on cultural values between different populations of similar income levels has also been discussed [28]. 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.

A number of ethical considerations have been noted from this style of research [5], with certain concerns arising from trials making use of app store repositories that are 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 might be recorded on their usage of the software, their unique identifiers and their location? We have previously reported that 70% of surveyed users did not know that the application was part of a research trial [20], despite the presentation of a terms and conditions page on first launch. This echoes previous findings on the number of users reading terms presented in desktop applications [11]. Even if users did want to read this information, language barriers may prevent it; although we regularly provide translations 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.

A HYBRID APPROACH TOWARDS MOBILE SOFTWARE TRIALS

The hybrid approach presented in this paper 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 hybrid approach proposes releasing an application globally to attract a large user base and concurrently running a local trial, where it might be possible to record more rich data from participants. Data recorded from each user group can feed into and generate research questions for the other, in this way leveraging the strengths of both types of trial.

This approach would obviously not be suitable for every type of study. It is appropriate for applications running on commodity platforms which have public 'app store's. If trials involve specialised sensors or other nonstandard hardware, this method would not be feasible. Additionally, if researchers are only interested in particular types of relationships (for example parent-child), it might not be

practically possible to recruit a large number of participants with such connections by deploying to an app store. However, as will be described below, the hybrid method affords researchers the opportunity to handpick such particular social demographics for their local trial if desired, as well as seeing what will occur 'naturally' with a global release.

The following section introduces the WCP application and describes the user trial using the hybrid methodology.

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 we apply the techniques of the mass participation approach 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. We considered that releasing a research application based solely on testing MANET communications, an attempt at altruistic message passing through a number of intermediary nodes for example, would perhaps not attract huge numbers. Rather we attempted to create an appealing application, which could be used by an individual player in isolation and which we were confident would be 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.

We were interested in whether groups of socially-connected users would all download and use the application when it was released in this way, and whether we would observe similar or different uptake of MANET features among these users as compared to a local user group that was hand-picked to have known social links.

System description

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

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 leaderboard. Figure 1 shows screenshots from the application.

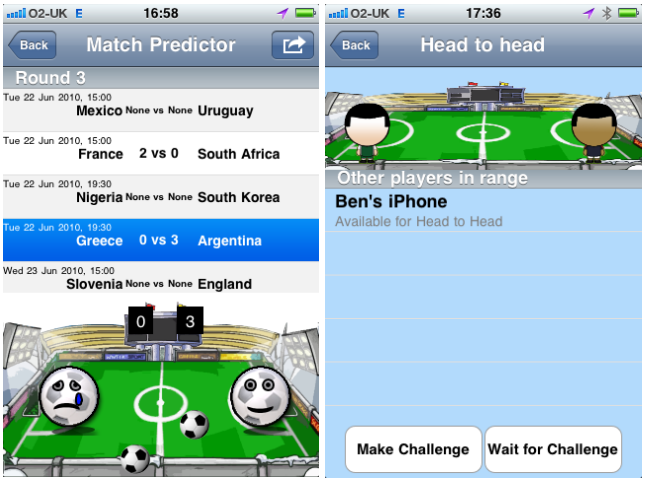

Figure 1. Screenshots from the WCP: the Match Predictor for the current round of matches (left) and the Head to Head page that allowed play via ad hoc networking (right).

As explained above, a purpose of the trial was to study the uptake of ad hoc networking functionality, so the application also included a peer-to-peer feature, 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 leaderboard. 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.

User trial

Although the WCP 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 methodology combining large and small-scale trials. As such, the WCP application was distributed to both locally recruited participants and via an iOS APT software repository [21] 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 users 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.

Local Participant Group

As we were particularly interested in investigating how social bonds between users impacted on game participation and use, we looked to recruit local users who had existing social ties. We were keen that this group should have regular day-to-day contact, to maximise the opportunities for use of the peer-to-peer functions. We also sought a smaller group of satellite players with no connection to the group of friends and colleagues. 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. In order to find a group with day-to-day social connections we asked volunteers 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 a 20 minute 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 but 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.

Global Participant Group

WCP 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 studies is the difficulty in quantifying the number of users involved in a trial [22]. In the local trial, we can identify a user as someone who was given the software, paid for participation and interviewed about their experience, and we can report the number of such participants with confidence. With an application released through an online software repository, this becomes more complicated as there are many possible ways to count users. Statistics provided from the 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.

Both the local and global deployments of the application logged participant usage data. The data logged includes activities within the game, such as moving between application screens, and general contextual information, such as location. Uploaded data was timestamped and stored on a database on a central server. To protect the privacy of participants TLS was used to encrypt data sent between phones and the server.

Before using the application, users were required to agree to the terms and conditions stating that their usage would be logged. A contact email address was also supplied for users to opt out of the trial at any time. This information was presented in four different languages. Users also had to agree via the standard iOS request for the application to use his/her location. The game was not 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 who provided locations played the game in Africa, the continent hosting the tournament, although there was activity recorded at five different World Cup stadia in South Africa.

On first launching the application, users were prompted to provide a username for use on the leaderboard and an email address so they could be contacted in the event of winning a prize. 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 we have to be aware that the reported answers cannot be verified. 80% of users provided an age, with most in their 30s or younger. The gender distribution was heavily male-biased, with 10297 male and 509 female.

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 requesting simple demographic information, we presented users with short questionnaires, allowing them to enter free text into a form within the application using the device's 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.

User Self-Selection Bias.

We note that in packaging the peer-to-peer functionality under scrutiny within a football predictor game we might have introduced a bias in the sample of users. For example, the vast majority of our 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, as opposed 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 who would be most likely to use peer-to-peer functions in applications more generally.

USAGE OF AD HOC NETWORKS

WCP was designed to test uptake of peer-to-peer networking. The game was completely playable without using ad hoc networks, but had this functionality presented as an option for 'head-to-head' play, where co-located players could swap predictions via Bluetooth for bonus points. The head-to-head mode was encouraged both through this 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 a further 5 points. This game mechanic was designed to provide motivation to use this feature. We were interested in whether we would see high uptake or whether, even with these incentives, participants would not be sufficiently motivated to overcome the obstacles to using this feature; namely, needing 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.

Our 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 co-located 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 eachother 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, compared to a global trial average of 0.01. 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*".

CONDUCTING HYBRID MASS PARTICIPATION TRIALS

Our primary goal in 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." [10] This section, drawing from our experiences of several app store style deployments and our 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 trial in a way in which these biases can be overcome.

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 with one group or the other. This goes some way to making a hybrid trial economically viable in research terms with, of course, the caveat that the software under evaluation must be able to run on standard hardware on a platform providing an 'app store'.

Commodity Hardware and Release-Quality Software

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 use their own devices, so an email with an app store link or a few minutes with the user and an Internet connection are enough to kit out a participant for the trial.

However, in general, software released in this fashion must be more polished and stable than more standard research prototypes [7], which obviously incurs greater expense in terms of implementation and testing. Although, this, in turn, greatly reduces the amount of technical support necessary during the trial – potentially producing less down time and therefore gathering more results. And where problems do occur the local participants are available to perform quick and accessible testing.

Numbers with ease; Interviews with ease

With the public distribution to a wide audience, there is less pressure to reach the 'magic number' of local participants 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, as described above.

More Interactive Design Cycle

One advantage of a hybrid trial method is in utilising the benefits of both groups of users as part of the software design cycle. Following release, developers may wish to evaluate current opinion, present new ideas, or probe the users for suggestions for desired features. This might begin by speaking with local users to hear current thoughts and future directions in which they would like to see the software develop. Having gathered a number of ideas from the local group, developers could then present some of these options to the global user base to vote upon to determine which suggestions are most popular.

Conversely, a mechanism can be provided to allow all users to submit suggestions for new features for the application. In our 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, and explore the subtle effects these modifications would make to the application.

In this way, the benefits of both groups of users are being exploited – the opportunity for in-depth discussions, and the ability to present design ideas to a 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. We use the two groups together, but utilise the strengths of each to maximise the benefits offered by this hybrid methodology.

Recommendations for Hybrid Trial Research

Here we discuss 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. We also discuss the different ethical responsibilities researchers have towards participants in each style of trial, and the restrictions they present. The discussion is formed around the following four recommendations, and we draw examples from our WCP trial to illustrate the points.

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 2. As can be seen, 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.

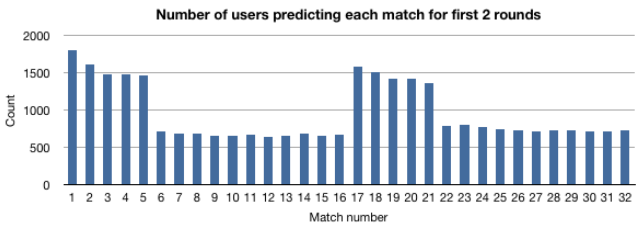

Figure 2. The number of users predicting each match

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 1, left) could be scrolled down. This solved our 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 solve this, 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 searching the web.

2. Use the Large to Verify the Small

As an inverse to guideline 1, the hybrid trial allowed us to use the global users to verify the generality of behaviour observed during detailed analysis of observations and interview transcripts from local participants. One of the most obvious advantages of running a large trial is the much larger number of users that can be expected, as compared to a more traditional local deployment. This offers the researcher increased confidence in making claims based on consistently observed behaviour, and, as shown below, reduces the risks of a researcher coming to incorrect conclusions if a local trial has non-representative users or the users are affected by experimenter effects.

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' [26]: subtle conscious or subconscious cues a researcher might give users that affect performance. We assert that 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 our trial of the hybrid method, we 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%. This disparity could be seen as simply reflecting the different sampling procedures used to create the two trial groups – one group specifically selected to be composed of socially connected users and another recruited via an app store with no forcing a desired social topology. Yet even ignoring the head-to-head aspects, the number of matches predicted in the game's single-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 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, we would have gained a very different impression of the features' popularity and therefore, it could be argued, gathered very misleading results.

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 WCP 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. McMillan et al. [20] reported that 70% of users had not understood that they were part of a trial. If users are unwilling to or, due to cognitive capacity or language skill, unable to read presented terms, it is infeasible to expect them to understand the possible consequences of logging or academic publication.

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. Our 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 we satisfy our ethical responsibility as evaluators, 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 participation 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, we approached our mass participation users differently than the local users. We engaged with the mass participation user group 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: we limited our direct interaction to non-compulsory survey questions, as opposed to requesting interviews, and examined aggregate logged data 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 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. harder to read a participant's reaction to a subject matter and stop if necessary. While we encourage researchers to continue to explore novel ways to meet their ethical responsibilities, we suggest that by using the full range of engagement possibilities with participants in this manner, researchers can feel more confident that they are pursuing an ethically sound research path.

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 use of the software we advise that 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, our results suggest that co-located social groups are not guaranteed to all adopt an application, even when strong incentives are given for using social features.

As shown, our results from the WCP trial indicate that the ad hoc network functionality was used to a reasonable degree within our 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 close by 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. It seems that handpicking participants matching a desired social topography is far more likely to lead to usage of features designed for a co-located group than relying on users acquired 'by chance' in a global release.

CONCLUSIONS

The use of 'app store's for research trials has become popular, yet researchers report several drawbacks to this approach. We show that the hybrid trial methodology presented here, combining use of a large-scale deployment with a local trial, can be a powerful tool, going some way to addressing these shortcomings. We designed an iOS application to test our proposed methodology, and studied it with a very large-scale trial. Based on these experiences of the hybrid method, we suggest that it offers a useful means to alleviate the weaknesses of both local and global trials.

For researchers conducting a local trial of mobile software, using our hybrid approach would:

- Allow findings based on consistently observed behaviour to be reported with greater confidence
- Mitigate problems resulting from 'outlier' users in a small sample leading to misleading conclusions
- Mitigate bias resulting from experimenter effects

For researchers who wish to run a mass participation-style trial, the hybrid approach will:

- 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

An important component of the hybrid method is a means for managing the ethical responsibilities 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. Again we emphasise that further work needs to be done in this area, for example assessing whether information available to participants after the trial has ended, e.g. on a web site, may serve to acceptably 'debrief'.

This study has also shown that regular usage of ad hoc networking in the WCP application only took place among those with pre-existing social ties and regular co-location. Despite offering generous prizes and a game design that encouraged peer-to-peer usage, there was very little uptake of this feature among the global users. We suggest that the

serendipitous use of such ad hoc networking technologies should not be expected within current applications. This also suggests that research questions relying upon any pre-existing social topology among co-located participants should be carried out with a selected local group.

Although this kind of hybrid methodology may have applicability in other application areas, we suggest that it is particularly worth exploring in future mobile software research. Local context, which is vital to mobile and ubiquitous computing 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.

REFERENCES

1. Abowd, G.D., Mynatt, E.D.: Charting past, present, and future research in ubiquitous computing. ACM Trans. Comput.-Hum. Interact. 7, 29-58 (2000)
2. Ames, M., Naaman, M.: Why We Tag: Motivations for Annotation in Mobile and Online Media. CHI (2007)
3. Bell, M and Chalmers, M., and Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown, B., Rowland, D., Benford, S., Capra, M., & Hampshire, A., Interweaving Mobile Games with Everyday Life, Proc. ACM CHI, pp. 417-426. (2006)
4. Campbell, A., Eisenman, S., Fodor, K., Lane, N., Lu, H., Miluzzo, E., Musolesi, M., Peterson, R., Zheng, X. Transforming the social networking experience with sensing presence from mobile phones. Proc ACM Conf on Embedded network sensor systems. 367-368 (2008)
5. Chalmers, M., McMillan, D., Morrison, A., Cramer, H., Rost, M., Mackay, W.: Ethics, Logs and Videotape: Ethics in Large Scale User Trials and User Generated Content. Ext. Abd, Proc CHI (2011)
6. Crabtree, A., Benford, S., Greenhalgh, C., Tennent, P., Chalmers, M., Brown, B.: Supporting ethnographic studies of ubiquitous computing in the wild. Proc conference on Designing Interactive systems. (2006)
7. Cramer, H., Rost, M. and Bentley, F. An Introduction to Research in the Large, Special Issue IJMHCI (2011).
8. Davies, N., Mitchell, K., Cheverst, K. and Blair, G. Developing A Context Sensitive Tourist Guide Proc. HCI for Mobile Devices (1998)
9. Denscombe, M. The good research guide. Maidenhead: Open University Press. (2003).
10. Denzin, N.K. *The Research Act,* 3rd edn. Englewood Cliffs, NJ: Prentice Hall. (1989)
11. FAST Federation Asks: Do you know what you're agreeing to? www.fastiis.org/resources/press/id/304/
12. Feather, N. T., Values and income level, Australian Journal of Psychology, 27(1), (1975)
13. Girardello, A. and Michahelles, F.: AppAware: which mobile applications are hot? : Proc. Int. conference on HCI with mobile devices & services. ACM, 431-434
14. Henze, N., Poppinga, B., Boll, S. Experiments in the Wild: Public Evaluation of Off-Screen Visualizations in the Android Market. Proc NordiCHI (2010).
15. James, N, Busher, H.: Credibility, authenticity and voice: dilemmas in online interviewing. Qualitative Research 6: 403-420 (2006).
16. Jay, T., Stanton Fraser, D.: The role of a cohort in the design and evaluation of pervasive systems. Proc. ACM Conf on Designing interactive systems. 31-39 (2008)
17. Khelil, A., Becker, C., Tian, J. and Rothermel, K.: An epidemic model for information diffusion in MANETs. Proc Int workshop on modeling analysis and simulation of wireless and mobile systems. ACM, (2002)
18. Kjeldskov, J., Skov, M., Als, B. and Høegh, R.: Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. MobileHCI, 529-535 (2004)
19. Kurkowski, S., Camp, T., Colagrosso, M.: MANET simulation studies: the incredibles. SIGMOBILE Mobile Computing and Comms Review 9 50-61 (2005)
20. McMillan, D., Morrison, A., Brown, O., Hall, M. and Chalmers, M.: Further into the Wild: Running Worldwide Trials of Mobile Systems. Pervasive Computing, Vol. 6030. 210-227 (2010)
21. McMillan, D., Morrison, A., Chalmers, M. Comparison of Distribution Channels for Large-Scale Deployments of iOS Applications, IJMHCI 3(4), 1-17 (2011)
22. Morrison. A., Reeves. S., McMillan, D., Chalmers M. Experiences of Mass Participation in Ubicomp Research. Research In The Large Workshop, ACM International Conference Ubiquitous Computing (2010).
23. Oliver, E. A survey of platforms for mobile networks research. SIGMOBILE (2009)
24. O'Shea, T., Lamming, M., Chalmers, M., Graube, N., Wellner, P. and Wiginton, G,.: Expectations and Perceptions of Ubiquitous Computing: Experiments with BirdDog, a Prototype Person Locator, Proc. BCS/IEE Conf on IT and People (1991)
25. Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W., Kurtz, A., Hall, R., Hursey, J., Toscos, T.: Why it's worth the hassle: the value of in-situ studies when designing Ubicomp. International conference on Ubiquitous computing. Springer-Verlag, 336-35 (2007)
26. Rosenthal, R. Experimenter effects in behavioral research. New York, Appleton-Century-Crofts, (1966)
27. Schleicher, R., Shirazi, A. S., Rohs, M., Kratz, S., & Schmidt, A. WorldCupinion Experiences with an Android App for Real-Time Opinion Sharing During Soccer World Cup Games. *IJMHCI, 3*(4), 18-35 (2011).
28. Strøm, G., Interaction Design for Countries with a Traditional Culture: A Comparative Study of Income Levels and Cultural Values. People And Computers XIX — The Bigger Picture, 2, 301-316 (2006)
29. Sturges, J. and Hanrahan,K.: Comparing Telephone and Face-to-Face Qualitative Interviewing: a Research Note. Qualitative Research April 4: 107-118 (2004)
30. Want, R., Hopper, A., Falcão, V. and Gibbons, J.:The active badge location system. ACM Trans. Inf. Syst. 10 (1992)