## Guest Editorial Preface

# Special Issue:
# An Introduction to Research in the Large

*Henriette Cramer, Mobile Life Centre at SICS, Sweden*

*Mattias Rost, Mobile Life Centre at SICS, Sweden*

*Frank Bentley, Motorola Mobility, USA*

## ABSTRACT

*Distribution of mobile applications has been greatly simplified by mobile app stores and markets. Both lone developers and large research and development teams can now relatively easily reach wide audiences. In addition, people's mobile phones can now run advanced applications and are equipped with sensors that used to be available only in custom research hardware. This provides researchers with a huge opportunity to gather research data from a large public. Evaluation and research methods have to be adapted to this new context. However, an overview of successful strategies and ways to overcome the methodological challenges inherent to wide deployment in a research context is not yet available. A workshop was organized on this topic and this special issue to help address these topics. This introduction provides an overview of strategies and opportunities in 'research in the large', while providing an introduction to challenges in ethics and validity as well.*

## INTRODUCTION

The increasing popularity of mobile smartphones, along with the convergence of these devices on a few key platforms are providing researchers with new ways to distribute mobile applications to larger audiences. By testing with a larger set of users, these deployments create the opportunity to receive more quantitative data about use as well as data from different locations around the world. This is in contrast to most mobile field trials over the past decade in which systems were deployed to a small group of (usually local) participants who would be loaned a target device for the duration of the study. It is hoped that these evaluations will allow for more naturalistic uses.

Today's mobile phones are much more capable than their predecessors, and because of this large numbers of people are able to run an ever increasing amount of applications. These applications can take advantage of rich

sensors on the device such as location, acceleration, ambient audio, imaging, and network access. Many systems that previously required distributing sensor packs and smart devices to participants can now be tested with the millions of people already owning a capable device all over the world.

Before the current app stores (the Apple App store and Google's Android Marketplace being the most successful), it was difficult to deploy a mobile system publicly. While a few researchers tried to release an app on their websites, often complex legal agreements had to be forged with mobile operators or phone manufacturers in order to obtain access to restricted APIs and build something interesting. This was beyond the grasp of most mobile HCI research teams. However, now both lone developers and large research and development teams can reach wide audiences relatively easily. The first steps towards wide deployment of research prototypes and gathering of data from large audiences are already being taken in various ways (e.g. Campbell & Choudhury, 2009; Good et al., 2007; McMillan et al., 2010; Kittur et al., 2008). Large-scale research can create huge opportunities, but also world-scale challenges. With large deployments come large amounts of (mostly quantitative) data. No serious discussion has occurred on the validity of this data compared to existing controlled research methods. There are not protocols, no standardization has occurred, and both the marketplaces and platforms are developing and changing rapidly. Nor is it clear whether when distribution on a large scale is possible, if it is really the best method to answer specific research questions.

Indeed, a selected number of researchers are taking advantage of the opportunities of wide deployments. In the fall of 2010, we organized a workshop at the 12th International Conference on Ubiquitous Computing on the topic of Research in the Large (Cramer et al., 2010). The workshop drew participants from industry and academia to address many of the issues that are raised in this special issue. We held discussions on methods for recruiting users, trusting data from a large deployment, analyzing demographics, and ethics for conducting research in this new way. These discussions led to the call for this special issue and the expanded articles that appear in the following pages.

Our workshop, and this special issue, has attempted to highlight the challenges that remain in scaling user studies from small evaluations to full-fledged deployment in the wild. Which research questions are best answered by large-scale research? Which app stores are better than others for distribution and how does device fragmentation restrict deployment? How does one promote their new system? What are the demographics of the users that sign on to these early-phase trial applications and how does this affect the validity of the results? Beyond practical questions, we hope to explore the ethical issues of large-scale research and matters of informed consent in research that might involve millions of participants. How can non-research participants that research teams will never meet understand the data that is being collected and make an educated decision about participating in the experiment? Understanding the issues involved and a deeper insight in these platforms is currently needed to make the most of this opportunity.

The articles in this special issue provide many examples of successful (and not successful) large-scale deployments of research systems. Through this discussion they address topics of ethics, data logging and analysis, mixed-methods studies that combine qualitative data with usage logs, and strategies for recruiting users. These topics are important for anyone attempting to launch a research project to a broader audience.

## WHY LARGE SCALE?

There are many research questions that require a large, statistically significant, population of users to answer. While in-lab studies and

small field trials can help to uncover usability concerns and to see a wide variety of behaviors in detail, large-scale studies are better suited at understanding precisely how a new application or service works for many people. Questions about typical use, network effects, and system adoption are best answered when users can choose to use a system when and how they choose in a more naturalistic manner. App stores and large-scale distribution provide for these conditions. Mobile applications can now also be used to collect large amounts of data from their users and their devices and can target users in specific locations of interest or provide a large overview of location-based data (e.g. Anderson et al., 2007; Chang et al., 2010).

Large-scale distribution can also be used to test the performance of an algorithm or system in a wide variety of contexts. With appropriate feedback to the research team when something does not work as expected, researchers can quickly learn scenarios that need additional focus. Small-scale studies cannot provide this large amount of data on successes and failures in thousands or millions of individual situations.

In addition to use of the system itself, large-scale deployments can help to answer larger questions about the user experience involving the wide ecosystem of network operators, device manufacturers, other users, developers and distribution and payment models, which all ultimately contribute to using a new application (McMillan et al., 2010). Factors such as steps needed in installing applications, paying for services, integration with other applications, media, and data on the phone and policies of service providers cannot all be considered in small, local studies.

The new distribution channels offer a great help in reaching the large and varied groups of end-users needed to gain such crucial insight. Not taking advantage of the available distribution channels and app stores would be missing an important opportunity. However, great community effort is required to overcome the new challenges that accompany the opportunities. And care needs to be taken in conducting the research such that the findings are trustworthy and helpful.

## CONDUCTING LARGE SCALE RESEARCH

Conducting a large-scale study brings many added complications to both development, and study design. Unlike a small-scale study where researchers are in the room or just a phone call away when issues arise, large-scale deployments require a much more robust system. Also issues of informed consent and demographics of participants can make or break a particular deployment.

### Developing for Distribution

When a person participates in a small-scale research project from a university or industrial lab, there are expectations that they hold about the early-stage of the system. Researchers can sit down with them personally and explain the current state of development and appropriately set expectations. However, large-scale deployments do not come with these one-on-one sessions and users typically expect something polished and (mostly) bug-free. This adds increased burdens on research teams as they are effectively developing a "product" to be released to large numbers of users.

This development then requires additional work in ensuring support for the wide variety of brands, device types, hardware limitations and platforms that users might own. Developers need to decide whether they make a choice for a specific platform or reaching a wider audience and supporting multiple. For instance, the experiences of Zhai et al. (2009) in bringing a mobile research prototype to the market via wide distribution, illustrate a number of challenges and opportunities that the choice for wide deployment via a particular platform (e.g. iPhone) bring. Later in this issue, McMillan et al (in press) explore two different deploy-

ment opportunities on the iPhone platform and the differing demographics of users in each. Releasing an app can also involve some time to obtain approval from the store itself. Sometimes, as in Zhai et al. (2009), this can only take a week, but less positive experiences are reported (Wired, 2009).

Platform choice can also imply specific design choices for the research team. Platforms differ in the ability for users to install and update applications or the particular platform functionality that is available. Being able to run threads in the background or access specific types of data might differ significantly between platforms. Understanding these differences and the security policies of each platform is a lot for almost any research team to be familiar with before even beginning to write the code. Miluzzo et al. (2008) provide a comparison between platforms on some of these issues, but such overviews are quickly out of date and keeping up requires quite some effort.

Beyond choice of platform, wide deployment introduces other deep technical issues. How should mobile applications be instrumented to better collect meaningful usage data? And how should this data be analyzed and used? How robust should the applications be in dealing with a wide variety of networks and different sensor hardware on different platforms? Upgrading an app to fix bugs or address user concerns is also something that is commonly done with large deployments, but depending on the severity of changes made, it might not be possible to directly compare usage data from users on multiple versions of the client. Calls to the server then need to be instrumented with client version to filter out data from users who have not upgraded. Many issues like this start to arise once an application is in 'the wild'.

While large-scale deployments often do not involve the typical costs of reimbursing participants for their time, they also come with their own infrastructure and personnel costs that are not present in small field or lab studies. With an increase of users potentially comes large server operating costs, costs of promotion in app stores, and additional tech support for users who may be experiencing issues. In addition to this is the cost of time to wade through large amounts of data over time to analyze and find patterns of behavior. This is much more work than looking at data from a dozen users over a few weeks in a traditional small scale field trial. How do we handle these costs? Releasing a research app may now require development of a business model as well.

## Getting Representative Users

Wide deployment enables collection of large amounts of data, but we need to make sure we actually get data that is representative and reliably answers our research questions. Relatively few ways are available to limit deployments when using open distribution channels, making it impossible to explicitly 'recruit' for a specific demographic. The choice for a specific platform might also mean reaching a specific type of people (to invoke some stereotypes: more wealthy or design-oriented users on the iPhone platforms, geeks for the Android market). The problems go deeper, as most mobile marketplaces do not share the demographics of their users with application developers, only aggregate counts. How can researchers ensure that their data is coming from an appropriate mix of users or analyze data by demographic to spot key differences in use? Asking for demographic data when a user first starts using a new app can be the difference between them continuing on or abandoning that app entirely. It is also not clear that user-provided demographic data can always be trusted as was discussed in our workshop.

Another aspect of attracting users relates to how the app is presented. In our workshop, we explored different research groups' experience with launching new systems as 'research prototypes', 'public betas', or just as another app in the marketplace. The way an app is described may greatly impact both the total count of users who download the app and their demographics.

Just putting an app 'out there' in an app store without any other promotion likely will not yield many users. How should applications be promoted and how does this affect research results? Researchers may use Facebook ads, tweets, and other means to introduce potential users to the system. The choice of keywords or demographics used to target these ads can serve as a way to recruit new populations to use the application, but might also start to bias the sample of participants in various ways. When analyzing data from these deployments it is important to consider the demographics of the sample. Raw usage data can often yield quite different results than data that is normalized based on demographics. This un-normalized data can give quite a different picture of application or feature use if not carefully accounted for through recruiting or filtering.

Putting apps out there also bring both chances and added complexities in dealing with exposure. Publicity can both increase an app's popularity, but can also affect usage and users' evaluation of a service. If an app gains great momentum, possible effects on developers' and researchers' time (and research focus) need to be considered. Research in addition involves risks and 'mistakes', which may have a negative impact on evaluations and reputations. Public feedback and media attention may impact the data we gather and may also affect future projects and releases. Strategies are needed to promote apps and to gain from user feedback, while also dealing with potential backlashes.

## Collecting the Right Data

Moving from research in the small, where voicemail diaries or other ethnographic-style means can often be enough to understand usage, to research in the large, where this type of data collection is impractical, it is necessary to build in some type of logging facilities in mobile applications. Often, researchers are interested in knowing how and when the application is used including the use of particular features or how long users spend in the application. In order to get this information, mobile applications need a type of instrumentation to capture interactions and upload them to a central server that the researchers can access. Many options can be experimented with in regards to different types of logging, including periodically capturing screen shots, logging every click, logging screen show/hide times, etc. No clear set of data has been established as the most useful and toolkits to log this type of data are still in early stages. The use of this data is critical to understanding how an application is being used in daily situations.

In addition to instrumentation, often it is desirable to get feedback from users in the form of short surveys. As mentioned above, demographic data is critical in understanding the use of an application that is deployed to many people, but collecting it reliably in a field deployment is still a challenge. Some researchers have introduced simple easy-to-answer questions in surveys asking for demographic data. If this answer is correct, it is more likely that the demographic data provided is also correct. Surveys can be displayed at regular intervals or after particular interactions in order to get more information from users. However, if the application is seen by users as a 'real' application and not as a research prototype that they are evaluating, prompting with surveys may not be seen as desirable and users may abandon the application entirely. Carefully probing for more information in a wide distribution is still an open research area.

Strategies on combining large deployments and triangulation with for example smaller ethnographic-style studies (Ames et al., 2007) will be necessary to increase data validity. In addition, we need strategies for gaining more reliable user feedback, and dealing with such feedback as it comes in. Different platforms provide different analytics tools and ways of gathering feedback; user comments in app stores being one such example. The problem from a research perspective however is how valid and reliable these comments actually are, how we should analyze them and whether we should react. By for example using app store comments for iterative development, we could

increase user satisfaction with our apps. But how do we then deal with updates and version management? When are we getting caught in a customer service cycle instead of focusing on our research questions?

## Ethics

The mechanics of executing research in the large are quite different from executing a small scale study. Public betas and other forms of large-scale research are often displayed to the public more like a product than like a research prototype. Because of this, many ethical issues about conducting the research and using data from users arise (Chalmers et al., 2011). Currently, researchers and institutional review boards are struggling to create guidelines for ensuring that participants in large-scale research studies remain informed about the purpose of the research, anticipated risks, and data collection/anonymity procedures.

When conducting research in the small, researchers can sit down with each participant and talk through a consent form. This form describes the study itself, any anticipated risks, that data will be kept anonymous but will also likely be published, and other important facts such as who to contact if the participant believes that something untoward is occurring in the research. Usually participants must explicitly agree to these terms and sign a form, often with a chance to opt in or out to the researchers' ability to use photos, sound recordings, videos, or other personal data from the study in presentations and research publications. By sitting down with participants and explaining each part of the form, they are more likely to think about the terms of the study and understand the implications of those terms. Researchers are also right there to answer any questionsHowever, when conducting research in the large, there is no opportunity to sit down with each participant. Often, the terms of the research and data collection are presented as fine print in an End User License Agreement (EULA) that may or may not be actively displayed as a part of the download and installation procedure of the new application. Even if actively displayed, research from Good et al. (2007) shows that users don't often understand the content of a EULA and that if the terms are actually explained to them, many users would un-install an application that they would have otherwise used. This raises serious ethical implications for research in the large, especially in countries or institutions that require explicit informed consent.

In addition to understanding risks and data collection procedures, participants should also know about the purpose of the research. But often research in the large does not explain this to the user, and they might think that they are just using a 'normal' commercial product like most applications in an app store. Do they understand how their use is being monitored? Do they understand that communication they might think is 'private' through the system might end up publicly displayed in a research paper or presentation? These are all important ethical considerations to consider before launching a research project as a public beta and something that institutional review boards and prudent researchers should ponder.

In some ways, research in the large can be seen as analytics on a large system, much like what Google or Facebook do with the large amounts of data that they receive. The Facebook Data team often publishes interesting findings gleaned from analyzing usage logs, such as stats on the recent midterm elections in the United States (Chang et al., 2010). Is mobile research in the large somehow different from this type of research? As more academic-minded researchers, do we have an obligation to ensure that our users are better informed about data collection and research purposes? Even if clearly disclosing terms of the study results in fewer users signing up for the service or downloading the application, it can be considered our responsibility as researchers to ensure that all of our participants are fully informed. In addition, research on a worldwide scale also means having to deal

with international and intercultural differences in terms of regulations, practices and norms (Henderson & Ben Abdesslem, 2009).

## THIS SPECIAL ISSUE

The papers in this issue fall into three main themes: distribution, data analysis, and validity with some papers providing contributions across the themes. McMillan et al. (in press) provide an analysis of two ways to distribute an iOS application: through the official Apple App Store or through third party repositories. This analysis highlights decision criteria that research teams should analyze if deciding to deploy a new application for the iOS platform. Schleicher et al. (in press) present their experiences with releasing a research application on the Android Market. They highlight important issues that researchers must take into account when conducting research in the large, in addition to giving accounts to what they learned about their application, WorldCupinion.

On the data-analysis side, Morrison and Chalmers (in press) describe a tool to visualize real-time results from a large-scale deployment as well as a way to see patterns of use across individual users. This system shows great promise in identifying patterns and combining more qualitative methods with the quantitative data that is being collected.

Finally on validity of results, Coulton and Bamford (in press) give accounts of a longitudinal study of how app stores behave with experiments including two apps with more than 1.5M downloads. They show how the number of new downloads are affected by events like updates, changes in presentation, etc. Henze et al. (in press), discuss five large-scale deployments that they have been involved in and discuss practical details of what has and hasn't worked in order to attract users and receive meaningful data when using an app as an experimental 'apparatus'. They discuss differences in use observed in an 'in the wild' evaluation versus a more controlled field experiment as well as the ethics of conducting research with unknown participants.

We hope that these articles will help to form a beginning of a community around large-scale deployments and that the lessons from these authors can be taken forward to help improve the quality of future work in this area. We are hopeful that as more research is completed in this domain, best practices and updated methods will emerge that can more reliably lead to valid and repeatable results.

## ACKNOWLEDGMENTS

We would like to thank all participants in the 'Research in the Large' workshop at UbiComp 2010 for the inspirational presentations and discussion, our co-organizers Didier Chincholle and Nicolas Belloni, as well as the program committee and reviewers for this special issue for their helpful comments. Henriette Cramer's work was partly carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

## REFERENCES

Ames, M., & Naaman, M. (2007). Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 971-980). New York, NY: ACM Press.

Anderson, I., Maitland, J., Sherwood, S., Barkhuus, L., Chalmers, M., Hall, M. et al. (2007). Shakra: Tracking and sharing daily activity levels with unaugmented mobile phones. *Mobile Networks and Applications*, *12*(2-3), 185-199.

Campbell, A., & Choudhury, T. (2009). Toward societal scale sensing using mobile phones. In *Proceedings of the NSF Workshop on Future Directions in Network Sensing Systems*.

Chalmers, M., McMillan, D., Morrison, A., Cramer, H., Rost, M., & Mackay, W. (2011) Ethics, logs and videotape: Ethics in large scale user trials and user

generated content. In *Proceedings of Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press.

Chang, J., Bonta, J., Qian, F., Shrenk, N., Li, D., & Conner, A. (2010). *How voters turned out to Facebook.* Retrieved from http://www.facebook.com/data#!/notes/facebook-data-team/how-voters-turned-out-on-facebook/451788333858

Coulton, P., & Bamford, W. (in press). Experimenting through mobile 'apps' and 'app stores'. *International Journal of Mobile Human Computer Interaction*.

Cramer, H., Rost, M., Belloni, N., Chincholle, D., & Bentley, F. (2010) Research in the large: Using app stores, markets and other wide distribution channels in UbiComp research. In *Extended Abstracts of the ACM International Conference on Ubiquitous Computing.* New York, NY: ACM Press.

Good, S. N., Grossklags, J., Mulliga, K. D., & Konstan, J. A. (2007). Noticing notice: A large-scale experiment on the timing of software license agreements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 607-616). New York, NY: ACM Press.

Henderson, T., & Ben Abdesslem, F. (2009). Scaling measurement experiments to planet-scale: Ethical, regulatory and cultural considerations. In *Proceedings of the 1st ACM International Workshop on Hot Topics of Planet-Scale Mobility Measurements* (pp. 1-5). New York, NY: ACM Press.

Henze, N., Pielot, M., Poppinga, B., Schinke, T., & Boll, S. (in press). My app is an experiment: Experience from user studies in mobile app stores. *International Journal of Mobile Human Computer Interaction*.

Kittur, A., Chi, E., & Suh, B. (2008). Crowdsourcing user studies with mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 453-456). New York, NY: ACM Press.

McMillan, D., Morrison, A., Brown, O., Hall, & Chalmers, M. (2010). Further into the wild: Running worldwide trials of mobile systems. In *Proceedings of the International Conference on Pervasive Computing* (pp. 210-227).

McMillan, D., Morrison, D., & Chalmers, M. (in press). A comparison of distribution channels for large-scale deployments of iOS applications. *International Journal of Mobile Human Computer Interaction*.

Miluzzo, E., Oakley, J., Lu, H., Lane, N. D., Peterson, R. A., & Campbell A. T. (2008, November 4). Evaluating the iPhone as a mobile platform for people-centric sensing applications. In *Proceedings of the International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems*, Raleigh, NC.

Morrison, A., & Chalmers, M. (in press). SGVis: Analysis of data from mass participation Ubicomp trials. *International Journal of Mobile Human Computer Interaction*.

Schleicher, R., Sahami Shirazi, A., Rohs, M., Kratz, S., & Schmidt, A. (2011). WorldCupinion: Experiences with an android app for real-time opinion sharing during soccer world cup games. *International Journal of Mobile Human Computer Interaction*.

Wired. (2009). *Apple's appalling approach to iPhone app approvals.* Retrieved from http://www.wired.com/underwire/2009/08/alt-text-apples-appalling-approach-to-iphone-app-approvals/

Zhai, S., Kristensson, P. O., Gong, P., Greiner, M., Peng, S. A., Liu, M. L., & Dunnigan, A. (2009). 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* (pp. 2667-2670). New York, NY: ACM Press.

*Henriette Cramer is a researcher at SICS and the Mobile Life Centre in Stockholm, Sweden. Her research includes location-based services, research through wide distribution of apps, using existing services and user-generated data for research purposes, and people's interaction with autonomous and adaptive systems. She was co-organizer of the 'Research in the Large' workshop on wide distribution of (mobile) research apps at UbiComp 2010 associated with this special issue, and its follow-up at UbiComp 2011, as well as a workshop at CHI 2011 on ethics in large scale trials.*

*Mattias Rost is a PhD student at Mobile Life and SICS. His interests are location based services, analyzing larger data sets generated by users of such services, and presenting local content in interesting ways, mobile mashups, and wide distribution of research apps. As developer of 'the best bad app ever', app store user comments are a special interest. He coorganized the Ubicomp 2010 'Research in the Large' workshop, and the followup the year after at UbiComp 2011. He is also the co-organizer of the CHI'11 workshop on ethics in large scale trials.*

*Frank Bentley is a Principal Staff Research Scientist at the Motorola Applied Research Center and co-teaches a class at MIT called Communicating with Mobile Technology. He's interested in strengthening strong-tie social relationships at a distance using mobile devices and in using ethnography and field deployments to understand how new technology is integrated into peoples lives. He was a co-organizer of the Ubicomp 2010 and 2011 'Research in the Large' workshops.*