My App is an Experiment: Experience from User Studies in Mobile App Stores

Niels Henze
University of Oldenburg
Oldenburg, Germany
niels.henze@
uni-oldenburg.de

Martin Pielot
OFFIS - Institute for
Information Technology
Oldenburg, Germany
martin.pielot@offis.de

Benjamin Poppinga
OFFIS - Institute for
Information Technology
Oldenburg, Germany
benjamin.poppinga@offis.de

Torben Schinke
Worldiety GbR
Oldenburg, Germany
torben.schinke@worldiety.com

Susanne Boll
University of Oldenburg
Oldenburg, Germany
susanne.boll@
uni-oldenburg.de

ABSTRACT

Experiments are a corner stone of HCI research. Mobile distribution channels such as Apple's App Store and Google's Android Market have created the opportunity to bring experiments to the end user. Hardly any experience exists how to conduct such experiments successfully. This article reports about five experiments that we conducted by publishing Apps in the Android Market. The Apps are freely available and have been installed more than 30,000 times. The outcomes of the experiments range from failure to valuable insights. Based on these outcomes we identified factors that account for the success of experiments using mobile application stores. When generalizing findings it must be considered that smartphone users are a non-representative sample of the world's population. Most participants can be obtained by informing users about the study when the App had been started for the first time. Because Apps are often used for a short time only, data should be collected as early as possible. To collect valuable qualitative feedback other channels than user comments and email have to be used. Finally, the interpretation of collected data has to consider unpredicted usage patterns to provide valid conclusions.

Author Keywords

App Store, study, field study, in the wild, mobile application store, experiment, observation, apparatus, Android Market

INTRODUCTION

Mobile application stores such as Apple's App Store and Google's Android Market revolutionized the distribution of applications for mobile devices. This distribution channel lowered the gateway hurdle dramatically and opened the market for small companies and engaged hobbyists. Mobile application stores – for the first time – enable virtually any developer to easily reach hundred thousands of mobile users. Recently researchers discovered this opportunity and began to publish prototypes via mobile application stores.

It has been argued that the "easy access to such a potentially wide audience could radically alter the nature of many UbiComp trials" [Morrison et al., 2010]. In the tradition of UbiComp research most attempts to distribute prototypes via mobile application stores focus on the evaluation of prototypes (e.g. [Zhai et al., 2009, Girardello, 2010, Michahelles, 2010, Gilbertson et al., 2008]). Proof-of-concept prototypes are developed and the large number of users is used to demonstrate the successfulness of the respective application. Feedback is mainly gathered to understand the nature of the respective prototype.

In the tradition of psychology and social sciences Human Factors and Human-Computer Interaction research in contrast focus on understanding the human. Commonly, controlled experiments, quasi-experiments and observations are used to derive general findings. As in psychology, prototypes are often just the apparatus to investigate a research question. The psychologist Danziger describes an apparatus as a tool for "exposing experimental subjects to controlled and precisely known forms of stimulation" and "for recording and measuring responses" [Danziger and Ballantyne, 1997]. In previous work we showed that Apps distributed to thousands of users can successfully be used as an apparatus for controlled experiments [Henze and Boll, 2010, Henze et al., 2010].

In this paper we report our findings from five studies we conducted by publishing Apps in the Android Market. The paper first presents these Apps, the research questions they address, and the outcomes. In the subsequent sections we then discuss our general findings and conclusions on the participants, the quantitative and qualitative data, and ethical aspects. We conclude with aspects that should be considered when conducting experiments in mobile application stores.

Pre-print of the paper "My App is an Experiment: Experience from User Studies in Mobile App Stores" accepted by the International Journal of Mobile Human Computer Interaction (IJMHCI) – special issue on Research in the Large.

CASE STUDIES

In order to investigate different mobile HCI topics we conducted five studies, which actually use an App as apparatus and were published via mobile application stores. All Apps have been implemented for the Android platform and are therefore available for a range of users and devices. Table 1 gives an overview of the studies that are described in the following.

name	installs	samples	time	type
SINLA	≈1737	8	8.5 mo.	quasi-exp
PocketNavigator	9,149	670	6 mo.	quasi-exp.
MapExplorer	6,372	4,197	6 mo.	experiment
Poke the Rabbit	5,708	5,103	5.5 mo.	experiment
Tap It	7,811	6,907	2 mo.	observation

Table 1. Overview about the five studies we conducted.

SINLA: Off-screen visualizations for augmented reality

In Augmented Reality the visualization of nearby points of interest (POIs) is commonly done by displaying a small mini-map to provide an overview as the user moves around. However the 3D augmented environment and the 2D mini-map have different reference systems. Therefore, interpreting the mini-map and align it with the augmented environment demands special mental effort. A number of techniques have been developed for digital maps to visualize off-screen objects that are currently beyond the screen [Zellweger et al., 2003, Baudisch and Rosenholtz, 2003, Burigat et al., 2006]. We adapted an existing arrow-based technique for visualizing off-screen handheld Augmented Reality. In a lab study we compared this technique with a state-of-the-art mini-map [Schinke et al., 2010]. Based on our findings we included three off-screen visualizations, a mini-map, 3d arrows and a combination of a minimap and 3d arrows in our prototype (shown in Figure 1) resulting in three conditions. The aim of the study was to validate whether users have age or gender specific preferences for a visualization technique.

Figure 1. SINLA screenshots of the two visualization techniques arrows and mini-map. The third condition is the combination of arrows and mini-map.

First of all the prototype is a simple handheld Augmented Reality application that displays nearby POIs as blue balls located at the appropriate virtual position on a live camera viewfinder. To make the prototype (shown in Figure 2) useful for real users we created a function to search for nearby POIs. The user can filter the results by searching for keywords and selecting a maximum search radius. As soon as POIs are in range the user becomes supported by displaying the off-screen graphics.

Every user is considered as a potential participant, thus all

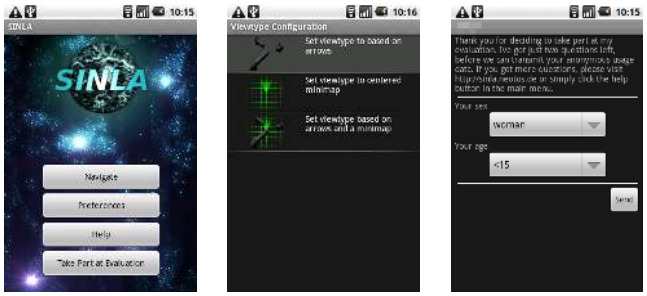

Figure 2. SINLA's menu, preferences, and questionaire.

relevant data is recorded locally without any remark. After five minutes of total usage a message appears that the user can take part in a study and as a reminder a new button is shown in the main menu. If the user took part the button disappears but after another five minutes of use the button reappears and the user can take part again.

Using SINLA we want to collect long-term usage data based upon a quasi-experiment as the study design. We continuously measure the time, each visualization is used. To avoid a systematic influence the initial visualization is chosen randomly on the first start up but the user can freely select another one. If the user decides to take part, a short questionnaire is shown to select a gender and the age from six categories. This data and the measured times for each visualization, in addition to a unique device hash and the current date, are submitted to our server.

The application was first published without any logging functionality containing earlier drafts of our off-screen visualizations in the Android Market on August 30, 2009. It has been updated on November 14, 2009 to provide the described features for the study and was accessible until September 29, 2010. Despite the fact that the application was available in the market for over one year, only 2,853 users installed one of the two versions and 1,737 of them were able to participate using the updated prototype. We collected only 8 samples, which made it impossible to derive any conclusions. Also no users took part more than once. Revising our study design it is likely that the attempts to get the user's attention for participating in the study are not obtrusively enough. The prototype probably also does not motivate long term usage. Another conclusion, taking the Android Market comments into account, is that our prototype does not work well on different android devices and platforms.

PocketNavigator: Conveying geographic cues with tactile feedback

More and more often, we can find pedestrian navigation systems integrated in modern smartphones. Inspired by the established car navigation systems they are able to show a user's location on a map and highlight a route to a destination. Some of these applications provide turn-by-turn instructions through text, visualisations or speech output. However, in some situations visual or audio feedback is not

appreciated by a pedestrian. Tactile feedback as navigation aid has been proposed and studied by several groups, e.g. [van Erp et al., 2005, Pielot and Boll, 2010]. However, existing studies mostly focus on artificial settings and tasks. The question that remains unanswered is if tactile navigation feedback works in non-artificial everyday situations.

At a first glance, the PocketNavigator (see Figure 3) is an ordinary pedestrian navigation application. A scrollable map, like e.g. available in Google Maps, is shown and the user's location is displayed. Furthermore, a waypoint-based shortest route can be calculated for any destination. The route is displayed as an overlay on the map, but is also shown as a visual arrow pointing towards the next waypoint, using the device's integrated compass. In addition, the PocketNavigator provides tactile patterns, conceptually showing towards the next waypoint (see [Pielot et al., 2010]). The aim of the study was to analyse how the tactile feedback will be used in the wild and how it affects the navigation performance (e.g., navigation errors, disorientations). For the PocketNavigator we tried different techniques to ask for permission to log data. In early versions of the application an opt-in checkbox in the about/tutorial view is used. In later versions the logging is also advertised through re-appearing popups, asking for participation in the study.

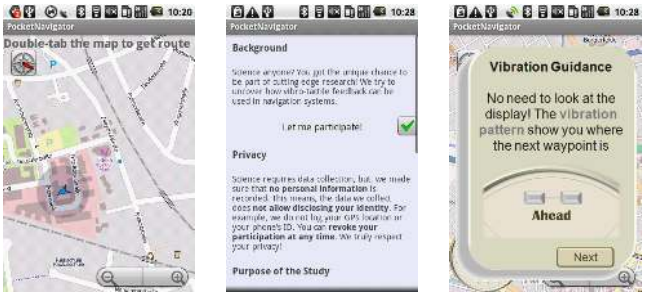

Figure 3. Screenshots of TavtileNavigator's main view, about screen, and information about the tactile feedback.

For the application we are interested in long-term everyday use results. Thus, no artificial data collection task has been integrated, but the participants are expected to use the application when they actually need navigation assistance. The tactile feedback serves as independent variable with the conditions *on* and *off*. The PocketNavigator uses a quasi-experiment as study design, as every participant is free to turn the tactile feedback on and off as desired. Important settings and configurations are logged into a file every second and are transmitted to a server every 120 seconds. In detail we log: touch interactions, speed, device orientation, compass angle, loudness, light level, and the current configuration and state of the application. We do not log any personal or private data (e.g., phonebook, SMS, user location). Some of the logged parameters are based on frequently occurring data (e.g. acceleration), which we do not log as raw data. Of particular importance are the logged navigation errors, disorientations, and the device posture, which we identified as an important measure for our results. Each of these high-level values is based on a combination of basic values.

The application is available in the Android Market since April 15, 2010 and has been installed 9,149 times. 670 participants agreed to participate in the user study. However, most of them did not navigate in the foreseen way or not at all. Most of the users tried and tested the application once. If the application has been started multiple times, reoccurring use-cases were e.g. driving in a vehicle or watching the map. 19 participants fulfilled the criteria of seriously navigating as pedestrian at least twice. On average these participants used 2.47 routes. They spent 390.79 seconds per route on average. In 93.90% the participants followed the route presented by the navigation system. The tactile feedback has been used by 8 users on 9 different routes. Having collected such a small amount of data only shows that conducting experiments via mobile application stores does not automatically yield in a large data set. Here, we believe, the reasons were that people do not navigate as often as they e.g. play games and that our method of asking for the users' consent was very conservative.

MapExplorer: Comparing off-screen visualizations with a tutorial

The aim of this study was to compare different off-screen visualization techniques for digital maps (e.g. Halos [Baudisch and Rosenholtz, 2003] and arrows [Burigat et al., 2006]. Previous work conducted studies with static maps and did not consider tasks where users can dynamically interact with the map by panning it [Baudisch and Rosenholtz, 2003, Burigat et al., 2006]. Furthermore, the conclusions are based on studies conducted in a lab with few participants that share similar backgrounds. To compare the three previously studied [Burigat et al., 2006] off-screen visualizations shown in Figure 4 we implemented a location based App.

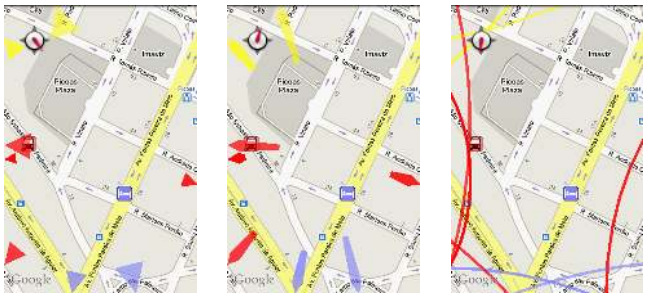

Figure 4. Screenshots of the three visualization techniques Halo, stretched arrows, and scaled arrows used by the MapExplorer.

To collect usage data as early as possible and to be able to compare the visualizations we decided for a tutorial which mimics a well defined task similar to the tasks in lab experiments. Using one defined task should improve the repeatability and reduce the effect of other influences. The tutorial appears when the application starts for the first time. After an introduction users should execute a simple find-and-select task using each visualization. While executing the task a map containing 10 randomly distributed POIs is shown. The tutorial's task is to select the red rabbit. The map can be explored by panning it with the finger. A POI is selected by tapping on it. After completing the tutorial the application

offers the standard functionalities of a location-based application. Users can search for nearby POIs and access details about them. When starting the MapExplorer for the first time the user learns that the App logs data (see Figure 5). The user can opt-out by deselecting a preselected checkbox.

The App's tutorial is designed as an experiment with repeated measures. The off-screen visualization technique is the independent variable resulting in three conditions. The order of the conditions is randomized to reduce sequence effects. While the user executes the tutorial the task completion time, the number of map shifts, and the number of errors are logged. After finishing the tutorial the time spend with each off-screen visualization is measured and we also measure if the user interacts with the application or not. Furthermore, users can fill the feedback form shown in Figure 5. In addition, we collected the user's time zone and the selected locale (e.g. en_US or de_DE).

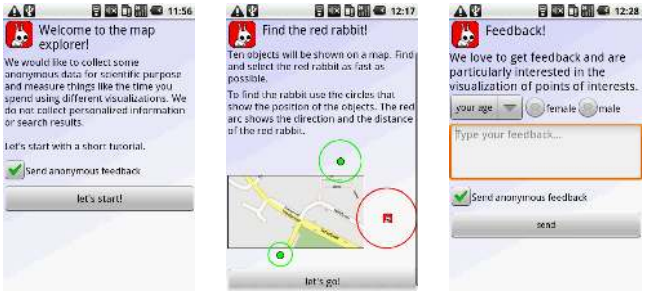

Figure 5. Screenshots of the MapExplorer's introduction, tutorial instruction, and the feedback form.

We published the application in the Android Market on April 1, 2010. Till September 29, 2010 the App was installed 7,664 times and we collected data from 4,197 users. Analysing the data we found that users need significantly more time and map-shifts with Halos ($p < .001$) to complete the tutorial (further details can be found in [Henze and Boll, 2010]). However, investigating the data in more detail shows that a number of users needed much more time to complete the tutorial than one would expect (e.g. longest time spend using Halos was 100 seconds). Reconsidering our design it might be assumed that instead of measuring the pure task completion time the results are affected by the "interestingness" of the visualizations. From informal tests we can report that some users explore the map much longer using Halos than using the other visualizations. Furthermore, our results are limited because users had no previous training and most users performed the tasks only once with each visualization.

Poke the Rabbit: Evaluating Off-Screen visualizations with a game

Based on the ambiguous results from the tutorial-based approach we decided to repeat the evaluation of off-screen visualization techniques with a different apparatus. The main shortcoming of the previous experiment is that it is not clear if users' tried to accomplish the task in an efficient and effective way. Therefore, we designed a game using the same three visualization techniques (see Figure 6) as conditions. Compared to a tutorial a game has the advantage that it is natural to confront players with variations of the same task.

Figure 6. In-game screenshots of the three visualization techniques Halos, stretched arrows, and scaled arrows from Poke the Rabbit.

Before starting the game a short introduction explains how to play. The game starts with a stage of three levels each containing 30 objects, represented by "cute" rabbit icons. The objects are randomly distributed on a plane (much larger than the actual screen size) that can be paned much like a digital map. Each level uses a different off-screen visualization (see Figure 6). The task of the player is to "poke" as many objects as possible by tapping them with the finger in a certain time frame. Once an object is poked it fades to gray and a new object appears. If a player finishes the three levels he or she goes to the next stage where 20 objects are used and afterwards to a stage with 10 objects. The visualizations are randomized within a stage to reduce sequence effects. After finishing three stages the game starts from the beginning with more time to complete a level but also with more objects needed to successfully finish a level. All players are directly considered as participants. The user is never informed that the App is the apparatus of a study and that usage data is transmitted to our server.

Poke the Rabbit uses the same study design as the MapExplorer's tutorial. The study is an experiment with repeated measures and task repetition. The off-screen visualization technique is the independent variable resulting in three conditions. The order of the conditions is randomized to reduce sequence effects. We recorded the number of poked rabbits for each level played. In addition, we collected the user's time zone and the selected locale.

We published the game in the Android Market on April 14, 2010. Till September 29, 2010 the game was installed 6,098 times and we collected data from 5,103 devices. We found that the performance of the off-screen visualizations depends on the number of used objects (see [Henze et al., 2010] for more details). For 20 and 30 objects the arrow-based approaches significantly outperform Halos. For 10 objects, however, Halos outperforms both arrow-based techniques. We also found that the device has an effect on the players performance (e.g. using the Motorola Sholes results in 13% higher performance than using a HTC Hero $p < 10^{-9}$). With the vast amount or data in our hands we assumed that we will be able to also analyze learning effects. Because of the multiple varied variables (e.g. duration of a level, num-

ber of objects, and required performance to advance to the next level) and the players' uncontrolled behaviour (players that perform badly might quit playing soon) we were, however, not able to analyze learning effects.

Tap It: Assessing users' touch performance

Following Poke the Rabbit, we conducted another study with the aim to investigate the touch behaviour of smartphone users. With Tap It we want to assess the touch performance for different target sizes similar to the work by Park et al. [Park et al., 2008]. By collecting a huge amount of data we aim not only at determining the error rate and reaction time for different screen locations and target sizes. More, we want to derive a model for predicting the users' performance that takes the touch history into account.

After starting the game, the player has to touch appearing white rectangles before they are fully visible (see Figure 7). Different patterns of rectangles appear from a single rectangle, over a number of connected or randomly distributed rectangles, to the whole screen filled with rectangles. If the user touches a rectangle it disappears and points are added to the user's score depending on his/her speed. As soon as a pattern is completed the next pattern appears. After completing a number of patterns the player is rewarded with a "badge" and advances to the next level with smaller rectangles. After four levels the player advances to the next theme and basically repeats the same procedure at a higher speed and an increased number of patterns. We implemented a global and a local high score list as well as the badges to increase the players' motivation. Players are informed that they will take part in a study when the App is started for the very first time. It is not possible to opt-out without to not play the game or turn off the internet connection.

Figure 7. Screenshots of Tap It. The images on the left and in the centre show in-game screenshots with two and four appearing rectangles and the right image shows the global high score list.

This study is a controlled observation with a defined task steered by an apparatus. It is intended to record the users' behaviour to derive a model that predicts the touch performance. While playing, the levels including the appearing rectangles are logged. To assess the users' touch behaviour each touch is recorded together with a timestamp and the state of the visible rectangles. We also collected the user's time zone and the selected locale.

Tap It has been published to the Android Market on July 31, 2010. Till September 29, 2010 the game was installed 8,495 times and we collected data from 6,907 devices resulting in 7,284,263 touch contacts. Not surprisingly we found significant differences between players that use different devices. We also found that the position of the preceding rectangle affects the touched position for the following rectangle. The average touch distribution is skewed towards the previous target and the position of the preceding rectangle affects the error rate as well as the users' speed.

DISTRIBUTION OF USERS

To derive general conclusions from user studies that are globally applicable it can be argued that the sample of persons must reflect the whole population. A number of studies investigated the importance of the diversity of participants. Evers and Day, for example, analyzed the role of culture in interface acceptance [Evers and Day, 1997]. They showed that fundamental differences exist between cultures regarding the user interface beyond obvious factors, such as, language and characters. Another example is Simon who showed that gender and culture has an important effect on the perception of web sites [Simon, 2000]. Young reviewed the literature that addresses the integration of culture in the design process and points out that "there is room for improvement" [Young, 2008]. Our initial expectation [Henze and Boll, 2010, Henze et al., 2010] (and we are not the only ones [Korn, 2010, Morrison and Chalmers, 2010]) was that by deploying Apps via a mobile application store we will get access to a worldwide audience. This would enable to derive conclusions that are not only applicable to a particular country or culture but to the global population.

Analyzing our data we found that the participants are less diverse than we expected. Figure 8 shows which locale the prototypes' users use. The results are quite consistent over the different prototypes (we only collected the user's locale for three of the prototypes). With 63%-79% English is by far the most common language followed by German (5.42%-7.45%), French (2.72%-5.29%), and Spanish (1.62%-6.04%). The most common non western languages are Chinese (1.93%-3.44%) and Korean (0.32%-2.84%). In general western languages accounted for more than 90% of the results for the three prototypes. This is consistent with the time-zone that users use (see Figure 9). 85.89% of all users have an American or European time zone.

Market research shows that 66% of the Android users are in the United States [AdMob, 2010b]. Contrary to our data the report from May 2010 says that 13% of the Android users are in China (compared to 1.93%-3.44% for our prototypes). There are a number of potential reasons for this divergence (e.g. our Apps address a particular audience). We, however, assume that the main reason is that we did not translate any of the Apps into Chinese. On the other hand, we internationalized the MapExplorer to German without a noticeable effect. Further looking at market research we see that in January 2010 Android users were mostly male (73%) while other platforms had an almost equal gender split (iPhone: 57% male and webOS: 58% male) [AdMob, 2010a]. Market research from Nielsen [Kellogg, 2010] analyzed Smartphone users in the United States and it can be seen that this sample

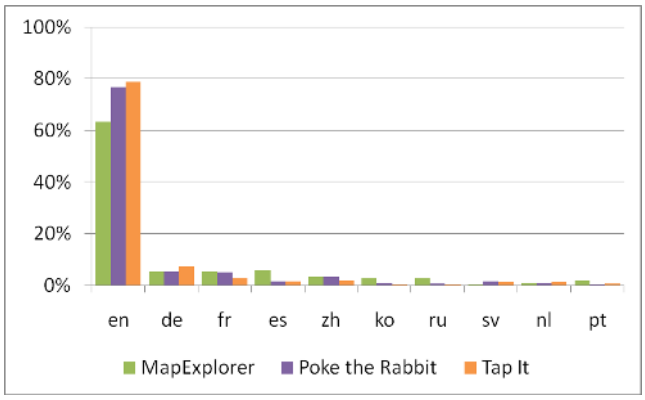

Figure 8. Percentage of users with the respective locales for three of the prototypes (n=15739).

does not even reflect the US population at all. In particular, the average Android users as well as smartphone users in general have a considerably higher income than the average US citizen.

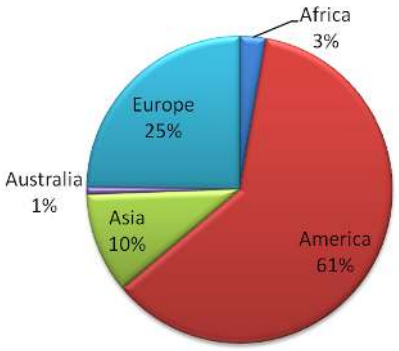

Figure 9. Fraction of users from different continents based on the respective time zone (n=15608).

Looking at market research but also by looking at our data, it can be concluded that general claims about participants' diversity are simply misleading. Smartphone users, and in particular Android users, are obviously far from being a perfect sample of the global population. As users have to actively install the respective apparatus themselves this further shifts the sample towards more tech-savvy people. General conclusions about the global population cannot be derived. However, depending on the aims of a study the market population might be more relevant than the overall global population.

COLLECTED DATA

When using Apps as an apparatus it is necessary to collect data about the users' performance and behaviour. Compared to lab studies, researchers cannot directly influence the users' behaviour. It cannot be ensured that users perform the tasks as often as desired and in the intended way. Furthermore, in contrast to most lab studies no personalized data is collected (at least this is the case for the Apps discussed in this paper). It might therefore not even be necessary to ask users to fill an informed consent form an ethical point of view.

Informing users

On the one hand, researchers want to collect as much data as possible, but on the other hand following the principle of data economy only the necessary data should be collected. Ethical consideration and possibly even legal requirements might also impair the options to inform users. There are no clear guidelines about the way to inform users about the fact that data is collect or if it is necessary to ask if the App is allowed to collect data. Therefore, different techniques have been used by the Apps used for the five studies. Figure 10 shows the ratio of installation (according to the Android Market) to the number of samples received on our server. The fraction of users that contributed data differs dramatically for the five prototypes. While the Apps differ in many aspects the most important factor is certainly the way to inform and ask the user about collecting data.

SINLA uses the most conservative and complicated approach compared to the other Apps. If the App is used for more than five minutes the user is informed that s/he can take part in a study and a new entry is added to the main menu. After selecting the new menu entry, filling an additional form, and finally pressing the send button, data is transmitted to our server. With this approach only 0.46% of the installations resulted in a log file. During the iterative development of the PocketNavigator we tried different techniques. In early versions an opt-in checkbox hidden in an about view was used. In later versions an additional reoccurring popup asks for participation resulting in an average return rate of 7.32%. The MapExplorer asks the user for permission to log data on the App's very first screen using a pre-selected checkbox (see Figure 5). This approach resulted in a return rate of 54.76%. Tap It also informs the users when the App is started for the first time using a popup. There is, however, no option to opt-out but quitting the game. This results in a return rate of 81.31%. With 83.68% Poke the Rabbit has the highest return rate. The game never asks or even informs the player. The fraction from the total installation for Poke the Rabbit is thus likely the upper boundary of what amount of data samples can be expected. The missing 16.32% likely consist of persons that only installed the App without ever playing a single level and persons that never played with an active internet connection.

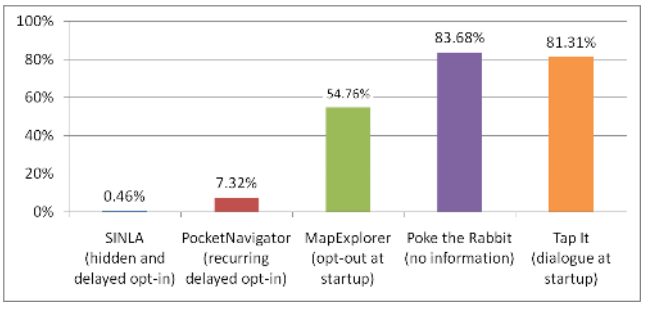

Figure 10. Percentage of installations that lead to a data sample.

Even though the five Apps differ and probably attract differ-

ent user groups we assume that the most important difference regarding the number of collected data samples is how the user is informed and asked about data logging and the true purpose of the App. The results of the PocketNavigator and especially SINLA show that hiding the option to opt-in the study dramatically reduces the amount of collected data. Presenting the option to opt-out when the App is started as we did it for the MapExplorer can be a good compromise that reduces the number of data samples only by around 30%. Comparing the results of Poke the Rabbit and Tap It shows that telling users that they are going to be part of a study without the option to opt-out only marginally reduces the amount of results. Thus, a simple popup at the Apps start is an opportunity to act ethically without losing too much data. Future work should, however, validate our results by comparing different ways to inform the user with prototypes that randomly choose one of the alternatives when starting the App. Thereby, the potential effect of the App on the number of samples could be cancelled out.

Amount of collected data

The amount of people that contributed to the results is only one factor that must be considered if looking at the amount of collected data. The second factor is the amount of data collected from the individual users. Depending on the study, data samples from persons that use the App for a short time (e.g. SINLA or the PocketNavigator) can be meaningless. Figure 11 shows the fraction of users that played a certain number of levels for Poke the Rabbit and Tap It. For the MapExplorer the Figure shows the time users spend with the App in $\frac{seconds}{20}$. In all three studies most users or players are engaged in the App for only a short time. 50% of the Tap It players played seven levels or less and 46.02% played only one or two levels of Poke the Rabbit. However, on 115 devices (1.67%) more than 100 levels of Tap It where played and on nine devices (0.18%) someone played more than 100 levels of Poke the Rabbit. 40.39% of the persons that used the MapExplorer used it for less than 60 seconds and only nine (0.21%) used it for more than one hour.

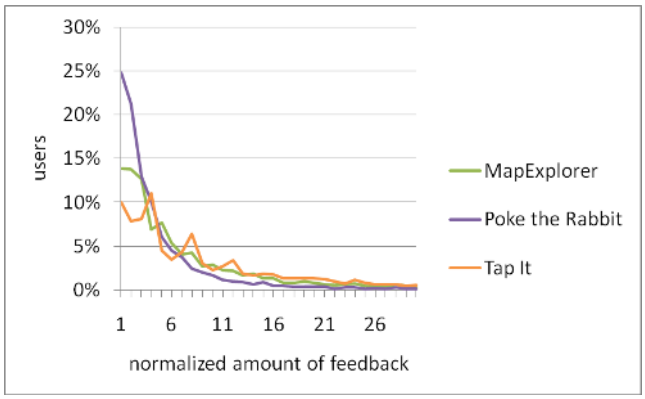

Figure 11. Amount of feedback received from users of the respective Apps. We normalized the amount of feedback for comparison purpose. For Poke the Rabbit and Tap It the x-axis represents the number of played levels and for the MapExplorer the x-axis is the time in seconds divided by 20.

For the three analyzed studies the majority of the users only tested the Apps for a short time. This prevents from analysing which off-screen visualization users prefer in the long-run using the data collected by the MapExplorer. The two games (and the MapExplorer's tutorial) on the other hand are designed in a way that even short term users contribute meaningful results. Especially with Tap It even the very first played level provides relevant data. Thus, for studies that do not need long-term user involvement it is important to collect results as early as possible and design the App accordingly. Thereby, it is feasible to collect the results before users even notice that they do not like or do not want this particular App. If long-term involvement is needed for a study it is necessary to increase the total number of users (e.g. by addressing multiple platforms) and/or increasing the number of long-term users (by increasing the quality and usefulness of the App). Both involve putting major effort in the development of an App and it might be considered to use conventional studies instead.

QUALITATIVE FEEDBACK

In their textbook Cooper, Reiman and Cronin claim that understanding the user "cannot be achieved by digging through the piles of numbers that come from quantitative study" [Cooper et al., 2007]. While quantitative data collected in experiments is used to identify that a cause results in an effect, qualitative data can help to understand why the cause results in the effect. When conducting experiments in the lab it is common to collect qualitative data either during or after the experiment.

In order to collect qualitative data we used four different approaches to receive feedback. The used feedback channels are comments from the Android Market, a feedback form inside the MapExplorer, providing email contact information in the market description and submission forms on the PocketNavigator's website. The key findings are summarized in the following two subsections.

Comments from the Market

The Android Market allows users to write comments for installed Apps by filling a form provided in the Android Market. Users can also rate installed Apps on a 5-point scale. We collected comments and ratings for the five Apps and clustered them into the following seven categories shown in figure 12: *nonsense*, *usage problem*, *misconception of prototype*, *dissatisfaction with prototype as a product*, *technical problem*, *satisfaction with technical aspects* and *satisfaction with prototype as a product*. If a comment contains ambiguous statements we decided upon the rating to which category the comment belongs.

In total only 0.4% to 0.8% of the users who ever installed an application also rated it and from these users only 16%-51% left also a comment. In general, most users rated the Apps as real products. Therefore, they report technical problems but provided little insight for the addressed research. It is noticeable that many users (36%-66%) that commented on an App reported technical incompatibilities with their Android phone or with a particular Android version. This shows how difficult it is to implement an App that runs on a variety

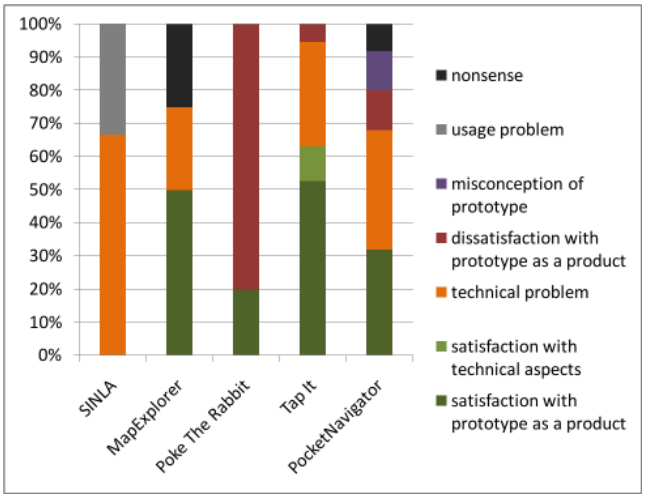

Figure 12. Distribution of clustered comments among the prototypes.

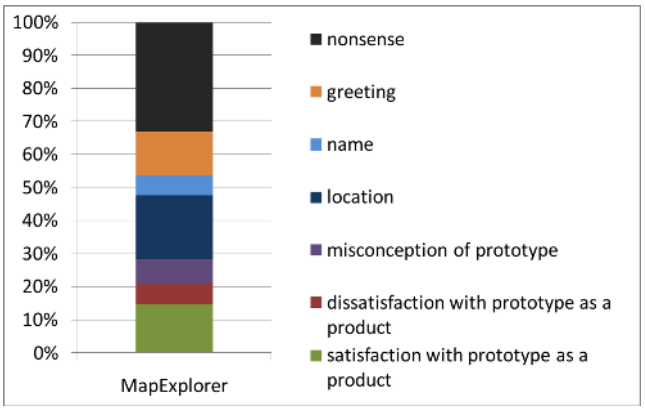

Figure 13. Distribution of clustered comments of the MapExplorer's feedback form.

of devices and different versions of a platform (e.g. four widespread Android versions). No comment mentioned privacy issues or worried about data protection. There is also no comment that provides insight in the addressed research questions. Only one comment reveals that the user is aware that Tap It is used for a study. A likely reason that we did not collect useful comments for our research is that information about the addressed questions is not prominently displayed inside the Apps. A lack of interest in the scientific background could be another reason. It can, however, also be argued that comments from the Market are not the adequate tool to collect qualitative feedback.

Feedback from other channels

A constraint to publish applications in the Android Market is to supply an email address as a contact option for users. Thus, user of all Apps can send comments and requests via mail. However, we received only emails for the PocketNavigator and besides SPAM all mails are feature requests.

Another concept for collecting feedback was used by the MapExplorer that contains a simple feedback form. In total 67 comments were collected, however, the results are as useful as the Android Market comments without any valuable feedback. Nevertheless, we were surprised what people entered into the feedback form. The clustered results are shown in Figure 13. Besides a high amount of nonsense comments, 25% percent submitted a name or their address although they have never been asked for it.

Instead of using a form inside the application, the TactileNavigator provides a form on the projects website. People are encouraged to visit the site by a description and a button inside the App that opens the phone's browser. In total 22 comments were logged but half of them were advertisements and the others were multiple submissions of redundant comments that are identical to comments from the Android Market.

All in all, four different feedback channels were used to collect qualitative feedback from users. However, none of these channels provided useful experiment-specific feedback. Most of the feedback contains either real garbage which does not fit at all to the App or assessed the prototypes as real products.

It can be concluded that qualitative feedback does not come for free and none of the obvious options provides valuable information per se. A viable approach described by McMillan et al. is rewarding users for providing feedback [McMillan et al., 2010] using in-game badges or bonus. They also describe that it is feasible to get in direct contact with participants, e.g. using Facebook and providing vouchers for participation in phone interviews. If, however, in-deep qualitative feedback is needed it should be considered if lab studies can not only provide richer feedback but are also more cost efficient.

VALIDITY

When experiments are conducted in the lab the number of participants is often limited and the tasks are artificial. This can make it difficult to generalize the findings to other populations or situations other than the artificial tasks. Since applications evaluated in mobile application stores reach a large number of users that use the App in "natural" settings, findings in the study can be generalized more confidently. How well findings can be generalized is expressed by the term external validity. The external validity is threatened, when findings have only been observed in very controlled settings as in a lab study. The mobile HCI community, for example, usually conducts studies in the lab even though a mobile or natural context would influence the outcome [Kjeldskov and Graham, 2003]. Using applications in "natural" settings means highly varying conditions, such as time, location, noise, or contemporary task. This contributes to a high external validity.

External validity comes with a counterpart, the internal validity. Internal validity describes how well the findings can be attributed to the experimental manipulation of the independent variable. The internal validity is high when the manipulation can be held responsible for the observed effects.

However, there may be huge wealth of confounding factors, such as the environmental conditions or bias in the sample, that can distort the findings significantly. As experiments conducted via mobile application stores offer less control over these conditions, the internal validity is highly threatened.

For the conducted studies we found that the design of the experiment and the unpredictability of the usage are the two biggest threats to the internal validity.

Experiment vs. Quasi-Experiment

In a true experiment, conditions are randomized. This means, that the participants are assigned to a condition randomly rather than choosing one. For an application, this means that it would switch between the conditions automatically. This, however, requires defying the participants' control over which condition is active. In some cases, such as Poke the Rabbit and MapExplorer, where the condition changes as part of the game or the tutorial, this will not bother the user. If the application is however meant for productive use, such as SINLA and the PocketNavigator, it is difficult to force the user into a certain mode, and would surely cause many users to uninstall the application.

Consequently, some experiments conducted via mobile application stores cannot be designed as true experiments, but have to be quasi-experiments. In the case of the PocketNavigator, for example, people are allowed to turn the tactile feedback on and off whenever they want. As the tactile feedback serves as the independent variable, conditions are not randomly assigned anymore. The disadvantage of such a quasi-experiment is that it is harder to rule out confounding variables. For example, the tactile feedback could mostly be turned on by very curious people. At the same time, these people navigate differently than the rest of the population. Thus, findings could be caused by the tactile feedback or by the curiosity of the participants, and it is virtually impossible to rule such influences out completely.

Unpredictable Usage

In the lab the experimenter can make sure that the participant conducts the task as scheduled. If conducting experiments in an application store, the experimenter has no control over how the task is conducted.

When the experimenter designs the experiment, s/he often has a certain usage pattern in mind that is being tested. With respect to the games Poke the Rabbit and Tap It, the designer anticipates that the user plays the game. For the PocketNavigator, we anticipated that it would be used as route guidance by pedestrians.

However, the unsupervised use of the apps offers many opportunities for unforeseen usage. For example, user may get interrupted while playing the game and they may not shut the game down before putting the mobile device aside. The application still assumes the experiment is running. If such events are not detected, the experimenter might infer that some users had severe difficulties in performing the given tasks.

In the MapExplorer example, we observed an increase in the usage time in one condition. However, this does not necessarily mean that the participants had difficulties using it, but they also could have been fascinated by the visualization and tested it out.

In the case of the PocketNavigator we envisioned the PocketNavigator to be a pedestrian route guidance system and wanted to study its use as such. However, from the log data we learned that the majority of the people never compute a route, so it is not used as a route guidance system at all. Other data shows travel speeds beyond 50 km/h, which is the speed that cars usually travel outside of the city. So the application was not used by a pedestrian but in the car. Our approach was to discount all data where the user had not computed a route and was not moving with walking speed. Nevertheless, it is impossible to be sure that no unforeseen usage is biasing the data.

This bias is a huge threat to the internal validity. In general, an experimenter that chooses to conduct experiments in an application store has to analyze the data very carefully for such unforeseen usage and process it accordingly. All conclusions drawn from the findings should clearly highlight this limitation.

ETHICAL CONSIDERATIONS

As for every conducted user study, legal regulations and ethics need to be considered. However, there are several ways how ethics can be approached, e.g. through official regulations, law or personal experience. While official regulations are not yet explicitly covering studies through mobile phones in the large scale, they look quite defensive and do not leave much room for interpretation.

To create a comprehensive data protection system throughout Europe, the Organisation of Economic Cooperation and Development (OECD) developed the "Guidelines on the Protection of Privacy and Transborder Flows of Personal Data" with the aim to protect the individuals' privacy [Organisation for Economic Co-operation and Development, 2002]. Seven principles evolved and are incorporated into an EU directive (Directive 95/46/EC). Each of these principles addresses the collection of personally identifiable information. A subject should be given *notice* that data is being collected. The collected data should only be used for a specific *purpose* stated to the subject. Each subject has to give *consent* for the data collection. The collected data should be kept *secure*. The participants should exactly know, who is collecting the data (*disclosure*). Furthermore, the subjects should be able to *access* their own data and make corrections to inaccurate data. Finally each subject should have a method available to hold the data collector *accountable* for these principles.

Some information that is stored or is available on a mobile phone is definitely considered as personally identifiable information. A prominent example is the phone number or

the phone's international mobile equipment identity (IMEI). Both are unique and can probably be tracked to a personal identity. This data should obviously not be accessed or logged in any way without asking the user for permission. A user's location is something which can be traced to particular buildings, which then might be identified as work place or home. The addresses can be looked up in address books and identify a user as a specific person. Therefore, none of our applications log any exact location information.

Most of the information we observe is either already non-person specific (e.g. the touch behaviour) or the information is abstracted to a level that an identification of a specific person is not possible (e.g. loudness instead of complete audio stream). With respect to the users, some of our applications explicitly consider some of the principles from the EU directive and *notice* the participants about the logging or even request a *consent* before the observation starts. A good trade-off between a satisfying return rate and a reasonable consent from the participant is to inform the user about the logging, while he agrees implicitly by continuing using the application. As no information that allows the identification of subjects is collected for the studies described in this paper we assume that the Apps comply with EU regulations.

CONCLUSIONS AND FUTURE WORK

In this paper, we reported from five experiments we conducted by publishing Apps in the Android Market. We showed, that it is possible to acquire a large number of users, e.g. in the case of Tap It we obtained nearly 7,000 participants contributing over 7,000,000 data points in two months only. The Apps enabled us to conduct experiments with real users in real usage situations without the huge effort that conventional methods would require. Two of the studies already yielded results that have been accepted by peer-reviewed HCI conferences [Henze and Boll, 2010, Henze et al., 2010]. Furthermore, the five case studies allowed us to get general insight, identify challenges and provide guidelines for further experiments using mobile application stores.

Our first insight is that the participants we obtained are mostly English-speaking users from the United States. The findings obtained by such a sample cannot be generalized to e.g. the world's population. Thus, when interpreting the data, researchers either have to highlight that the findings are only valid for this certain population, or they have to draw a more representative sub-sample. Furthermore, testing for differences between the different user populations allows checking whether cultural differences exist.

Secondly, when users have to opt-out of the study instead of allowing them to opt-in, a much larger fraction of the users actually takes part in the study. In e.g. the case of the PocketNavigator the opt-in method lead to a hardly useful small number of participants. However, the challenge remains to act legally and ethically, while still attracting as much participants as possible. The most successful approaches are to inform the user about the study on the first start of the application and offer the possibility to opt-out.

Another finding we made is that many users may use the applications only for a short period of time. This might be difficult if the study intends to study experienced users and not beginners. It therefore is important to collect data as early as possible and to motivate long-term use e.g. by providing badges, high-scores, or similar. Furthermore, the App has to offer a user experience that comes close to commercial products to promote extended use.

The qualitative feedback we received from the users was largely useless. The comments from the market were mostly complaints about errors and the application's usability. Mails mostly requested new features. Still, qualitative feedback is the key to understand the collected data. McMillan et al. [McMillan et al., 2010] tested two promising approaches by providing incentives, such as badges or achievements, for giving qualitative feedback and tried to contact users directly.

During the analysis of our data we found plenty of examples where Apps are used in unforeseen ways, such as the user leaving the device lying on the table with the App still running. Thus, it has to be expected that results always contain artefacts or noise, which makes it difficult to obtain valid results from the data. Researchers therefore have to filter the data before analysing it. Furthermore, it should be considered to design the App in a way that makes unforeseen usage less likely, e.g. by providing incentives for the desired usage.

As a bottom line, we found that experiments in the market allow to gain insight that otherwise would only be possible to obtain with an enormous amount of effort. However, our case studies also illustrated that such studies are not automatically successful. Future work should try to uncover pitfalls and establish further guidelines for conducting experiments in mobile application stores successfully. Furthermore, it has to be investigated, how the results obtained by experiments in application stores related to those results obtained by lab studies.

ACKNOWLEDGMENTS

This paper is partially supported by the European Commission within the projects InterMedia (FP6-038419) and HaptiMap (FP7-224675). We thank our colleagues for their support and we thank the anonymous users for using our Apps.

REFERENCES

AdMob (2010a). AdMob Mobile Metrics: App Usage Survey. http://metrics.admob.com/wp-content/uploads/2010/03/AdMob-Mobile-Metrics-Jan-10-Survey-Supplement.pdf.

AdMob (2010b). AdMob Mobile Metrics: Metrics Highlights. http://metrics.admob.com/wp-content/uploads/2010/06/May-2010-AdMob-Mobile-Metrics-Highlights.pdf.

Baudisch, P. and Rosenholtz, R. (2003). Halo: a technique for visualizing off-screen objects. In *Proceedings of CHI*, pages 481–488.

Burigat, S., Chittaro, L., and Gabrielli, S. (2006). Visualizing locations of off-screen objects on mobile devices: a comparative evaluation of three approaches. In *Proceedings of MobileHCI*, pages 239–246.

Cooper, A., Reimann, R., and Cronin, D. (2007). *About face 3: the essentials of interaction design*. Wiley-India.

Danziger, K. and Ballantyne, P. (1997). Psychological experiments. *Pictorial history of psychology*, pages 233–239.

Evers, V. and Day, D. (1997). The role of culture in interface acceptance. In *Proceedings of Interact*, pages 260–267. Citeseer.

Gilbertson, P., Coulton, P., Chehimi, F., and Vajk, T. (2008). Using "tilt" as an interface to control "no-button" 3-D mobile games. *Computers in Entertainment*, 6(3):1–13.

Girardello, A. (2010). AppAware: serendipity in mobile applications. In *Proceedings of MobileHCI*, pages 479–480.

Henze, N. and Boll, S. (2010). Push the study to the app store: Evaluating off-screen visualizations for maps in the android market. In *Proceedings of MobileHCI*, pages 373–374. ACM.

Henze, N., Poppinga, B., and Boll, S. (2010). Experiments in the Wild: Public Evaluation of Off-Screen Visualizations in the Android Market. In *Proceedings of NordiCHI*.

Kellogg, D. (2010). iPhone vs. Android. `http://blog.nielsen.com/nielsenwire/online_mobile/iphone-vs-android`. The Nielsen Company.

Kjeldskov, J. and Graham, C. (2003). A review of mobile HCI research methods. *Proceedings of MobileHCI*, pages 317–335.

Korn, M. (2010). Understanding Use Situated in Real-world Mobile Contexts. In *Proceedings of the Workshop on Research in the large*.

McMillan, D., Morrison, A., Brown, O., Hall, M., and Chalmers, M. (2010). Further into the wild: Running worldwide trials of mobile systems. *Pervasive Computing*, pages 210–227.

Michahelles, F. (2010). Getting closer to reality by evaluating released apps? In *Proceedings of the Workshop on Research in the large*.

Morrison, A. and Chalmers, M. (2010). SGVis: Analysis of Mass Participation Trial Data. In *Proceedings of the Workshop on Research in the large*.

Morrison, A., Reeves, S., McMillan, D., and Chalmers, M. (2010). Experiences of Mass Participation in Ubicomp Research. In *Proceedings of the Workshop on Research in the large*.

Organisation for Economic Co-operation and Development (2002). *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*.

Park, Y., Han, S., Park, J., and Cho, Y. (2008). Touch key design for target selection on a mobile phone. In *Proceedings of MobileHCI*, pages 423–426.

Pielot, M. and Boll, S. (2010). Tactile Wayfinder: comparison of tactile waypoint navigation with commercial pedestrian navigation systems. In *Proceedings of Pervasive*, pages 76–93.

Pielot, M., Poppinga, B., and Boll, S. (2010). PocketNavigator: vibro-tactile waypoint navigation for everyday mobile devices. In *Proceedings of MobileHCI*, pages 423–426. ACM.

Schinke, T., Henze, N., and Boll, S. (2010). Visualization of Off-Screen Objects in Mobile Augmented Reality. In *Proceedings of MobileHCI*, pages 313–316.

Simon, S. (2000). The impact of culture and gender on web sites: an empirical study. *ACM SIGMIS Database*, 32(1):18–37.

van Erp, J. B. F., van Veen, H. A. H. C., Jansen, C., and Dobbins, T. (2005). Waypoint navigation with a vibrotactile waist belt. *ACM Transactions on Applied Perception*, 2(2):106–117.

Young, P. (2008). Integrating culture in the design of ICTs. *British Journal of Educational Technology*, 39(1):6–17.

Zellweger, P. T., Mackinlay, J. D., Good, L., Stefik, M., and Baudisch, P. (2003). City lights: contextual views in minimal space. In *Proceedings of CHI*, pages 838–839.

Zhai, S., Kristensson, P., Gong, P., Greiner, M., Peng, S., Liu, L., and Dunnigan, A. (2009). Shapewriter on the iPhone: from the laboratory to the real world. In *Proceedings of CHI*, pages 2667–2670.