Hit It! – An Apparatus for Upscaling Mobile HCI Studies

Niels Henze
University of Oldenburg
Escherweg 2
26121, Oldenburg, Germany
niels.henze@uni-oldenburg.de

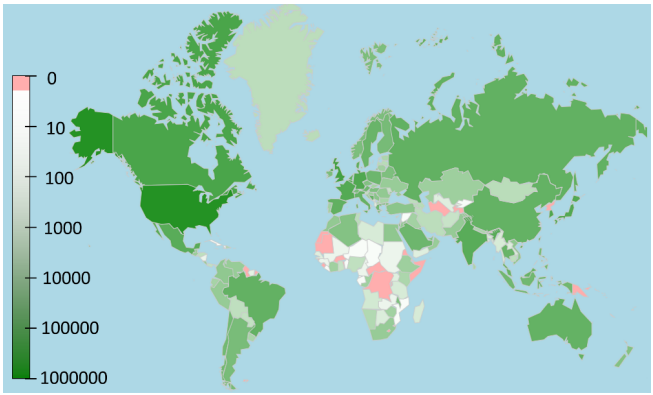

Figure 1: Number of installations of the game *Hit It!* around the globe on a logarithmic scale.

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract
Mobile HCI studies are often conducted in a highly controlled environment and with a small convenient sample. Such common studies can have a high internal validity but often lack external validity. The findings cannot always be generalized to the behaviour of real users in real contexts. In contrast, researchers recently started to use apps as an apparatus for mobile HCI research. Publishing apps in mobile application stores enables to study large samples in their 'natural habitat'. We present the game *Hit It!* that has been installed 427,255 times and, more importantly, served as a tool to conduct a number of HCI studies. The amount of collected data enabled us to study users' touch behaviour, approaches to ask for consent, the best time to deploy games, and approaches for providing tactile feedback. We discuss advantages of the approach and argue that *Hit It!* enabled findings that would be impossible to obtain using traditional studies.

Keywords
Mobile HCI, user study, app store, large-scale, game

ACM Classification Keywords
H.5.2 [User Interfaces]: Evaluation/Methodology

General Terms
Design, Human Factors, Experimentation

Introduction

Human subject studies are a cornerstone of human-computer interaction (HCI) research. As HCI researchers we often aim to conduct studies where we control all contextual factors in order to isolate certain aspects. Especially when conducting controlled experiments we aim to ensure that the obtained measurements are due to our manipulation and not due to other factors. Well designed and executed, such experiments lead to studies with high internal validity. Experiments, however, can have a high internal validity and be very reliable without necessarily having much to do with humans' real life behaviour [3]. In particular, if studying mobile HCI contextual factors can have a large influence on the user's behaviour. Mobile devices' screens are, for example, hardly readable in sunlight, typing performance is affected by walking speed, and interaction is affected by users' fragmented attention [13]. Still, Kjeldskov and Graham showed in 2003 that the mobile HCI community often conducts studies in the lab [11] even though a mobile or natural context would often influence outcome. Revising the proceedings of MobileHCI 2011 [4] we assume that most studies are still conducted in the lab. Furthermore, significantly more males than females take part in the studies and most subjects have a technical background. Therefore, mobile HCI studies often lack external validity. We cannot be sure that the findings can be generalized to the average human being using mobile devices in the real world.

Around 2010, mobile HCI researchers, including ourselves [7], started to explore app stores as a tool to conduct human subject studies. Mobile applications and games are designed to measure the users' behaviour and are published in a mobile application store. In contrast to commercial apps the apparatus is specially designed to answer research questions. As the apparatus is freely available it can be installed by thousands of users. Therefore, they are used by real users in their real life context, which ensures a high external validity. In this paper we present the game *Hit It!* that served as a highly successful apparatus for a number of different studies. The game has been installed over 400,000 times by players from all over the world. In the following section we present the design of the game. Afterwards, we describe how the game enabled us to analyse a large amount of touch data, to show how the users' touch performance can be improved, test Fitts' Law, compare different approaches for providing tactile feedback, investigate dialogues to ask for consent, and argue about the best time to deploy apps. We finally discuss the advantages and limitations of our approach and argue that it enables findings that would otherwise not be possible to obtain.

A Game as an Apparatus

During the design of the game we had to find a balance between providing a game that is worth playing and an apparatus that collects meaningful data. Based on our experience with using apps as an apparatus [7] we consider stringent tasks, users' motivation, and data collection as crucial aspects for the success of this approach in terms of the data validity and widespread usage.

Game play

The game play is inspired by the very controlled study by Park et al. [14]. Circles are presented to the player and the task is to touch these targets. The game is structured in three stages: stars, water, and fire. Each stage contains four levels and each level consists of multiple micro levels. In most micro levels, one circle is presented to the player (see Figure 4). The player advances to the next micro level as soon as the target is hit. Every fifth micro levels consists of multiple targets (see also Figure 4).

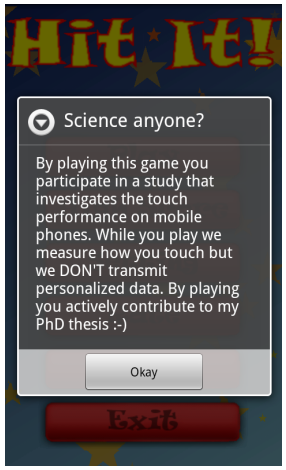

Figure 2: Dialogue to inform players about the study.

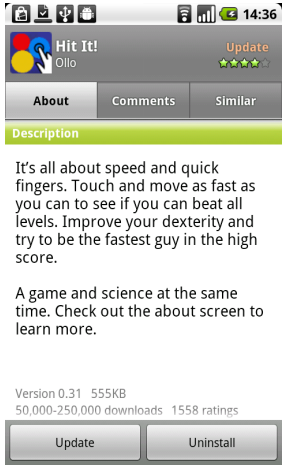

Figure 3: Game's description in the Android Market.

We designed this task to analyse the behaviour if multiple targets are presented. As soon as a target is successfully hit it disappears. The player must hit all targets to advance to the next micro level.

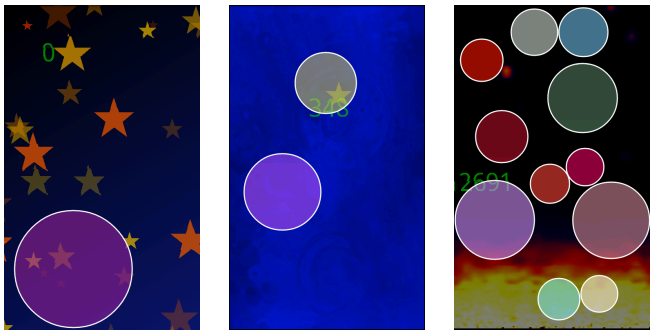

Figure 4: Screenshots with different numbers of targets

To make a game out of the basic task it must be challenging for the player. Thus, the player must complete a micro level in a certain timeframe. The time is reduced while the player proceeds through a level and a penalty point is received if a target has not been hit in this timeframe. Players lose after collecting three penalty points in one level. The faster a target is hit the higher the score a player gets. Targets with different sizes are used in each level and the average size is reduced from level to level within each stage. Levels of the first stage consist of 20 micro levels, 40 micro levels in the second stage, and 60 micro levels in the first three levels of the last stage. The very last level consists of 80 micro levels and is impossible to beat to encourage even good players to keep playing.

We made the game visually appealing to motivate intensive usage. We integrated animated backgrounds and a star highlights the score for hitting a target. The total score is shown in the background but continuously moves across the screen to minimize its effect on a player's average performance. A player receives a "badge" when successfully completing a level. To increase the long term motivation we also implemented a global high score list.

Measures and consent

We collect data about the used devices and the performance of the players. A unique identifier for each installation is derived from a device's "Android ID" using a hash function. Furthermore, we collect the user's locale, the device's type, the time zone, and the resolution. We log the position and size of the targets, the position of each touch event, and the time elapsed since the start of the micro level. The data is transmitted after a level is completed and later retransmitted if the transmission failed.

We do not collect data that allows identifying individual players or installations and we inform players about the fact that data is collected in order to act ethically and to conform to legal regulations. The modal dialog shown in Figure 2 tells players that they are about to participate in a study when the game is started for the first time. A note about the study is also included in the game's description in the Android Market (see Figure 3).

Conducted Studies

We published *Hit It!*[1] in the Android Market on October 31, 2010. We collect different kinds of data to conduct different studies simultaneously. In addition, the studies required us to alter the app to introduce different conditions by publishing an update. Depending on the design of the respective study, only new players take part in study introduced with the update. Until December 31, 2011 the game has been installed 427,255 times[2] and re-

[1] Hit It! in the Android Market: http://tiny.cc/HitIt
[2] Number of installations shown in Google's developer console

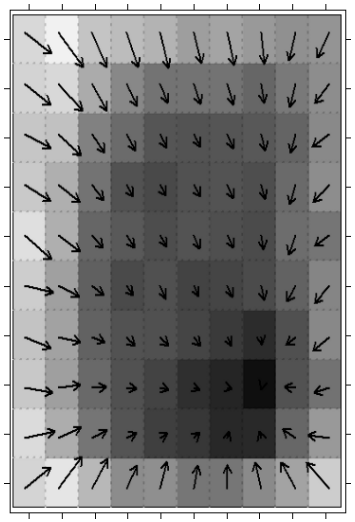

Figure 5: On average players do not touch a target's centre. Depending on a targets position the touch positions are systematically skewed towards a position in the lower-right of the screen.

ceived 3,524 ratings with an average 4.1 on the five point scale (the higher the better). The individual studies are, however, based on a subset of these installations. In the following, we provide an overview about the studies and provide a reference if the results have been published.

Analysing large amounts of touch data
The aim of the initial study, described in [8], was to record the touch behaviour of a large number of participants that execute a controlled task. We collected data from 91,731 different installations and the players produced 120,626,225 touch events. Using the data from micro levels with a single target we determined the error rates for different target sizes and positions. We showed that the error rate is lower in the centre of the screen and not surprisingly decreases for larger targets. The amount of data also enabled us to show that touch positions are systematically skewed. Figure 5 shows how the touch events are skewed towards the lower-right of the screen.

Compensation Function to Improve Touch Performance
As we found that users' touch events are systematically skewed we derived a compensation function from the data that shifts the users' touches to reduce the error rate. The resulting function has been evaluated by publishing an update of the game. The independent-measures experiment, also describe in [8], with further data from 12,201 installations showed that the function can reduce the error rate by 7.79%. We argue that such a compensation function could improve the touch performance of virtually every smartphone user.

Testing Fitts' Law
The micro levels with multiple targets are designed as what we expected to be a Fitts' law task. In [5] we tried to determine a model for these tasks by analysing 5,359,650 micro tasks from 63,154 installations. Using Fitts' law to find a model we found a very weak correlation and an implausible high index of performance across different devices. Further analysis showed a similar correlation between time and distance as with Fitts' law but only a very weak correlation with the targets' width. The results show that, for our tasks, the time to hit a target depends on the target's width. The distance to the target has a much smaller effect on the time to hit a target.

Comparing Approaches for Tactile Feedback
It has been shown that touch performance can be improved by providing tactile feedback using mobile phones' build-in vibrators [10]. We were, however, wondering if the simple tactile feedback used by commercial systems is also beneficial. We identified four different approaches to provide tactile feedback when the user's finger touches the screen. The phone can just do not vibrate if the screen is touched, it could always vibrate if the screen is touched, it could only vibrate if a target is hit, or it could vibrate if a touch missed a target. The four alternatives have been compared in an experiment with data from 27,102 installations. We show that players are fastest if the phone only vibrates if a target is hit but make the least errors if the phone only vibrates if a target is missed. It is better to provide no feedback than always vibrating whenever the screen is touched.

How to Ask for Consent
To conduct studies by publishing an app in an application store it is necessary to ask for the users' consent. One question that repeatedly came up in our group was how effective different approaches of asking for consent are. In [16] we report from a study that compares four alternative consent dialogs (see Figure 6): using an unchecked checkbox, a pre-checked checkbox, two buttons, and a single OK button. We conducted an experiment by

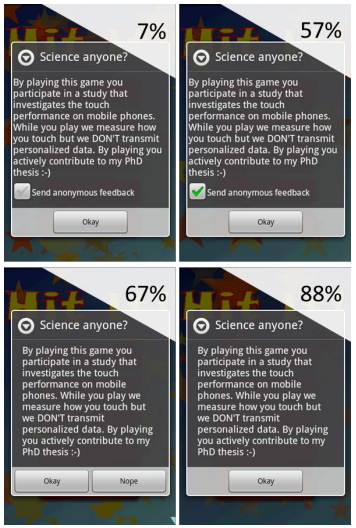

Figure 6: The four consent forms we compared in an experiment with 3,275 participants and the respective conversion rate.

randomly presenting one of the dialogues to a user. We found that all approaches except the unchecked checkbox are suitable to recruit a good share of the app users as participants (57%-88%). Considering ethical issues we, however, stress that further guidance for researchers conducting studies in app stores is needed.

Finding the Best Time to Deploy Apps
Because successful studies can require a large number of users, researchers might need to attract a large audience. The right timing when releasing or updating apps can considerably increase the number of installations. Using the time players start *Hit It!* for the first time we approximated when people install games. We also determined when developers deploy games by monitoring the Android Market. We combined data from 157,438 installations of the game and the observation of 24,647 published apps. Our results presented in [6] suggest that the best time to deploy a game is on Sunday evening.

Discussions

We presented the game *Hit It!* that we used to conduct six studies by publishing the game in the Android Market. The approach enabled findings we would be unable to derive from traditional studies. The initial study that analysed large amounts of touch data, for example, is very similar to the study conducted by Park et al. [14]. They conducted the study with 30 right-handed participants between 18-28 years. While no other information about the participants is provided we could further imagine that some of these participants might have been students at the authors' institution. Even with this relatively large and homogenous sample that participated in a very controlled lab environment, the results are affected by noise. Furthermore, some of the results are only obtained for specific regions of the screen, only three discrete target sizes have been considered, and only a single device has been used. Each additional factor that one might want to consider, such as different devices, different age groups, or different contexts, would multiply the number of participants required to find significant differences.

Considering realistic contexts in traditional studies might not even be possible because we know too little about how and when real users use their phones. Serious effort has been invested to learn about how people use their phones. Patel et al. [15] and Falaki et al. [2], for example, invested the effort to learn about mobile phone usage. Still, the conclusions are based on rather specific samples. In contrast, Böhmer et al. conducted a much larger study with an approach similar to ours [1]. They also published an app and collected data from over 4,100 participants from around the world (over 15 times more participants than e.g. [15, 2]). Thus, the most detailed investigation of mobile phone usage might have been conducted using the 'app store approach'.

There are, however, advantages and disadvantages of the large-scale approach using app stores. It is certainly not possible to develop a successful app for all tasks that we are interested in. While the sample one gets when studying navigation systems using a navigation system [17] is the perfect sample of the population that is interested in such systems it is not the perfect sample of the population. Still, these participants might represent a good sample of the relevant population. While it has been shown that it is possible to collect subjective data [12] collecting rich qualitative feedback might be easier in a controlled study. Finally, diverse usage contexts are not always desired if one wants to isolate the effect of a certain context. In particular, as it might not be possible to programmatically determine or control all contextual factors.

For the studies described in this paper we assume a high external validity while still maintaining a reasonable internal validity. E.g. for the conducted experiments we randomly assigned participants to the condition. Combined with a very large sample it can be assume that we can factor out individual differences and contextual effects. We are, however, highly interested to supplement the data with demographic and contextual information as in [12]. We therefore believe that games as well as other apps and widgets used as an apparatus and published via application stores are viable tools to supplement existing HCI research practices. We are particularly interested in using games to investigate the typing behaviour using virtual keyboards [9] and the usage of other mobile UI widgets. In general, we hope that using this approach can provide the external validity that current work often lacks.

References

[1] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In *Proc. MobileHCI*, 2011.

[2] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. Diversity in smartphone usage. In *Proc. MobileHCI*, 2010.

[3] A. Field and G. Hole. *How to design and report experiments*. Sage publications London, 2003.

[4] N. Henze. Analysis of User Studies at MobileHCI 2011. http://nhenze.net/?p=865, 2011.

[5] N. Henze and S. Boll. It does not fitts my data! analysing large amounts of mobile touch data. In *Proc. INTERACT*, 2011.

[6] N. Henze and S. Boll. Release your app on sunday eve: finding the best time to deploy apps. In *Proc. MobileHCI*, 2011.

[7] N. Henze, M. Pielot, B. Poppinga, T. Schinke, and S. Boll. My App is an Experiment: Experience from User Studies in Mobile App Stores. *IJMHCI*, 2011.

[8] N. Henze, E. Rukzio, and S. Boll. 100,000,000 taps: analysis and improvement of touch performance in the large. In *Proc. MobileHCI*, 2011.

[9] N. Henze, E. Rukzio, and S. Boll. Observational and experimental investigation of typing behaviour using virtual keyboards on mobile devices. In *Proc. CHI*, 2012.

[10] E. Hoggan, S. Brewster, and J. Johnston. Investigating the effectiveness of tactile feedback for mobile touchscreens. In *Proc. CHI*, 2008.

[11] J. Kjeldskov and C. Graham. A review of mobile hci research methods. In *Proc. MobileHCI*, 2003.

[12] D. McMillan, A. Morrison, O. Brown, M. Hall, and M. Chalmers. Further into the wild: Running worldwide trials of mobile systems. In *Proc. Pervasive*, 2010.

[13] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile hci. In *Proc. CHI*, 2005.

[14] Y. Park, S. Han, J. Park, and Y. Cho. Touch key design for target selection on a mobile phone. In *Proc. MobileHCI*, 2008.

[15] S. Patel, J. Kientz, G. Hayes, S. Bhat, and G. Abowd. Farther than you may think: An empirical investigation of the proximity of users to their mobile phones. In *Proc. UbiComp*, 2006.

[16] M. Pielot, N. Henze, and S. Boll. Experiments in app stores – how to ask users for their consent? In *Proc. Workshop on Ethics, logs & videotape*, 2011.

[17] M. Pielot, B. Poppinga, W. Heuten, and S. Boll. Pocketnavigator: Studying tactile navigation systems in-situ. In *Proc. CHI*, 2012.