App Stores: External Validity for Mobile HCI

Niels Henze
University of Stuttgart | niels.henze@vis.uni-stuttgart.de

Martin Pielot
Telefónica Research | pielot@tid.es

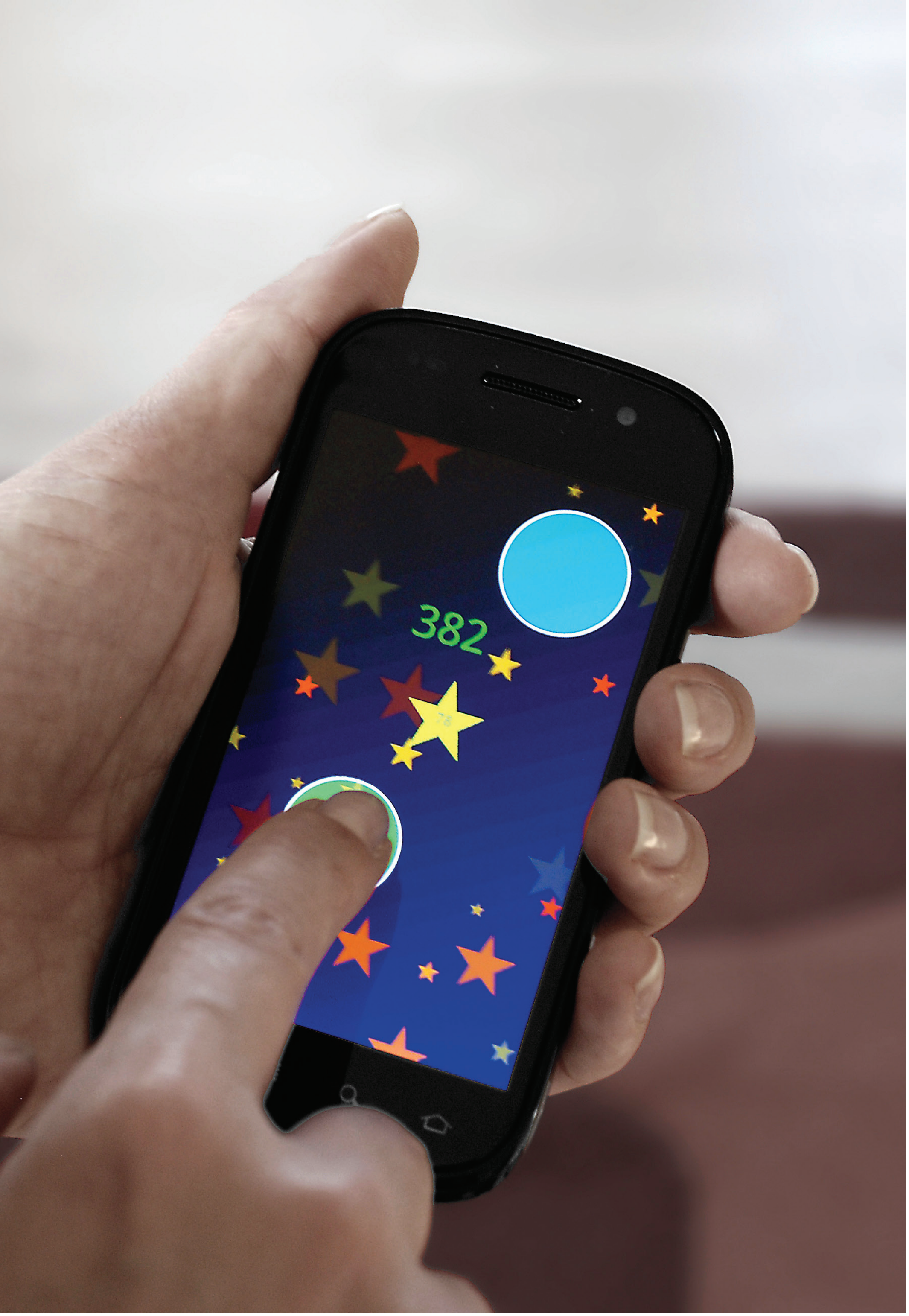
Studies in HCI research are often conducted in a highly controlled environment with a small, convenient sample. Such studies can have a high internal validity but often lack external validity. That is, the findings themselves can be reliable but cannot always be generalized to real contexts. To address this problem, researchers recently started to use app stores to bring mobile HCI research into people's lives, as opposed to bringing people into the lab. Here, we present two studies in which we published apps on Google Play. This allowed us to collect usage data from thousands of people who used the apps in their "natural habitat," which leads to high external validity of the findings. We argue that app stores are powerful tools for increasing the ability to generalize results when studying mobile and pervasive systems.

Human subject studies are the cornerstone of human-computer interaction research. Experiments are a common type of study used to uncover cause-and-effect relationships. Participants are exposed to a number of almost identical situations, in which experimenters manipulate one or several aspects: the suspected causes. Then, the experimenters measure changes in a number of observations: the assumed effects. If situations are the same except for the manipulated aspect, this aspect is the only possible cause for the observed effect. Unfortunately, it is a non-trivial challenge to keep situations absolutely the same because every study suffers from unsystematic variance; results may differ due to the time of the day, the weather, or an unlimited number of other factors.

Internal validity describes to what extent we are certain that an effect was caused by a suspected cause. Internal validity can be increased by ruling out unsystematic variance. To limit unsystematic variance, one can conduct all trials in a stable environment. For example, Bergstrom-Lehtovirta et al. used a treadmill to study the effect of walking speed on how well students can target visual shapes on phones' touchscreens [1]. Thus, they could rule out many sources of unsystematic variance, such as weather conditions. However, drawing conclusions from these findings on how people interact with phones in daily life is challenging.

Confined samples and highly specific contexts are typical threats to a study's external validity, the extent to which results can be generalized. For example, outside the lab, participants would need to obey environmental cues, get distracted by traffic, and face different weather conditions. Internal validity and external validity are often competing aspects. To increase the external validity while maintaining a high internal validity, additional variables need to be controlled with the same rigor that walking speed is controlled. Considering realistic contexts in traditional lab studies is therefore often not even possible because we know too little about what the realistic contexts are. Even if we do know the important factors, the number of participants required to determine statistically significant results grows exponentially with the number of considered factors.

Instead of conducting the study inside a sterile lab, the ideal solution for these challenges would be to conduct studies in all the situations where the studied behavior actually occurs. Taking the above example, we ideally would study typing behavior during the normal



▸ Playing our game Hit It!

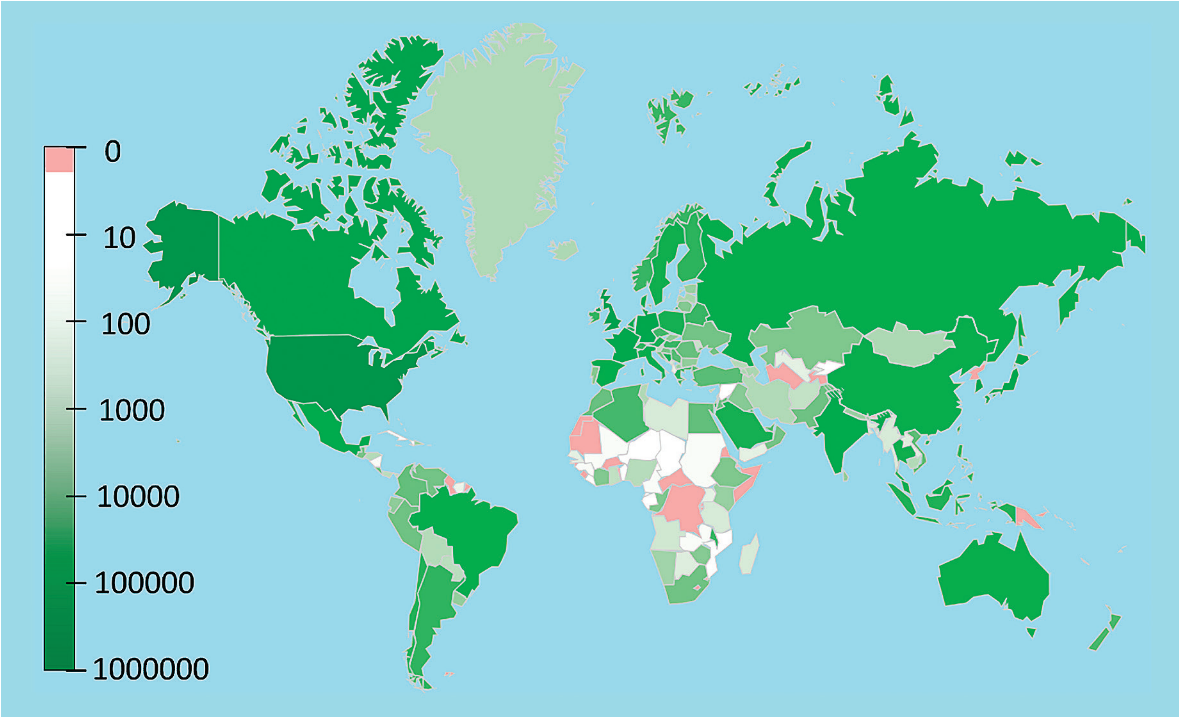
day of as many smartphone users as possible. This would allow us to catch all different types of users (tech-savvy, elderly, left-handed, etc.) and usage contexts (at night, in traffic, during a bus ride, etc.). This leaves us with the challenge that testing all relevant situations will often not be possible due to lack of resources. Or, to put it differently, how can a grad student recruit thousands of participants from all over the world and study them in all the situations they face in daily life?

Mobile HCI Research in the Large

On July 10, 2008, Apple launched the iOS App Store. It was the first open unified distribution channel for smartphone (iPhone) apps. The App Store and similar stores, such as Google Play and Windows Phone Marketplace, dramatically lowered the hurdles for developers. HCI researchers realized that this could extend the external validity of research by allowing them to conduct studies with a diverse sample of users and usage contexts.

Just as in traditional studies, researchers develop an apparatus for their study. But instead of using the apparatus for a controlled study, the apparatus is embedded into an app, which is then published on a mobile application market. Publishing the app can attract a large number of users and thus a large number of participants for the study. Early examples used app stores to collect user feedback on novel interface artifacts via the reviewing system of Apple's iOS Store. However, the reviews were often rather short and lacking in depth.

Thus, McMillan et al. explored approaches to collect rich feed-



▸ Distribution of 400,000 installations of the game Hit It! around the globe.

back from users [2]. They studied the usability of a game that was downloaded more than 90,000 times from the iOS App Store. Different ways to obtain rich qualitative feedback were tested. They asked for demographic information in exchange for game tokens. Players were encouraged to log in to the game with their Facebook accounts, so that the authors could contact them. They even arranged a number of interviews via VoIP and identified significant opportunities to improve the game.

We advanced this idea by proposing to conduct experiments in the wild via app stores [3]. For example, we published an app that uses different visualization techniques to show the location of off-screen objects on maps. Users had to use all of the visualization techniques for the same task during a tutorial. The type of visualization served as independent variable. Instead of collecting qualitative feedback, we measured how efficient users completed the tutorial with the different visualizations. The results were in line with findings from previous local lab studies.

In the following, we elaborate on two of our previous experiments that used application markets to add external validity to findings from similar lab experiments.

Hit It!—An Apparatus for Upscaling Mobile Target Selection Studies

In our work, we are interested in the interaction with mobile devices' touchscreens and how we can improve it. While the interaction with touchscreens has been studied for years, understanding the low-level characteristics of touch remains a challenge. For example, in a target-selection study, Park et al. collected 750 touch events from each of their 30 participants [4]. The resulting 22,500 touch events appear to be a large dataset, but only at first glance. A much higher number of trials are needed for a detailed analysis. Considering just 10 x 10 screen locations and 10 target sizes would already result in 1,000 different targets, which means there were only 22.5 touch events recorded per target. As touch data is heavily affected by noise, a much larger dataset can be necessary to reveal significant effects when comparing different conditions.

In contrast, we aimed to study touch behavior with more participants and a higher external validity. We published a game for Android phones on Google Play (https://play.google.com/store/apps/details?id=net.nhenze.game.button2) that records the players' touch behavior [5]. Players simply touch circles appearing on the screen. Using this game, we collected data from 91,731 players who selected

120,626,225 targets. This equals 314 touch events per pixel on a state-of-the-art screen with a resolution of 480 x 800 pixels. The amount of data enabled us to show how touch positions are systematically skewed. On the basis of this data, we derived a function that shifts the users' input as compensation. We evaluated the compensation function by publishing it as part of an update of the game and collecting data from an additional 12,201 players. The results show that the compensation function significantly reduces the error rate.

The approach enabled findings that can hardly be derived from traditional studies. When examining target selection on touchscreens, the results are affected by noise, even when using homogeneous samples and controlled lab environments. Facing a large number of confounding factors, we discovered, for example, that introducing a compensation function has only a small effect. But even an improvement of just a few percentage points becomes relevant if the task is as pervasive as selecting targets on touchscreens. However, a large sample is needed to verify this effect. The largest amount of target-selection trials ever collected enabled us to describe the users' behavior very precisely. The amount of data enabled us to analyze the user behavior for all screen locations and a large variety of target sizes. In contrast to related studies, our dataset spans more than a hundred different smartphone models. Each additional factor that one considers, such as targets with different shapes and colors or different user groups,

▸ In our field experiments, we have shown that navigation instructions encoded in vibration patterns can significantly reduce distraction and allow people to focus on the environment.

would further multiply the number of participants required to find significant differences. Further, the study confirmed the effect for users all around the world. Controlled lab studies will not scale to a level that enables analyzing how these aspects affect each other. In contrast, using a simple game and hundreds of thousands of players enabled us to upscale previous work to find subtle effects for a task executed daily by a billion smartphone users.

PocketNavigator: In-Situ Field Studies Using App Stores

Another strand of our research explores *vibrotactile navigation systems,* which use the sense of touch to deliver navigation information. Researchers have been studying such systems since the late 1990s. They showed that feedback from custom vibrotactile displays can guide travelers effectively and efficiently to destinations. In our work, we developed a pedestrian navigation system, the PocketNavigator (https://play.google.com/store/apps/details?id=org.haptimap.offis.pocketnavigator), that provides instructions through a conventional smartphone by using its built-in vibration motor. On the basis of a traditional field study, which compared the PocketNavigator with a visual navigation system, we provided evidence that participants were less distracted when being guided by tactile feedback. Furthermore, the participants very much liked the general idea. However, our study also shares its limitations with this previous research. We did not study in-situ usage. Our users employed the app to fulfill the artificial task that we had assigned to them. They navigated to destinations because it was part of the study. They were enthusiastic but only after we had given them extensive training.

To investigate why researchers constantly show that tactile feedback provides a benefit, despite the lack of real-world adoption in the past few years, we conducted an in-the-wild version of the field experiment [6]. We refined the PocketNavigator and published it on Google Play. During a period of 11 months, we recorded 8,187 routes by 3,338 users. However, we could not limit how people used the application. In fact, for only a fraction of these 8,187 routes was the PocketNavigator used as a pedestrian navigation aid. This is already a crucial insight. It indicates that pedestrian navigation is not necessarily the primary motivation to use a navigation application. Our local study could never yield such findings, since our participants were forced to use the system as a pedestrian navigation system.

To keep only those routes where the PocketNavigator was definitely used as pedestrian navigation aid, we applied a strict set of filters. We finally considered 301 trips by 112 users. Vibration feedback was used in 29.9 percent of the trips. When vibration was enabled, users interacted significantly less with the touchscreen, looked less often at the display, and turned off the screen more often. Hence, we confirmed that the findings from the body of lab and field studies also apply to daily use in the wild.

In the local study, all participants had to test tactile feedback, and provided positive feedback about it in post hoc interviews. However, in the in-the-large study, tactile feedback was used in less than one-third of the trips. A participant in the local study provided a potential explanation: "When reading the information sheets, I never thought these vibration patterns would work. But in retrospect, it was much more intui-

Considering realistic contexts in traditional lab studies is often not even possible because we know too little about what the realistic contexts are.

tive than I expected." So, what the local study could not show is what percentage of people give up trying to learn how the novel feedback works if it is not sufficiently intuitive. The results of the local study showed that tactile feedback reduces distraction if used as the only modality—in other words, when the screen is turned off. In fact, the in-the-large study showed that when users enable vibration feedback, they turn off the screen more often. Hence, our local study showed the value of encouraging users to turn off the visual feedback, while the in-the-large study showed that the interface encourages this behavior.

Users from around the world used the PocketNavigator in all kinds of locations. Not only do the participants represent a large population, but the usage context also represents diverse situations. As, strictly speaking, an experiment is valid only for the context in which it was conducted, the local study cannot be generalized beyond Germans using the application in the center of Oldenburg, the city where the field study took place. The in-the-large study provides the external validity that was lacking in the local study. However, we had to trade external for internal validity. To avoid negative ratings, we allowed users to turn the tactile feedback on and off at any time. Hence, each participant "selected" her own experimental condition (visual feedback only, tactile feedback only, both). This experimental design is called *quasi-experiment* and is known to be a threat to internal validity. Nevertheless, the two studies together allow us to combine results with both high internal and external validity. This closes the gap and allows us to conclude that tactile interfaces, when used in daily life, truly have a positive impact on how much attention people pay to the environment.

External Validity for Mobile HCI and Beyond

HCI research, and even mobile HCI research, often focus on highly controlled experiments conducted in sterile environments. Often, we develop new interaction techniques, then we design the studies, and in the end we recruit our colleagues, students, and peers as participants. One could argue that we as a discipline mainly investigate how HCI researchers interact with digital devices in laboratories. This kind of problem is not new. It is common for psychology students to participate in experiments at their university to earn credit points for their degree. Consequently, the population that is best studied by psychologists is psychology students, and the best-studied context of use is the university lab. Mobile HCI researchers risk a similar judgment from neighboring disciplines. But it's not for a lack of alternative methods. A number of approaches for studying a wider population have been proposed in the past. Online questionnaires, crowdsourcing, and games with a purpose are some examples. However, they hardly increase the external validity notably, are usually still limited to small biased samples, and, most important for us, do not address the mobile domain.

For the two described large-scale studies, we assume high external validity while still maintaining a reasonable internal validity. Combined with a large sample, we assume that individual differences and contextual effects are factored out. A growing body of work uses the same approach. Mobile games, apps, and widgets are used as the experiment's apparatus and are published via application stores. This approach can be a viable tool to supplement existing HCI research practices. Research in-the-large can provide our field with the external validity that current practices fail to provide and overcome the focus on students that is evident nowadays. And it is not even limited to the mobile domain. App stores for desktop computers and TVs became popular recently and are ready to be exploited by HCI researchers.

ENDNOTES:

1. Bergstrom-Lehtovirta, J., Oulasvirta, A., and Brewster, S. The effects of walking speed on target acquisition on a touchscreen interface. *Proc. of the Inter. Conf. on Human-Computer Interaction with Mobile Devices and Services.* 2011.

2. McMillan, D., Morrison, A., Brown, O., Hall, M., and Chalmers, M. Further into the wild: Running worldwide trials of mobile systems. *Proc. of the Conf. on Pervasive Computing.* 2010.

3. Henze, N., Pielot, M., Poppinga, B., Schinke, T., and Boll, S. My app is an experiment: Experience from user studies in mobile app stores. *Inter. Journal of Mobile Human Computer Interaction.* 2012.

4. Park, Y.S., Han, S.H., Park, J., and Cho, Y. Touch key design for target selection on a mobile phone. *Proc. of the Inter. Conf. on Human-Computer Interaction with Mobile Devices and Services.* 2008.

5. Henze, N., Rukzio, E., and Boll, S. 100,000,000 taps: Analysis and improvement of touch performance in the large. *Proc. of the Inter. Conf. on Human-Computer Interaction with Mobile Devices and Services.* 2011.

6. Pielot, M., Poppinga, B., Heuten, W., and Boll, S. PocketNavigator: Studying tactile navigation systems in-situ. *Proc. of the Conf. on Human Factors in Computing Systems.* 2012.

ABOUT THE AUTHORS

Niels Henze is a senior researcher at the University of Stuttgart, Germany. He received his Ph.D. from the University of Oldenburg for his work on camera-based mobile interaction for physical objects. Henze is interested in large-scale studies using mobile application stores, interlinking physical objects and digital information, and multimodal interfaces.

Martin Pielot is an associate researcher at Telefónica Research, Barcelona, Spain. He received his Ph.D. on conveying spatial information via tactile displays in 2012. Pielot is interested in large-scale studies as a means to study his research on non-visual and ambient interfaces in-situ and in the wild.

DOI: 10.1145/2427076.2427084