Was it Worth the Hassle? Ten Years of Mobile HCI Research Discussions on Lab and Field Evaluations

Jesper Kjeldskov and Mikael B. Skov
Centre for Socio-Interactive Design, Department of Computer Science
Aalborg University, Selma Lagerlöfs Vej 300, DK-9220 Aalborg East, Denmark
{jesper, dubois}@cs.aau.dk

ABSTRACT
Evaluation is considered one of the major cornerstones of human-computer interaction (HCI). During the last decade, several studies have discussed pros and cons of lab and field evaluations. Based on these discussions, we conduct a review to explore the past decade of mobile HCI research on field and lab evaluation, investigating responses in the literature to the "is it worth the hassle?" paper from 2004. We find that while our knowledge and experience with both lab and field studies have grown considerably, there is still no definite answer to the lab versus field question. In response we suggest that the real question is not *if* – but *when* and *how* – to go into the field. In response we suggest moving beyond usability evaluations, and to engage with field studies that are truly in-the-wild, and longitudinal.

Author Keywords
Evaluation; study; lab; field; in-the-wild; in-situ

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Evaluation of technologies is generally considered one of the major cornerstones in interaction design and human-computer interaction, and it is well known that most HCI design processes include evaluation as a key component. This is also true for mobile HCI. When the field of mobile HCI began to evolve into a distinct area within human-computer interaction about 15 years ago, the issue of evaluation naturally appeared on the agenda almost immediately. In the proceedings of Mobile HCI 1998, Johnson encouraged researchers and practitioners to investigate further into the methods and data collection for evaluating mobile devices, and he suggested that *"the conventional usability laboratory would not be able to adequately simulate such important aspects as the weather and could not easily provide for the wide range of competing activities and demands on users that might arise in a natural setting"* and he continued by saying that *"data collection methods would be needed that were outside the common range of usability studies"* [28].

MobileHCI '14, September 23 - 26 2014, Toronto, ON, Canada
Copyright 2014 ACM 978-1-4503-3004-6/14/09…$15.00.
http://dx.doi.org/10.1145/2628363.2628398

Despite these initial suggestions, Kjeldskov and Graham's survey of mobile HCI research between 2000-02 showed no research focusing on mobile evaluation methodology, and that 71% of all evaluations of mobile devices and services were done in the lab [35]. In direct response to this, we conducted a comparative study on field and lab evaluations of a mobile system with the purpose of investigating the value of evaluating usability in the field [36]. In this study we found – surprisingly – little added value in the field setting, prompting us to ask the somewhat provocative question "is it worth the hassle?" This study raised a heated debate when presented at the 2004 Mobile HCI conference, as observed by Iachello and Terrenghi [25], and sparked a long lasting discussion in the mobile HCI research field of what methods and techniques are appropriate.

Our comparison study [36] influenced a number of follow-up studies contrasting lab-based approaches with field-based ones, and somewhat polarising the research field into two distinct camps of thought, one taking an ethnographic research approach, the other one taking a usability engineering one. As a result of this, as of February 2014 the "is it worth the hassle" paper has 191 citations according to Google Scholar, and it is fair to say that it has had a strong impact on the research field by putting the discussion of empirical methodology in mobile HCI on the agenda – as originally suggested by Johnson [28].

Following up on these discussions, we have reviewed the publications that have responded to the *"is it worth the hassle"* question posed in 2004 [36], and considered whether posing this provocative question a decade ago *was worth the hassle* in terms of subsequent research on the topic. We have done this to investigate and understand how the last ten years of mobile HCI research discussions on lab and field evaluations have unfolded – what we have learned and where we are at today, what challenges we are faced with in the discussion of conducting lab and field studies, and what opportunities for future thinking in this discussion have emerged.

BACKGROUND

The discussion about where and how to do evaluations in mobile HCI is based on the distinction between field and lab studies in research methodology. As space here is limited we are not going to provide a lengthy discussion of research methodology definitions, but only briefly outline what is meant by research in the field and in the lab as this distinction is important for the following discussion.

Field studies are characterized by taking place in "the real world" with researchers spending considerable amounts of time in the real social and cultural context of their study. Data is typically gathered through observations, interviews and surveying techniques. The major advantages are the gathering of large amount of rich and grounded data, and a high level of *ecological validity*. Disadvantages are unknown biases, unknown external validity/generalizability – and typically *low level of control.*

In contrast, lab studies take place in controlled environments created for the purpose of research. Data is typically gathered with precise instruments, such as video recording, logging and questionnaires, and the studied phenomena is placed in an artificial environment where it cannot be disturbed from the outside. Major advantages are the ability to focus on detail, with high replicability, and with large *experimental control.* Disadvantages are the limited relations to the real world, unknown external validity – and typically *low level of ecological validity*.

The "Hassle" paper

In 2004 we published a comparative study of evaluating the usability of a mobile system in a field study and in a lab study [36]. In that paper, which we will refer to as the "Hassle" paper, the purpose of the study was firstly *"to compare the outcome of evaluating the usability of a mobile system in a laboratory setting and in the field in relation to identified usability problems and time spent on conducting the evaluations"* and secondly *to "describe two techniques used for 1) improving the realism of laboratory settings by including mobility and context, and 2) supporting high-quality video data collection when evaluating usability of mobile devices in the field"* [36].

Both evaluations involved six professional nurses with the same amount of work and IT expertise. All nurses used and interacted with the same mobile system (a context-aware handheld patient record), and in both evaluations close-up display video was recorded using a wireless micro-camera mounted on the mobile device. The field evaluation took place in a hospital during ordinary work activities at the ward over a couple of days. The lab evaluation took place in a traditional usability laboratory that was transformed to simulate a hospital ward with a hallway and several rooms (furnished with beds, tables etc.) and with actors acting as patients. The data analysis produced two lists of usability problems – one for each setting. Each problem was classified as cosmetic, serious, or critical, and it was noted how many test subjects it had been experienced by. Time spent by the investigators was calculated from a log for both conditions. These metrics on usability problems and time spent were used for comparing the two conditions.

We found that the lab study revealed more usability problems than the field study, where the lab condition found 36 usability problems while the field found 23 usability problems. Out of the total number of usability problems (N=37), 14 problems were unique to the lab, and only one problem was unique to the field. Looking at the 14 unique lab problems, 9 were classified as serious and one as critical. When considering effort spent, the lab took 34 person-hours while the field took 65 person-hours.

Based on these empirical findings, the Hassle paper discusses the added value of evaluating usability in the field compared to in a lab and state that *"quite surprisingly, our study shows that compared to setting up a realistic laboratory study evaluators achieve little added value when taking a usability evaluation of a context-aware mobile device into the field"*. Since the lab evaluations were able to identify the exact same usability problems, except one, and using less effort in man-hours, we pose a confronting question of whether such field evaluations are "*worth the hassle*". We do, however, not provide a direct answer to the question in the 2004 paper. However, we summarize that simulating context in a lab evaluation can facilitate solid identification of usability problems, and that the lack experimental control in the field can undermine the ability to focus an evaluation on specific parts of a system. Ultimately, we suggest that *"expensive time in the field should perhaps not be spent on usability evaluation if its possible to create a realistic laboratory setup including elements of context and requiring mobility"* and that *"field studies may instead be more suitable for obtaining insight needed to design the system right in the first place"* [36].

As argued in the introduction, the Hassle paper triggered a debate about lab and field evaluations in mobile HCI and it has been cited and used significantly for the past decade. We wish to investigate what kinds of discussions we have seen in this period and what we have learned about field and lab studies in mobile HCI.

METHOD

We have conducted a literature review examining the research discussions on lab and field evaluations in mobile HCI following on from the Hassle paper [36] from 2004. According to Google Scholar, the Hassle paper was cited 191 times as of February 2014. Out of these citing publications we were able to obtain 165 electronically. These included conference papers, workshop papers, journal articles, book chapters, and Ph.D. theses. 12 of these publications turned out to be in a language we could not read (e.g. French, Spanish, Chinese, Finish) and were therefore disregarded. The remaining 153 publications were then printed, catalogued and reviewed.

We reviewed each publication by identifying what topic it was addressing, what its contribution was, where the Hassle paper had been cited, and what it had been cited for. We then gave each publication a number of key words that described it briefly (e.g. "lab", "field", "experiment", "longitudinal"). After having reviewed roughly half of the publications, we used the key words to group the papers. We then continued reviewing the rest of the papers using these groups in addition to individual key words. After all publications had been reviewed and grouped, we went through each of them again for a "sanity check" on our first round of reviewing. As the final step we clustered the groups into 3 overall themes. The 3 identified themes covered 142 of the publications (Table 1). Another 7 publications cited the Hassle paper for (sometimes obscure) reasons that were not related to its core focus and content (such as mobile devices having small displays), and in 4 of the publications, the paper was listed in the references, but was not actually referred to in the body text.

Theme	No. of publications
Using lab or field	N=62 (24)
Comparing lab and field	N= 16 (13)
Discussing lab and field	N=64 (22)
Total	N=142 (59)

Table 1. Themes and their number of publications. Numbers in parentheses count the publications referred to in this paper.

The first theme of "Using" covered publications where the Hassle paper is referenced in studies using either field or lab for evaluating a mobile device or service. Within these publications 23 reported findings from the lab and 39 from the field. The lab studies very often involved a simulation of the real use context and cited the Hassle paper for the appropriateness of doing so. The field studies typically reported from an evaluation of a system in real world use, usually voicing the authors' disagreement with the Hassle paper. The second theme of "Comparing" covered publications where an actual empirical study had been carried out that investigated into the relative strengths and weaknesses of lab and field approaches. The third theme of "Discussing" covered publications that took up the topic of where and how to study and evaluate mobile devices and services, typically reviewing the literature and proposing new methods and techniques for the lab or the field.

In the following we will take a closer look at the different research represented by these themes, exemplified by selected representative papers.

FINDINGS

Using Lab or Field Studies

Out of the total 142 publications, we identified 62 papers that present mobile HCI studies conducted in either the lab or in the field, typically using the Hassle paper (and others) to justify the chosen evaluation setting.

Lab studies

The findings presented in the Hassle paper appears to have inspired research on exploring ways of increasing realism in controlled environments, in similar ways to the imitation of a hospital ward in a usability lab in [36]. Of the 23 papers reporting from lab evaluations, 14 involved some kind of context simulation. These simulation studies range from high-fidelity setups such as the use of a ship simulator in [38], to experimental setups where the level of simulation is rather low. As an example of the latter, Holzinger et al. report from a study where the real life environment of an ambulance officer is simulated by the participant sitting on a chair in an office holding the PDA in his hands without laying down the elbows [23].

The group of lab simulation studies shows great variation in what aspects of use context is being simulated, and great creativity in how they are simulated. Firstly, a notable body of lab studies simulates the physical real world environment in great detail. Examples of this include the work of Alsos and Dabelow [2] who carried out a usability evaluation of three PDA based medication systems in a full-scale model of a section of a hospital ward allowing them to effectively collect data from 56 simulated ward rounds with 14 physicians. In a similar manner, Holone et al. [21] evaluate an indoor navigation system for wheelchair users in a "close to real world setting" by turning two floors of a building into a controlled environment that could be observed by the researchers. In the work of Vastenburg et al. [60] evaluations were carried out in a dedicated living room laboratory simulating a home environment.

Some lab studies attempt to reconstruct different ambient aspects of a real world setting like noise, signs of potential danger, or presence and activity of other people. Examples of this include the work of Kondratova et al. [40] who simulate elements of a construction site that would potentially influence a technician's use of a multimodal data entry. Similarly, Lumsden et al. [45] simulate a city street using a surround sound system in the lab to deliver recorded city street noise, and projections on the floor to create virtual obstacles that the user should try to avoid when walking around the lab. Such simulation of mobility is also found in other lab studies, for example in the work by Wilson et al. [62] and Maly et al. [48] where test subjects were asked to walk and navigate a track with a number of obstacles, or in Banard et al. [6] where they are asked to walk on a treadmill. An example of including the presence and activity of other people in a lab study is found in the work of Leitner et al. [44] who simulate a traffic accident to users of an emergency response system.

Most of the rest of the papers referring to a lab evaluation justify this by the need for more control, replicability, easier data capture, and less time required. Only 2 papers justify the lab approach with reference to the Hassle paper by stating simply that the lab is as good as the field, or that field studies have not proven to be better.

Field studies

Most field study papers report from empirical studies where researchers have introduced different forms of experimental control in a natural environment. Using the research method categories of [39] these studies can be characterized as "field experiments", covering the body of *"natural setting research where a number of independent variables are manipulated in the study of a particular phenomenon under controlled but realistic conditions"*. Of the 39 papers reporting from a field study, 17 involved some controlled experimentation such as usability tests, randomized trial, or quasi experiments in real use contexts.

A highly cited field experiment paper is that of Oulasvirta et al. [51] who investigated the fragmented nature of attention when interacting with mobile devices in real world settings, comparing the performance of 28 subjects across two conditions. Data was collected by means of wireless cameras capturing views of the user, the mobile device and the physical surroundings. By comparing their findings to results from earlier lab based studies of the same matter they were able to provide evidence for a difference between lab and field findings, and thereby justification for the importance of field experiments. Another example of a field experiment is the work of Dearman et al. [16] where 48 subjects were assigned to one of three mobile technology conditions in the field, and asked to carry out three different scenarios of rendezvousing with a partner. Data was then collected by means of field notes, audio recordings, data logging, questionnaires and interviews, allowing the researchers to gain insight about the participants' behavior, interactions, performance, and opinions. Finally, Howell et al. [22] report from a controlled field experiment where 56 subjects were divided into two groups in a study with three independent variables of interface metaphors and context of use for a speech-based mobile city guide.

Another notable body of mobile HCI field study research responding to the Hassle paper reports from what can be characterized as "field ethnographies" using *"qualitative and quantitative approaches to natural setting research where the researcher is present in the field from full-scale ethnographic studies of phenomena in their social and cultural context to smaller scale observational studies and contextual inquiry"* [39]. Of the 39 field study papers, 11 reported the use of ethnographic techniques with minimal researcher involvement and no controlled manipulation of variables. Some, but not many, of these studies are in-depth and longitudinal qualitative enquiries of potential areas for mobile technology deployment, such as Wilson et al.'s study of medical shift handover [63], Tolmie et al.'s study of researchers' deployment of UbiComp technologies in private homes [59], and Skattør' study of Norwegian building constructors using a PDA system over four months [56]. The largest group, however, report from shorter studies of prototype systems in use in realistic settings, using various ethnographic-style observational techniques. These include Kalnikaite et al. [32] reporting from the use of a mobile shopping application in a supermarket, Davies et al. [15] reporting from the use of a mobile guide at a historical site, and Wilfinger et al. [61] reporting from the use of an interactive TV system in people's homes.

Another eleven papers are "field surveys" using *"natural setting research where survey techniques such as questionnaires, diaries, log files, interviews etc. are use for data collection rather than the researcher being present in the field"* [39]. These 11 papers report from studies where mobile systems have been deployed in real world settings without researcher presence. A comprehensive example of this is the study by Streefkerk et al. [57] where a notification system for police officers was deployed with 30 users for four months using primarily questionnaires and data logging for data collection. In fact, data logging have played an important role in several field surveys making use of the range of sensors built in to mobile phones today. As an example, Larsen et al. [43] deployed a mobile media player for two weeks collecting data about what media was being played when, where, and in what social contexts. In another example, Jambon and Meillon [27] equipped a group of skiers with a self-performance system, which relayed usage data from a camera, accelerometer and GPS back to the researchers via wireless Internet.

Comparing Lab and Field Studies

As the second theme of research we identified 16 papers that empirically compare lab and field evaluation conditions for mobile systems. Among these some papers conduct studies where they directly compare field and lab conditions, for example [17, 30, 47, 50, 58] while other papers add extra conditions to their comparison and not only consider field and lab, for example eye tracking in lab-based testing [19], or heuristic inspection [18, 37].

In light of the passionate discussions generated by the Hassle paper when presented at Mobile HCI 2004, as observed in [25], it is interesting to observe that no study has aimed at replicating it to confirm or reject its findings. Quite the opposite, all of the comparison papers we have reviewed report from studies where the focus, test subjects, tested system, or the field and lab environments have been very different from the study reported in the Hassle paper. Three studies [17, 30, 50] are similar to the Hassle study in that they address the topic of field versus lab evaluations directly, but they nevertheless differ quite significantly in their experimental setup. For example, in [50] the test subjects were all apprentices at a technical school, and not professional/experienced workers. In [17] the study compared test subjects seated at a table (lab) with test subjects on a train (field). In [30] the study was carried out in an office district, the system was a mobile system for transferring files, and the number of test subjects was considerably higher (20 per condition versus only 6 in the Hassle study). Thus, while replication is fundamental in many other scientific disciplines (i.e. medicine or physics), it seems to play an insignificant role in mobile HCI.

Despite the lack of actual replication, some of the reviewed studies claim to confirm the findings in the Hassle study, while others claim to reject them. Several studies, for example [5, 18, 19, 26, 30, 47], find that properly conducted *lab* usability evaluations can identify many of the usability issues (i.e. usability problems) identified in field evaluations. In a sense, these studies repeat the question posed in the Hassle paper if it is worth doing usability evaluations in the field? As an example, Kaikkonen et al. [30] state that field-testing may not be the optimal way for testing a mobile user interface, as they found no difference in the number of problems between the two test settings. On the hand, Duh et al. [17] found that the field led to the identification of additional usability problems compared to the lab. This was also the case in the study by Nielsen et al. [50] who coarsely state in their title that "it's worth the hassle!" regardless that their field condition was a simulation in a controlled environment (with great resemblance to the *lab* condition in the Hassle paper), thus making it questionable if any claims about real world use can be made from this study. The observation of added value in the field in a comparative study is, however, validly supported by Baillie and Schatz [4] who found that the overall system usability (effectiveness, efficiency, and satisfaction) was rated more highly in the field than in the lab, thus showing a difference between lab and field not addressed in the Hassle paper.

Several comparison papers address the lab and field discussion in quantitative studies focusing comprehensively on statistical differences, for example [5, 29, 58], and as a consequence include a higher number of test subjects than in the Hassle study (which had 6 in each condition). Kaikkonen et al. [30] reports from 20 subjects in each condition, and Barnard [5] from 41 in each. In relation to this, Kaikkonen et al. discuss the number of subjects needed and argue that while six subjects may be adequate for conducting a usability test, the number of subjects needed is different when *comparing* approaches, and that a higher number of subjects *"can increase the power of the test to find differences between two test settings"* [30].

While almost all 16 comparison-papers explicitly apply the terminology lab and field to describe their study conditions, it is quite obvious that there are rather varied understandings on what constitutes a field environment and what constitutes a lab environment. In [58] the lab condition was set up to resemble parts of a sports stadium, while in [50] the lab evaluations was done in a room where subjects were sitting at a table. In [30] the field evaluations were done in the a city center during rush hour, while in [50] the "field" environment was a technical high school warehouse that was *"similar to real working environment"*. Similarly to [50] for the comparison between lab and field environments Khanum et al. [34] *"created two labs at in the school, one for field testing sessions, and one for laboratory testing sessions"*. In our discussion section, we will return to such use and understanding of field and lab.

Several studies conclude that field studies and field evaluations are time-consuming and costly [18, 26, 30]. Kaikkonen et al. report that studies showed that usability field-testing required double the time compared their lab testing [30], while Jambon et al. found that their field tests required almost triple the time compared to lab tests [26].

Discussing Lab and Field Studies

As the third theme, we identified 64 papers that discuss the need, opportunities, implications, or limitations of field and lab evaluations in mobile HCI. Compared to the two other groups of papers (using and comparing) these papers go beyond the more practical use of lab or field studies, but they don't empirically compare the two conditions as the 16 comparison papers. Having said that, these 64 papers are nevertheless rather different in their discussion focus and approach, but they typically integrate a quite strong and detailed discussion on field or lab evaluation approaches in mobile HCI and ubiquitous computing.

From a research methodology perspective, we identified both theoretical and empirical discussion papers. E.g. some papers pursue theoretical perspectives where they outline opportunities or challenges of conducting field or lab-based evaluations. For example several papers are extensive paper reviews [1, 3, 7, 10, 11, 20, 55]. As an illustrative example, Carter et al. describe a comprehensive study on ecological validity in ubiquitous computing stressing that we as designers and researchers should focus on evaluation methods, and prototyping tools, that support realistic use in realistic settings [11]. Other papers are empirically based and typically tend to focus their discussions on either field evaluations or lab evaluations, for example [8, 9, 49, 54]. Brown et al. illustrate how we as researchers need to re-focus our approaches but also views of solid studies when evaluating and understanding experimental systems in the wild [8]. They argue that replication of studies is impossible and infeasible for field trials, and they reject those researchers who advocate more standardized approaches to trials. They argue this by the fact that *"social settings involving humans and technology contain far too much variability to be reproducible in any straightforward way"*.

As noted by Iachello and Terrenghi [25], the lab versus field discussion is associated with passionate and strong opinion and world-views. This is also observable in the group of discussion papers reviewed. In a number of papers the authors argue quite strongly for conducting field studies or evaluations in the field, for example [8, 41, 52, 53, 54]. Explicitly exploring the field approach, TOCHI ran a special issue in 2013 on "The Turn to the Wild" [12] where the editors and contributing authors focused on in-the-wild studies that *"seek to understand and shape new technology interventions within everyday living"* with the purpose of examining the insights, demands and concerns that this has for HCI theory, practice and design. Preceding this special issue, Rogers et al. [54] already in 2007 provided strong and well grounded argumentations on the importance of

field (or in-situ) studies in mobile HCI and ubicomp, in the directly responding paper called *"why it's worth the hassle"*. In this paper Rogers et al. stress that traditional evaluation methods and metrics (derived from laboratory settings) fail to capture the complexities and richness of the real world in which systems or technologies are placed and used. As an example it was specifically found that even something as simple as the changing nature of the physical environment, like particular time of year, can have quite significant impact on user experience.

As a consequence of these aspects, several researchers have set out to provide guidelines and techniques for improving studies and evaluations conducted in the field [7, 8, 31, 33, 55]. Roto et al. present and discuss best practices for capturing context when studying technology in the wild. They propose 18 practices, for example that you should identify and select realistic contexts for the tasks during the planning phase, and you should minimize the effects of research setup on participants and the context during the data collection phase [55]. Kaikkonen et al. also identify social context as important and argue that studies should consider the social location of the evaluation, e.g. other people if making phone calls [31]. On the other hand, Burghardt et al. provide a discussion of techniques for collecting data in the field, for example thinking-aloud, video recording, interview [9]. Brown et al. suggest that we should re-focus our paper method sections going away from illustrating replicable results to be more explicit about the natural contingencies and events that happen during trials and that these are vital in understanding the different trials contexts [8]. Finally, some provide inspiration on how to capture user interaction and experiences from a more low-level and practical point of view, for example hardware and equipment configurations for data collection [24, 52].

While the importance of field studies is rather evident (as shown above), several papers bring up field study obstacles or challenges. Two issues seem to play a significant role in these discussions namely lack of control and cost. Kellar et al. stress the lack of control when conducting studies in-situ and identify control challenges in the field, for example weather, social considerations [33]. Secondly, several researchers point to the fact that field studies are nevertheless costly in terms of time and effort spent [53, 54]. Thus, partly as a result of these challenges, some papers work with, and argue in favor of lab evaluations for mobile computing for example [7, 13, 14, 42, 46]. Kray et al. discusses lab studies using immersive video as a technique [42] and Lumsden et al. explore how to conduct meaningful lab-based usability evaluations of mobile systems [46], which is also proposed by Dahl et al. [13, 14] who suggest to simulate the use setting and environment – in their particular case they explore the role of fidelity in recreating hospital settings in laboratories, like in [36]. Finally, Billi et al. propose a methodology for evaluating usability of mobile system including mobile heuristics [7].

DISCUSSION

We have now illustrated ten years of mobile HCI research on field and lab evaluations through three identified themes namely using, comparing, and discussing. We will now turn our attention towards issues that span the three themes and address these issues in the following sections. We will start by looking at how far we have come and where we are today. We will then take up a number of issues that we feel are important to address in order to re-rail the discussion. Lastly we will put forward some points for future thinking in the lab-field study discussion, and some questions that we believe needs to be addressed and considered in our research community in order to move forward.

Status in 2014: Mobile HCI Evaluation Research

Looking back at ten years of research in mobile HCI, we think it is clear that empirical methodologies have been central. A very large body of research has discussed, at length, the pros and cons of different empirical methods for evaluating mobile technology, and several studies have made empirical comparisons between different approaches. Furthermore, as also pointed out in a recent mobile HCI research methods survey [39], there has been a huge increase in the amount of empirical studies in HCI (in the lab as well as in the field) – both absolute and relative to the amount of mobile HCI research as a whole. This research has involved significant effort into evolving our toolbox of empirical methods in mobile HCI. From the lab study side our community has arrived at new ways of simulating context, and from the field study side it has arrived at news ways of experimentation in situ. In that sense it is fair to say that we, as a research community, have responded quite well to Johnson's encouragements at the first Mobile HCI workshop in 1998 in terms of building a body of knowledge and experience with both lab and field studies that are *"outside the common range of usability studies"* [28].

It is, however, also quite clear that the research discussions and comparisons between lab and field approaches has not produced an answer to the question of whether mobile systems should be evaluated in the lab or in the field. There appears to be a general agreement that contextual realism plays an important role when evaluating mobile HCI, but this may be achieved both by simulating contextual factors in the lab or by taking the study "outside" into the field. Further, there appears to be agreement that researcher control plays an important role, but again this may be achieved both by experimentation in the field or by taking the study "inside" into the lab. This discussion basically comes down to the question of balancing *ecological validity* and *control,* and while the lab simulation and field experimentation research are quite distinct, and often done by researchers with very different backgrounds, its important to observe that they actually converge towards the same goal. Hence it appears that both approaches are valid, if paired thoughtfully with ones research aims (and claims), and if carried out well. If this is true, then the

important question is not *if* or *why* one should do lab or field studies, but rather *when* we should do what, and *how* we should then do it, so that its done well. These questions remains largely unanswered in a way that does not simply restate existing disciplinary doctrines. They are, however, not for us to attempt to answer here either, as they are much bigger than what one literature review can resolve. They are questions for the broader community of researchers to address through joint efforts. But creating a catalogue of guidelines and best practices in lab and field studies would certainly be a timely and relevant effort.

Different Understandings of the Field

As we illustrated earlier, some researchers argue strongly in favor of conducting field studies as laboratories potentially fail to capture the complexities and richness of the real world [54]. While we agree with this, we interestingly found extensive discrepancies in the way different studies applied (and thus understand) what constitutes a "field setting" in evaluations. The research studies reviewed in this paper involved quite diverse field settings, such as sitting in a train [17], walking around the center of a city [30], being in shops [19], sitting at a sports stadium [58] while others, for example [4] conducted their field study in the immediate area outside their research center building. This obviously raises the questions – *what is the field?* And *when can we say that a setting is in the field?* These essential questions remain partly unanswered.

Perhaps much more importantly, we think it is worth raising the question *does it make sense to engage in field versus lab discussions when we clearly lack common understandings of the things we compare*? It is quite clear that several of the 16 comparison papers are comparing different kinds of field and lab environments, and it makes little (or even no) sense to compare these papers against each other. For example, Nielsen et al. argue that *"it is definitely worth the hassle to conduct field evaluations"* [50] as a response to the Hassle study, but the field and lab conditions are very different in those two studies. Further, while several researchers argue for field studies as they increase realism and uninstructed use of technology [8, 54], other researchers strangely state that if we want to compare field and lab settings *"the field evaluation will have to be less realistic"* [50]. But less realism in field studies is in contrast with both Rogers et al. and Brown et al. who use the term "In the Wild" studies or in-situ studies to illustrate appropriation of technologies in the *real* world under *realistic* conditions. Such fundamental disagreements have highly influenced the discussion over the past decade.

Replication or Novelty

The HCI discipline (including mobile HCI) does not have a strong history of replicating previous studies. Wilson et al. state that replication is a cornerstone of good science where results can be validated to ensure a solid foundation for progress, but HCI research rewards novelty and is typically focused on new results [64]. This clearly seems to hold for mobile HCI research as well. In our opinion, some of the field versus lab discussions from the last decade seem to have failed as they have tried to achieve both replication and novelty at the same time – for example they have attempted to replicate the argumentation of field versus lab (e.g. tried to explore the added value of field studies), but at the same time they have aimed for novelty by evaluating a new type of system, in a different setting etc. But does novelty prevent replication, and does replication prevent novelty? Related to these concerns, we found that several of the comparison studies, for example [17, 30, 36, 50] apply usability problems as their primary metric for their comparison. As a result, the studies report quantitative data (number of problems and severity) that are easy to compare, but perhaps also leave out some of the richness that field studies offer (as illustrated in [54]). Further, usability problem identification and classification has been extensively criticized over the past years especially when applied in research studies for quantitative measures due to the evaluator effect.

Perhaps we need to refocus our discussion on field and lab studies to better reflect the inherent nature of our research discipline namely that we are concerned with discussing *the challenges, potential solutions and innovations towards effective interaction with mobile systems and services* [Mobile HCI conference series website]. As part of this discussion, we definitely need to investigate and understand how technologies are being used and adapted in real world settings – therefore **we need field studies**. But we should focus our field study research to better reflect and embrace the complexity and richness of real world interaction with technology as suggested by Rogers et al. [54]. As argued by Brown [8], we need to address the reality of in-situ studies including innovation in methods that are not replicable.

Beyond Usability and Usability Evaluations

Looking ahead, there are a number of points for future thinking in the lab-field study discussion that we would like to put forward for consideration. The first is to question whether *usability evaluations* are even what we ought to be doing in the first place when studying mobile HCI? In line with the argumentation by Rogers et al. [54] we think that a focus on *usability* simply fails to capture what it is that we really need to learn more about when we study our mobile interaction designs in use. We would argue that after 15 years of mobile HCI research and design, we have become pretty good at designing interfaces that people can operate on a mobile device in a mobile context. That is perhaps not the key research challenge anymore. Where the research challenge 15 years ago was to achieve usability on small displays and with limited means of input, processing power and network speed, for people away from their desk, the research challenge today, and what we need to learn more about, is about designing services, devices and interactions that fit well into people's complex lives, for work and

leisure, and that fit well with the abundance of other technologies that we surround ourselves with. This entails a shift from designing for interacting with individual devices, to designing for "orchestration" of *digital ecosystems* made up by a multitude of different systems and devices across ever-changing and overlapping contexts. In this challenge usability is just a basic condition, like bug-free code is. It will not get us there in itself, and therefore neither will usability evaluations – regardless of them being in the lab or in the field. Therefore we should also not use usability problems as a metric when comparing the performance of one method against another.

Beyond Non-Wild, Snap-Shot Field Studies

Moving beyond a focus on usability might be a useful prompt for approaching field studies in a different way. Rather than trying to "fix" the issue of limited control in the field by introducing experimentation, such as usability evaluations, why not consider going in the opposite direction and purposely let go of researcher control? Field experiments are fine as ecologically valid alternatives to lab experiments, but perhaps not as a controlled alternative to field ethnographies. As discussed earlier, the main value of the field is that it is *real* and perhaps *messy,* and not an amputated version of reality. That is perhaps also why the labels "in-situ" and "in-the-wild" have been adapted by some papers (e.g. [8, 12, 24, 27, 54, 55, 63]) as they are really much better at capturing the essence of what field studies should be about. So, just like a lab study without control and replicability would be considered a poor one, a field study that does not really take the researcher into an uncontrolled real world situation is perhaps not a good one either. When going out of the lab, we ought to actually make across the parking lot outside our buildings, and go all the way in to the wild. Studies in the field should embrace the wilderness and not be half-tame.

Moving beyond non-wild field studies of mobile systems should include a second element namely being longitudinal. As another piece of legacy from the tradition of usability evaluation, we have grown accustomed to grounding our knowledge in "snapshots of use" rather than repeated and sustained use over longer periods of time. This is not only true for the lab, but also for several field studies, especially the growing body of field experiments, but also most of the ones using field ethnographies for evaluation. If we are to address issues beyond usability and truly embrace going into the wild, we should also to start embracing longitudinal studies more, perhaps even entertain the thought of sometimes sacrificing some of the direct researcher involvement on order to stretch out the time in use of our systems in the field. Studies like that already exist amongst the group of field surveys described earlier, with [57] being a prime example of a longitudinal study in the wild that does not focus on usability. We definitely believe that more studies like that will give us valuable information on mobile systems use over the coming years.

CONCLUSION

"Was it Worth the Hassle?" We posed that question to investigate and understand what we have learned from the last ten years of mobile HCI research discussions on lab and field evaluations in the slipstream of the 2004 Kjeldskov et al. paper [36] that ignited much of this discussion.

We have shown that lab and field evaluation has been discussed extensively, and been a topic for many authors. We have also shown that over the course of 10 years of empirical evaluations our methodological toolbox have evolved substantially, and that today we have considerable knowledge and experience with both lab and field studies for mobile HCI. Since no answer to the lab versus field question seems to be found, we have argued that the important question is not *if* or *why* one should do lab or field studies, but rather *when* we should do what, and *how* we should then do it. As input to moving the discussion of empirical methodology forward, we have suggested that mobile HCI research should move beyond focus on usability and usability evaluation, that we should embrace field studies that are truly wild and longitudinal in nature in order to fully experience and explore real world use. Currently only a few examples of such studies exist.

To conclude, we believe that the last ten years of empirical work and research discussions of lab and field evaluations have been highly valuable for the mobile HCI research field, and therefore also that engaging with this topic of research has been *worth the hassle*.

REFERENCES

1. Abdulrazak, B. and Malik, Y. Review of Challenges, Requirements, and Approaches of Pervasive Computing System Evaluation. *IETE Technical Review 29,* 6 (2012), 506-522
2. Alsos, O.A. and Dabelow, B. A comparative evaluation study of basic interaction techniques for PDAs in point-of-care situations. *Proc. P-Health'10,* IEEE (2010), 1-8.
3. Axup, J. Building a Path For Future Communities. In *Handbook of Research on Socio-Technical Design,* (2008), 3-20.
4. Baillie, L. and Schatz, R. Exploring Multimodality in the Laboratory and the Field. *Proc. CMI'05,* ACM (2005), 100-107.
5. Barnard, L., Y, J. S., Jacko, J. A. and Sears, A. An empirical comparison of use-in-motion evaluation scenarios for mobile computing devices. *IJHCS 62* (2005), 487-520.
6. Barnard, L., Yi, J.S., Jacko, J. and Sears, A. Capturing the effect of context on human performance in mobile computing. *Pers Ubiquit Comput 11* (2007), 81-96.
7. Billi, M., Burzagli, L., Catarci, T., Santucci, G., Bertini, E., Gabbanini, F. and Palchetti, E. Unified methodology for evaluation of accessibility and usability of mobile applications. *Univ. Access Inf. Soc.,* 9 (2010), 337–356.

8. Brown, B., Reeves, S., and Sherwood, S. Into the Wild: Challenges and Opportunities for Field Trial Methods. *Proc. CHI'11,* ACM (2011), 1657-1666.
9. Burghardt, D. and Wirth, K. Comparison of Evaluation Methods for Field-Based Usability Studies of Mobile Map Applications. *Proc. International Cartographic Conference* (2011).
10. Carter, S. *Techniques and tools for field-based early-stage study and iteration of ubicomp applications: A dissertation proposal.* University of California, 2005.
11. Carter, S., Mankoff, J., Klemmer, S. R. and Matthews, T. Exiting the Cleanroom: On Ecological Validity and Ubiquitous Computing. *Human-Computer Interaction 23,* 1, (2008), 47-99.
12. Crabtree, A., Chamberlain, A., Grinter, R. E., Jones, M., Rodden, T. and Rogers, Y. Introduction to the Special Issue of "The Turn to The Wild". *TOCHI 20,* 3 (2013).
13. Dahl, Y., Alsos, O. A. and Svanæs, D. Evaluating Mobile Usability: The Role of Fidelity in Full-Scale Laboratory Simulations with Mobile ICT for Hospitals, *Proc. HCII'09,* Springer (2009), 232–241.
14. Dahl, Y. Seeking a Theoretical Foundation for Design of In Sitro Usability Assessments. *Proc. NordiCHI'10,* ACM (2010), 623-626.
15. Davies, N., Cheverst, K., Dix, A. and Hesse, A. Understanding the Role of Image Recognition in Mobile Tour Guides. Proc. Mobile HCI'05, ACM (2005), 191-198.
16. Dearman, D., Hawkey, K. and Inkpen, K.M. Rendezvousing with location-aware devices. *IwC 17* (2005), 524-566.
17. Duh, H. B., Tan, G. and Chen, V.H. Usability Evaluation for Mobile Devices: A Comparison of Laboratory and Field Tests. *Proc. Mobile HCI'06,* ACM (2006), 181-186.
18. Fiotakis, G., Raptis, D. and Avouris, N. Considering Cost in Usability Evaluation of Mobile Applications: Who, Where and When. *Proc. Interact'09,* Springer (2009), 231-234.
19. Gelderblom, H., Bruin, J. and Singh, A. Three Methods for Evaluating Mobile Phone Applications Aimed Users in a Developing Environment: AComparative Case Study. *Proc. M4D'12* (2012), 321-334.
20. Hagen, P., Robertson, T., Kan, M. and Sadler, K. Emerging research methods for understanding mobile technology use. Proc. OzCHI'05, CHISIG (2005), 1-10.
21. Holone, H., Mislund, G., Tolsby, H. and Kristoffersen, S. Aspects of personal navigation with collaborative feedback. *Proc. NordiCHI'08,* ACM (2008), 182-191.
22. Howell, M., Love, S. and Turner, M. The impact of interface metaphor and context of use on the usability of a speech-based mobile city guide service. *Behaviour & Information Technology 24,* 1 (2005): 67-78.
23. Holzinger, A., Schlögl, M., Peischl, B. and Debevc, M. Optimization of a handwriting recognition algorithm for a mobile enterprise health information system on the basis of real-life usability research. *Proc. ICETE'10,* Springer (2010), 97-111.
24. Høegh, R. T., Kjeldskov, J., Skov, M. B. and Stage J. Setting Up A Field Laboratory for Evaluating In Situ. In *Handbook of Research on User Interface Design and Evaluation for Mobile Technology,* ISR, 2008.
25. Iachello, G. and Terrenghi, L. Mobile HCI 2004: Experience and Reflection. *Pervasive Computing,* Jan-Mar (2005), 88-91.
26. Jambon, F., Golanski, C. and Pommier, P.J. Meta-evaluation of a context-aware mobile device usability. *Proc. UBICOMM,* IEEE (2007), pp. 21-26.
27. Jambon, F. and Meillon, B. User Experience in the Wild. *Proc. CHI'09 EA,* ACM (2009), 4069-4074.
28. Johnson, P. Usability and Mobility; Interactions on the move. Proc. Mobile HCI'98, GIST Technical Report G98-1 (1998)
29. Jumisko-Pyykkö, S. and Utriainen, T. (2011) A Hybrid Method for Quality Evaluation in the Context of Use for Mobile (3D) Television. Multimedia Tools and Applications, 55(2): 185-225.
30. Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T. and Kankainen, A. Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing. *Journal of Usability Studies 1,* 1 (2005), 4-16.
31. Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T., and Kankainen, A. Will laboratory test results be valid in mobile contexts? In *Handbook of Research on User Interface Design and Evaluation for Mobile Technology,* ISR, 2008.
32. Kalnikaite, V., Bird, J. and Rogers, Y. Decision-making in the aisles: informing, overwhelming or nudging supermarket shoppers? *Pers Ubiquit Comput 17* (2013), 1247-1259.
33. Kellar, M., Inkpen, K., Dearman, D., et al. Evaluation of Mobile Collaboration: Learning from our Mistakes. Technical Report 2004-13, Dalhousie University, 2004.
34. Khanum, M. A. and Trivedi, M. C. Comparison of Testing Environments with Children for Usability Problem Identification. *International Journal of Engineering and Technology 5,* 3 (2013), 2048-2053.
35. Kjeldskov J. and Graham C. A Review of Mobile HCI Research *Methods. Proc. Mobile HCI'03,* Springer (2003), 317-335.
36. Kjeldskov, J., Skov, M.B., Als, B.S. and Høegh, R.T. Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. *Proc. Mobile HCI'04,* Springer (2004), 61-73.

37. Kjeldskov, J., Graham, C., Pedell, S., Vetere, F., Howard, S., Balbo, S. and Davies, J. Evaluating the usability of a mobile guide: The influence of location, participants and resources. *Behaviour & Information Technology 24,* 1 (2005), 51-65.
38. Kjeldskov, J. and Stage, J. Exploring 'Canned Communication' for coordinating distributed mobile work activities. *IwC 18* (2006) 1310-1335.
39. Kjeldskov, J. and Paay, J. A Longitudinal Review of Mobile HCI Research Methods. *Proc. Mobile HCI'12,* ACM (2012), 69-78.
40. Kondratova, I., Lumsden, J. and Langton, N. Multimodal Field Data Entry: Performance and Usability Issues. *Proc. International Conference on Computing and Decision Making* NRC-CNRC (2006).
41. Korn, M. and Bødker, S. Looking ahead: how field trials can work in iterative and exploratory design of ubicomp systems. *Proc. UbiComp'12,* ACM (2012), 21-30.
42. Kray, C., Olivier, P., Guo, A. W., Singh, P., Ha, H. N. and Blythe, P. Taming Context: A Key Challenge in Evaluating the Usability of Ubiquitous Systems. *Proc. USE'07 Workshop at Ubicomp'07* (2007).
43. Larsen, J.E., Petersen, M.K., Handler, R. and Zandi, N. Observing the Context of Use of a Media Player on Mobile Phones using Embedded and Virtual Sensors. *Proc. NordiCHI'10,* ACM (2010), 33-36.
44. Leitner, G., Ahlström, D. and Hitz, M. Usability of Mobile Computing in Emergency Response Systems – Lessons Learned and Future Directions. *Proc. USAB'07.* Springer (2007), 241-254.
45. Lumsden, J., Kondratova, I. and Durling, S. Investigating microphone efficacy for mobile speech-based data entry. *Proc. HCI'07,* Springer (2007), 89-97.
46. Lumsden, J. and MacLean, R. A Comparison of Pseudo-Paper and Paper Prototyping Methods for Mobile Evaluations. *Proc. MONET'08* (2008), 538-457.
47. Lumsden, J., Langton, N., and Kondratova, I. Bringing the High Seas into the Lab to Evaluate Speech Input Feasibility: A Case Study. *Proc. SiMPE Workshop at Mobile HCI'10* (2010).
48. Maly, I., Mikovec, Z., Vystrcil, J., Franc, J. and Slavik, P. An evaluation tool for research of user behavior in a realistic mobile environment. *Pers Ubiquit Comput 17* (2013), 3-14.
49. Morrison, A., McMillan, D., Reeves, S., Sherwood, S., and Chalmers, M. A Hybrid Mass Participation Approach to Mobile Software Trials. *Proc. CHI'12,* ACM (2012), 1311-1320.
50. Nielsen, C. M., Overgaard, M., Pedersen, M. B., Stage, J. and Stenild, S. It's Worth the Hassle! The Added Value of Evaluating the Usability of Mobile Systems in the Field. *Proc. NordiCHI'06,* ACM (2006), 272-280.
51. Oulasvirta, A., Tamminen, S., Roto, V. and Kuorelahti Interaction in 4-second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI. *Proc. CHI'05,* ACM (2005), 919-928.
52. Oulasvirta, A. and Nyyssönen, T. Flexible Hardware Configurations for Studying Mobile Usability. *Journal of Usability Studies 4,* 2 (2009), 93-105.
53. Oulasvirta, A. Rethinking Experimental Designs for Field Evaluations. *Pervasive Computing,* Oct-Dec (2012), 60-67.
54. Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W., Kurtz, A., Hall, R. E., Hursey, J., and Toscos, T. Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. *Proc. UbiComp'07,* Springer (2007), 336–353.
55. Roto, V., Väätäjä, H., Jumisko-Pyykkö, S., and Väänänen-Vainio-Mattila, K. Best Practices for Capturing Context in User Experience Studies in the Wild. *Proc. MindTrek'11* (2011), 91-98.
56. Skattør, B. Training and Deployment as a basis for Usability Engineering of Mobile Systems. *Proc. ACHI,* IEEE (2008), 277-284.
57. Streefkerk, J.W., van Esch-Bussemakers, M.P. and Neerincx, M.A. Field Evaluation of a Mobile Location-Based Notification System for Police Officers. *Proc. Mobile HCI'08,* ACM (2008), 101-108.
58. Sun, X. and May, A. A Comparison of Field-Based and Lab-Based Experiments to Evaluate User Experience of Personalised Mobile Devices. *Adv. in Hum.-Comp. Int.,* Hindawi (2013), Article 2.
59. Tolmie, P., Crabtree, A., Egglestone, S., Humble, J., Greenhalgh, C. and Rodden, T. Digital Plumbing: the mundane work of deploying UbiComp in the home. *Pers Ubiquit Comput 14* (2010), 181-196.
60. Vastenburg, M.H., Keyson, D.V. and de Ridder, H. Measuring User Experiences of Prototypical Autonomous Products in a Simulated Home Environment. *Proc. HCII'07,* Springer (2007), 998-1007.
61. Wilfinger, D., Pirker, M., Bernhaupt, R. and Tscheligi, M. Evaluating and Investigating an iTV Concept in the Field. *Proc. EuroITV'09,* ACM (2009), 175-178.
62. Wilson, M.L, Russel, A., Smith, D.A. and schraefel, m.c. mSpace Mobile: Exploring Support for Mobile Tasks. *Proc. HCI'06,* Springer (2006), 193-202.
63. Wilson, S., Galliers, J. and Fone, J. (2007) Cognitive Artifacts in Support of Medical Shift Handover: An in Use, in Situ Evaluation. *IJHCS 22,* 1&2 (2007), 59-80
64. Wilson, M. L., Mackay, W., Chi, E., Berstein, M., Russell, D. and Thimbleby, H. RepliCHI - CHI should be replicating and validating results more: discuss. *Proc. CHI'11 EA,* ACM (2011), 463-466.