How can we transition from lab to the real world with our HCI and HRI setups?

Christian Dondrup, Lynne Baillie, Frank Broz, Katrin Lohan
School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh, UK
{c.dondrup; l.baillie; f.broz; k.lohan}@hw.ac.uk

ABSTRACT
In this position paper, we present the issues we and others have found when moving from the controlled lab space into the field. We therefore recommend some dos and don’ts for facing the challenge of transferring your research prototype from the lab to the real world. During this transfer, we often encounter crucial disconnects between our envisaged evaluation protocol and real life, often related to differing user expectations outside controlled experimental interactions. The redeployment of complex systems in unfamiliar (and often dissimilar) environments presents additional challenges.

In this paper, we present some example transitions in the fields of mobile HCI and HRI. Based on these experiences, we list the possible roadblocks that other researchers might encounter and provide guidelines and suggestions for dealing with frequently encountered issues. We hope that with this paper we can help stimulate the discussion on the pathway to undertaking scientifically reproducible evaluations in the wild.

Author Keywords
Reproducibility; HCI; HRI; Mobile HCI; Mobile Robotics

CCS Concepts
• General and reference → Reliability; Design; Performance; • Human-centered computing → Human computer interaction (HCI);

MOVING MOBILE HCI TO THE REAL WORLD

Semi-Controlled
In the early 2000s mobile devices began to have the ability to be used to access applications that had hitherto only been available in the workplace e.g. email. This led to some researchers thinking about whether the methods currently used in HCI to assess a system’s functionality and usability were still fit for purpose or needed to be changed. Kjeldskov et al. [13] undertook a study to understand this and they evaluated a new note taking mobile application for nurses: in the lab and the field (different nurses). The methods were: interviews (open ended questions) and task completion (time, errors, and satisfaction with tasks). It was envisaged that both methods would be used in the lab and in the field. However, in the end they did not do any tasks in the field but just did interviews. They then compared the results from the two studies and concluded that they found out the same information from both the lab and the field studies and therefore field studies were not worth the hassle (as they famously said in the title of their paper).

Some researchers questioned their findings and thought that if carefully chosen, and if the methods were consistently applied in both locations (e.g. lab and field), then differences may be found. The key they thought was in the consistency.

Baillie and Schatz [2] carried out such an experiment in which they undertook the same evaluation giving the same tasks to complete in the field as in the lab. They found significant differences regarding time on task and on whether or not users found certain modalities useful or not. Users wanted to complete tasks much more quickly in the field and if this was not possible, became very frustrated. As regards functionally, functions found to not be useful in the lab were very much appreciated by users in the field e.g speech to text. These are two key design issues e.g. task completion time and usefulness to the user of key functions.

This showed that undertaking evaluations in context was very important. Since then many others have found field studies useful. Our recommendation is that the key is in the consistent use of measures across the two contexts, which should be maintained whenever possible. This recommendation though may be almost impossible to maintain if the researcher is not present.

Uncontrolled
The above studies were still facilitated and supervised by researchers. One of the key issues facing us recently and in the future is how to design field studies that give us important design and usability findings for HCI and HRI. The key question we need to ask is: “what happens when the researcher is completely removed from the situation? What methods are still appropriate and more importantly will work and provide the findings that they need in terms of design and in terms of qualitative and quantitative results?”

Some researchers [10] have tried the approach of replicating to a certain degree the lab in the field. They set up a room in an assistive living complex with their kit and made the kit available at timetabled times and asked people to sign up. Thus, if there were any breakdowns, these could be fixed on the spot and so that the researcher was on hand. However, this was time consuming and ultimately not a full test of the system if it were to be used completely unassisted. However, this does meet the needs of researchers to get results. Our question
would be: is this in the long term the type of evaluation we want?

Another approach has been (Istop falls EU Project [6]) dominant in some EU projects and that has been to build a system which has complex interaction and assess it in the lab, but then to strip out most of the interaction and make only a very limited subset of it available in the field. Again we would posit that this is not really a full test of such systems.

MOVING TOWARDS HRI IN THE REAL WORLD

In 2015, Baker et al. [3], stated that over half of psychology studies fail the reproducibility test, this article was the start of a dramatic discussion in all relating fields like HRI and HCI to ensure that they can avoid the mistakes made in psychology. Already in 2014, Lier et al. [15], started their journey by producing a toolkit to support experimental studies in the interdisciplinary field of HRI to make them more reproducible [14], they demonstrated this by using their toolkit during a study that was run at different universities. Moreover, in EU Projects like the one outlined below reproducibility was tackled from different angles:

- **LIREC** (*Living with Robots and intEractive Characters* funded by the EU Robotics and Cognition programme under FP7, 2008 - 2012.), undertook the first attempt to run a long-term deployment, running the robotic platform “Shara” for 3 weeks in a lab environment autonomously (see [5, 16, 17]).

- **ITALK** (*Integration and Transfer of Action and Language Knowledge in Robots* funded by the EU Robotics and Cognition programme under FP7, 2008 - 2012), showed the strategy to repeat experiments both with the same participants over 3 different interactions and in 2 different locations (see [4, 7, 18]).

- **STRANDS** (*Spatio-Temporal Representations and Activities for Cognitive Control in Long-Term Scenarios* funded by the EU Robotics and Cognition programme under FP7, 2013 - 2017) [11], aimed at long-term autonomy for mobile robots. Over the 4 years of the project, the robot was deployed to an elder care home [9] and an office building for 30 - 120 days at a time with yearly increments. This included recurring experiments with the same group of participants over several months [12].

These strategies presented above where attempts to avoid the reproducibility and transferability downfall when producing HRI systems to be employed in the real world.

The drawbacks of the last two examples can be shown through the fact that it took the researchers in the ITALK project several months of integration to move from one lab to another. Whereas, in the STRANDS project it took the researchers the whole period of the project to reach a run time of up to 4 months at two different deployment sites. Specific examples of problems were different times of day or different days of the month imposing different constraints, e.g. the lighting and crowd density will differ and the environment might change slowly over time or very rapidly, i.e. moving furniture or events like a Christmas market as observed during the last deployment of the STRANDS project. Therefore, the scientific challenge presented itself especially when it comes to making software stable enough to run for this amount of time and develop approaches that can deal with the ever-changing environment. However, one cannot account for all possible sources of error during deployments, at the moment, systems still need supervision and fixing of hardware and software defects on-the-fly.

Overall, there are several technical improvements on toolkits like the one presented by Lier et al. [15] and integration technology like Jenkins an open source automation server that help to automate the non-human part of the software development process, with continuous integration, facilitating technical aspects of continuous delivery [21], and automated software testing [20]. Systems like github [23] that make it easy to share and collaboratively work on your software and allow it to be tied to continuous integration servers are a good step towards the right integration tools.

There are also attempts to move into different more lifelike environments for example competitions like the robocup and European Robotics League. These competitions attempt to give guidance for the transfer from the lab towards the real world. But in reality these competitions are somewhat artificial and very focused on specific tasks and, therefore, again one could fall into the trap of limited transferability of these systems. Projects like MuMMER [8] attempt to bring robotics to the wild by – in the case of MuMMER – putting a Pepper robot into a shopping mall interacting with a lay audience. This is, as mentioned above, not only challenging from a technological point of view, i.e. noise, lighting, network connectivity, etc. but also from an interaction point of view due to the unconstrained nature of the environment and types of interaction compared to the laboratories where the software is developed. For example, while laboratory experiments can expect a static physical environment, the STRANDS project encountered extreme environmental changes in both the physical layout and user traffic due to events such as Christmas markets.

DISCUSSION

We believe that truly full uncontrolled long-term field trials in public or unsupervised spaces are not being done because of the problems with the levels of prototype we build in research and because these trials can be time consuming to set-up, run, and have a high dropout rate [1, 19, 22]. Frustratingly, these costly deployments (using either a robot or a mobile device) may not always yield a sufficient amount of quality data or repeated interactions. Whether and how people interact with technologies in real environments is subject to numerous factors that might not come to light in laboratory studies.

Full field studies are possible, however, but there needs to be careful design of the evaluation protocol to ensure that useful insights are gained. However, even when doing this, we still found the following issues: if you give people too much additional paperwork (e.g. questionnaires) they will not fill them out. If you leave people alone with your system, it will probably have breakdowns: in terms of battery life, technical issues (as it is a prototype and probably hasnâ€™t been used
for a sustained amount of time) or they are research robots which are only meant to be used for short experiments.

Our recommendations are, therefore, that you must:

- Conduct WoZ or human-human mockups to test your data collection.
- Invest time in constructing a monitoring system that alerts you in case of errors or unresponsive components.
- Regularly check if users/people are actually interacting with or using the technology and be able to perhaps make adjustments on-the-fly. This usually means in person and seeing if you can collect log files.
- Be ready to deal with incomplete data, unused systems and dropouts.
- Create watchdogs that execute automated recovery behaviours for the most commonly occurring problems.
- Regularly check the current system state and help it recover from unrecoverable errors either remotely or physically.
- Have a person available to go to the deployment site who can intervene in cases where the researchers are not available or remote connections have broken down.
- If the device acts autonomously during an experiment, it might still require close supervision to intervene quickly in case of problems to be able to continue data recording.
- Especially for initial deployments, it may be useful to employ qualitative techniques for data analysis in order to gain insight into unexpected interactions between users and your system.

One key thing that helps mitigate some of the above is: 1) doing a pilot in the lab, 2) pilots in the field, and then 3) finally doing an assessment for one week to see if it is all still working before a completely unsupervised trial for a long period.

CONCLUSION
Moving from the controlled set up of the lab to the field needs careful assessment of the tasks, people, context and level of sustained interaction. This all needs to be done at the outset of the project at which time a thorough and robust set of user and interaction requirements should be drawn up. These must be based on realistic expectations of what is possible of the technology and in the context of use.

For successful deployment in real environments, preliminary testing in the target environment is key. In the case of robotics, the constraints imposed by the environment, lighting, network, layout of rooms, noise levels, etc., has to be taken into account throughout the development process. Software needs to be tested early on in the deployment environment to avoid over specialisation to lab conditions. Data needs to be collected in the actual environment and from actual interaction to be able to develop and train software components in the lab using realistic data. Even after following these best practices, experimenters should expect to encounter issues and interactions that they could not anticipate in the lab. Therefore, a flexible approach that can react to the complexity of unstructured data should be employed to inform refinements to the system that reflect the needs of real users.

ACKNOWLEDGMENTS
The authors would like to acknowledge the support of the EPSRC IAA under No. 455791 and the MuMMER EU H2020 project under grant agreement No. 688147.

REFERENCES


