Khai Truong *University of Toronto*

CONTROLLED STUDIES OUTSIDE OF THE LAB

Illustration, istockphoto.com

Often, evaluators study a computing system inside a laboratory setting to best gain an understanding of the effect of the system when different factors are manipulated. The laboratory setting allows evaluators to create not only the environment, but also the scenario in which a user study of system is conducted. Thus, the laboratory setting allows evaluators to control possible confounding variables and to develop insight about the cause-and-effect of the system when they manipulate specific usage factors. For example, it is clear that people often use mobile devices while walking. Thus, a laboratory study can be designed to test how well users might be able to interact with a mobile device while walking on a treadmill machine. Such a study, because it is conducted in a laboratory setting, would allow the evaluators to control the speed at which study participants would walk while using a mobile device, without fearing that participants must also pay attention to traffic or could be distracted otherwise.

Experiments conducted in the lab are sometimes referred to as controlled studies because evaluators can control different usage factors during the experiment. However, controlled studies do not need to occur only within a laboratory setting. Depending on the system being tested, evaluators may need to conduct the user study outside of the lab. For example, user studies of location-aware systems may require that participants interact with the system over many different locations. Such a criteria would be difficult to accomplish inside of the lab. However, if the evaluators want to be able to understand the effect of specific usage factors, they need to be able to control those factors in an outside-of-the-lab user study as well.

In this article, we discuss when and how to conduct controlled studies of mobile or ubiquitous computing systems outside of the lab. We first compare the goals of such user studies against deployment studies of systems outside the lab and experiments conducted with such systems inside the lab. We use this discussion to motivate why evaluators might conduct controlled studies of computing systems outside of the lab. We then present an example case study to explain how such studies might happen. We outline key considerations when evaluating systems in this manner. Finally, we review the benefits and limitations of this evaluation strategy.

CONTROLLED STUDIES OUTSIDE OF THE LAB VS. OTHER TYPES OF STUDIES

Experiments conducted inside of the lab offer many benefits. As mentioned earlier, they allow for precise control of study variables. Because they are done inside the lab, the evaluators can construct the test environment and scenario to include certain factors and define how those variables should be tested in the study. Evaluators would also be able to address potentially confounding variables in the study design as well. This ability to "control" the study variables allows the evaluators to isolate a key variable to study in order to measure its effect on some task. In the example given earlier, walking speed can be varied to understand its effect on how fast users might be able to read text on a mobile device. Of course, multiple variables can also be examined. For example, font size might also be something that the evaluators choose to vary and test in the study as well. Evaluators can use inferential statistical tests to analyze the quantitative data collected from the experiment. These tests can suggest how likely it is that the results could have occurred through chance. The scope of an in-the-lab user study also allows for it to be easily replicated. Because the results of a single experiment can not be generalized, a study might be repeated multiple times so that if the same results are obtained, there is reason to believe that findings from these studies are valid.

At the same time, in-the-lab user studies do not fully capture and represent real-life situations. For example, when a user is walking, she must pay careful attention to her surroundings and avoid obstacles. There might be loud noises and too much or too little sunlight, which could affect her ability to easily complete the task of reading. Additionally, there might be different types of reading content, display format, font types, etc., that could also affect her ability to read the material while walking. Thus, findings from an in-the-lab user study might not generalize beyond the context of the specific conditions in which the study was conducted.

Deployment studies can be used to answer questions left unanswered from in-the-lab experiments. In-the-wild deployments are often conducted by leaving a specific instance of a system for use by study participants over a prolonged period of time. By leaving that system in the hands of the users, it would be possible for evaluators to gain an understanding of how well that design of the system works when it is used in authentic situations. Furthermore, the nature of a deployment study means that potentially the system will be tested in many different contexts. The data gained from deployment studies enable the evaluators to gain an understanding of when, why and how a system is adopted and used. Quantitative data can be collected

to understand if the system helps the user complete tasks in different situations over time. Alternatively, baseline data can be collected before the introduction of the system in order to demonstrate the effect of the system. Evaluators may also deploy different versions of the system in order to gain an understanding of specific design factors as well.

Deployment studies, however, are not easy to carry out. First, they require fully implemented systems that work outside the lab. Second, they often need to be conducted over a long period of time. In particular, this length is used to provide participants with ample opportunities to interact with the system in different situations. A sufficient amount of exposure to the system is needed so that participants are able to gain an understanding of when, why and how to use the system and ultimately develop expertise and proficiency with using it. At the same time, it is entirely possible that over the course of a deployment study, participants still do not experience situations in which they would need to use the system or take the opportunity to do so. On the other hand, laboratory studies create opportunities for participants to develop expertise and proficiency with using a system by asking them to repeatedly use it in multiple trials. Through these trials, it is possible to examine the effect of the system on the completion of a task. Because the system would only be used in a limited scope, a fully implemented system is not required.

Controlled studies outside the labs fill the void between in-the-lab experiments and deployment studies. Evaluators can design such studies to create opportunities for study participants to test mobile and ubiquitous computing systems in specific environments and scenarios in a naturalistic manner. By doing so, it is possible to ensure that study participants have an actual chance to use the system. The system also would not need to be implemented in all possible contexts, but only these environments and scenarios. At the same time, this type of study allows the evaluators to assess the effect of the system with some amount of ecological validity – that is, external factors that can affect the users' ability to complete a task with the system are not removed from the study as could happen with an inside-the-laboratory study design.

Photo, istockphoto.com

Case study: Evaluating the effect of a Vocabulary Wallpaper Application

An essential aspect of learning a second language is the acquisition of vocabulary. However, acquiring vocabulary is often a protracted process that requires repeated and spaced exposure, which can be difficult to accommodate given the busyness of daily living. To address this challenge, we explored if rather than a single instance of reading or reviewing vocabulary for a prolonged period of time, the task of learning vocabulary can be partitioned into sessions that fit within these opportunities for microlearning. Furthermore, we explored how to create microlearning opportunities, which involve vocabulary that a learner will find engaging.

We investigated if a learner can implicitly acquire second language vocabulary through her explicit interactions with her mobile phone (*e.g.,* navigating multiple home screens) using an interface we developed called Vocabulary Wallpaper [1]. The Vocabulary Wallpaper application utilizes the device's mobility to provide learning material that can be relevant to the user's current context (*i.e.,* providing the user with vocabulary that is relevant to cafés when the person is located in a café).

This research requires a location-aware system to be built. Furthermore, the system must work across a number of different locations in order to provide the user with opportunities to learn vocabulary in different contexts. This meant that the study must be conducted outside of the lab. We generated the place-specific vocabulary using the Activity-Service engine [2] which uses community-authored content to characterize the potential activities a person can perform at a place. Using this method, we were able to systematically produce a contextually relevant vocabulary that is unique to a place (i.e., it is place-specific). However,

the method requires a large amount of community-authored content at each type of place to generate a usable place-specific vocabulary. As a result, it would not have been possible to test the system in a full deployment study.

Furthermore, to examine the effect of exposing participants to contextually relevant vocabulary meant that we needed to provide all participants with the same amount of opportunities to learn vocabulary. That is, if we conducted the research as an in-the-wild deployment study, there could be a chance that for some participants, they simply might not encounter any contextually relevant vocabulary learning moments if their lifestyle only takes them to locations where there were not enough community-authored content to generate vocabulary for the system to display.

We conducted a controlled outside-the-lab user study to evaluate the benefit of using a contextualized vocabulary to assist in second language vocabulary acquisition by enabling microlearning opportunities. In this study, we recruited 15 participants and provided each with a mobile device to use that was instrumented with:

- The Vocabulary Wallpaper to present the participants with the vocabulary and associated translations.
- A custom, Google Voice-like task widget that we developed to instruct participants where to go and what to do there.
- A Twitter widget and account so the participants could post status update messages, responding to the task we requested they perform.

The custom task widget monitors the participant's location and automatically provides the participant with instructions for where to go next and what to do there. We intentionally installed the task widget on the far left and the Twitter widget on the far right of the home screen to force users to interact with the device – creating opportunity for them to experience the Vocabulary Wallpaper application implicitly.

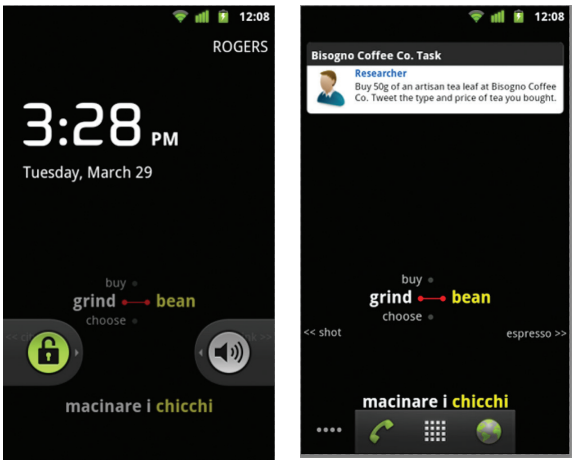

FIGURE 1. Vocabulary Wallpaper implemented as a Live Wallpaper for the Android OS to always show vocabulary content unless occluded by a running application.

The study consisted of five separate sessions. In sessions one to four, we asked the participants to visit the six businesses and to perform a task at each location. Examples of the tasks we asked participants to perform at each business type include:

- Coffee shop: Buy 50 g of an artisan tea leaf at Jobs Coffee Co. Tweet the type of tea you bought and the price.
- Cheese shop: Buy 100 g of Danish Blue Cheese at Cheese Boutique. Tweet the price of the Danish Blue.
- Bakery: Buy a French baguette at B-Bakery. Tweet the price of the French baguette.

Although tasks were repeated across participants for consistency, no task was repeated by a participant between sessions. We did not inform participants ahead of time of the tasks they were to perform. Instead, instructions were delivered to the participants through the smartphone that we provided to them.

Finally, to test whether presenting users with contextually relevant vocabulary will influence the learner's rate of vocabulary acquisition, we tested three types of vocabulary: those that are place-specific, or based on venue-type or are high-frequency words (and contextually independent). We conducted the study as described above using a within-subject design, where each of the three vocabulary types was tested by five participants.

DESIGNING CONTROLLED STUDIES OUTSIDE OF THE LAB

In some regards, controlled studies conducted outside of the lab are similar to controlled studies inside the lab. Thus, not surprisingly, the steps for designing controlled studies outside of the lab are very similar to those for controlled studies inside of the lab. The process begins with defining the research question. This involves formulating general questions of inquiry (*e.g.,* "Can a learner implicitly acquire second language vocabulary through her interactions with her mobile phone?") which are later refined into a testable hypothesis (*e.g.,* "A user will implicitly acquire a larger second language vocabulary through her interactions with her mobile phone when the words are contextually relevant"). Evaluators must then identify the dependent variables (or measures collected in the study) and the independent variables (or study conditions that the evaluators will deliberately vary during the study). In the case study presented above, the dependent variable is the number of vocabulary words learned by the participants. The independent variable is the contextual relevance of the vocabulary presented to the users. Once the evaluator has identified the independent variables, she must then determine the appropriate levels of treatment in order to test how changing or varying an independent variable affects the dependent variables. In the case study, there were three types of vocabulary: not contextually relevant vocabulary (based on high-frequency words) and contextually relevant vocabularies (words determined based on the venue-type or are place-specific). Based on this information, the

task that participants will be asked to perform multiple times is identified. The task is constructed to allow the evaluators to measure the dependent variables for different levels of treatment of the independent variables.

Controlled studies, when conducted inside the lab, give the evaluators full control over when and where the experiment will take place. Often the laboratory setting is a sterile room, containing only the apparatus that will be used by the study participants and equipment (such as video and audio recording devices, eye motion trackers, *etc.*) for monitoring and recording the participants interaction with the system being tested. This places the focus of the experiment solely on the task being performed by the participant and removes how different contexts may affect what is being studied. The laboratory may sometimes also be instrumented as a particular type of environment to allow for the testing of a system in a specific context. However, it is difficult to develop a laboratory study which tests a system for several different contexts. Furthermore, the laboratory setting will not be able to completely recreate real-life situations. The artificiality of the lab can affect of participant's behaviour. Additionally, researchers have shown that the choice of study environment can potentially bias participant behaviours as well [3, 4]. Thus, the value of conducting a controlled study outside of the lab lies in its ability to examine participant's behaviour in many real environments.

However, as part of designing the study, the evaluators must consider whether they will have a fully implemented and fully functioning system that participants will test with or not. Participants would be able to test a fully functioning system in any scenario. However, a partially functioning one may only work in a limited fashion, thereby limiting the number of scenarios in which participants will be able to use and test it. Thus, when considering how functional the system is, evaluators must decide the scenarios in which their system will be tested. Additionally, often controlled studies outside of the lab are used to collect user performance data (*e.g.,* the speed at which a participant performs a task) with a system and deployment studies are used to collect behavioural data (*e.g.,* when and why a participant uses the system).

CONTROLLED STUDIES OUTSIDE THE LABS FILL THE VOID BETWEEN IN-THE-LAB EXPERIMENTS AND DEPLOYMENT STUDIES

Evaluators must design the procedures of the controlled studies to enable all participants to use the system for the same amount of time to enable the analysis of how varying an independent variable affects the dependent variables being measured. In the example case study, the system was only partially functional. This limited the number of locations in which the system could be tested. We designed the study as a five-session study; during each session, participants received errands on their mobile devices that took them to six different businesses (in a similar manner as if they were asked to do so while out shopping by friends and family members). While the different businesses differed each session, the types of locations were the same across the sessions. This procedure ensured that all participants had the same number of opportunities to interact with the Vocabulary Wallpaper system over the same number of locations. This allowed us to analyze how much the system helps people implicitly acquire a second language vocabulary through microlearning moments if different types of vocabulary (contextually relevant vs. not contextually relevant words) were shown to them.

SUMMARY

In this article, we discuss when and how to conduct controlled studies of mobile or ubiquitous computing systems outside of the lab. Because controlled studies conducted outside of the lab are similar to controlled studies inside the lab, the general steps for both are similar to one another. In studies outside of the lab, how functional the system is and what scenarios it can work in should be taken into account in the study design. Evaluators must design the procedures of the controlled studies to enable all participants to use the system for the necessary amount of time to allow for the analysis of how varying an independent variable affects the dependent variables being measured.

Although controlled studies outside of the lab allow evaluators to test a system in real settings, they mostly provide an understanding about the system that helps user performance of a task in specific situations. While results from these studies have greater external validity than those from studies conducted in the lab, they can only provide answers to specific types of research questions. They do not provide an understanding about when and why the user would use the system in those situations and beyond. ■

Khai Truong is an associate professor in the Department of Computer Science at the University of Toronto. His research interests include human-computer interaction and mobile and ubiquitous computing. Specifically, he focuses on the design and development of ubicomp applications and evaluation of these systems' impact on daily life.

REFERENCES

[1] Dearman, D., and Truong, K.N. "Evaluating the implicit acquisition of second language vocabulary using a live wallpaper." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012.

[2] Dearman, D., and Truong, K.N. "Identifying the activities supported by locations with community-authored content." Proceedings of the 12th ACM international conference on Ubiquitous computing. ACM, 2010.

[3] Sotirakopoulos, A., Hawkey, K., and Beznosov, K. "I did it because I trusted you: Challenges with the Study Environment Biasing Participant Behaviours." Proceedings of SOUPS 2010 Usable Security Experiment Reports (USER) Workshop, 2010.

[4] Sunshine, J., Egelman, S., Almuhimedi, H., Atri, N., and Cranor, L.F. "Crying Wolf: An Empirical Study of SSL Warning Effectiveness." Proceedings of 18th USENIX Security Symposium, pages 399–432, 2009.