Khai Truong *University of Toronto* **Julie Kientz** *University of Washington*

Editor: Khai N. Truong

DEPLOYMENT STUDY LENGTH:

How Long Should a System Be Evaluated in the Wild?

There are many different ways to evaluate a novel interactive system. However, placing that system into the hands of real users and allowing them to use it as they would like in their natural environments may arguably be the best approach to understand if it "really" works. This is because findings learned from user studies conducted in the lab or a controlled setting are limited in external validity and therefore might not generalize beyond the studied usage scenario. Furthermore, the scenario used in a controlled study often lacks full authenticity, and thus it may not faithfully represent situations from the users' lives.

The issues described above point to the need for evaluators to examine how a system would perform in a variety of scenarios that match with when, where, and why the system would or would not be used in real life. To do this well would require the evaluators to first accurately determine different scenarios that would need to be tested with users. However, instead of placing this challenge upon themselves, evaluators can test the system by simply providing it to the target users. The users can be left to interact with the system however they would see it fitting into their everyday lives. This evaluation approach is often referred to as a *deployment study*[1].

Deployment studies have, as a benefit, the ability to collect ecologically valid data about the system. The data can provide insights such as if users would actually adopt the system and whether people will continue to use the system over time. They could also just as easily show that the system is a technology that people will ultimately abandon or never initially liked. Deployment studies can be used to help identify how users incorporate the system into their lives, revealing the different situations when they would use the system, along with when and why they reject the technology. This user evaluation approach also allows evaluators to understand how people actually interact with and use the system in different situations, as well as how well the system helps the users in these scenarios. These findings allow researchers and designers to understand the benefits that the system offers to users, when they would really require these capabilities, the shortcomings of the system, and how to improve it.

In this article, we discuss how to design a deployment study to learn information about a system. We discuss how the length of a study is dependent on a number of factors. We outline the types of information

Illustration, istockphoto.com

[1] Designers and researchers may also refer to variants of this approach as in-the-wild evaluations and field tests. Because it is possible to perform an evaluation of a system outside of the lab (i.e., "in the wild") but in a controlled manner and to field test a product in a limited number of scenarios, we choose to use the term "deployment studies" in this article instead of these other names to call attention to the fact that the system is deployed with users for them to use as they see fit in different situations in their own lives.

that can be learned over the "*length*" of a study. Additionally, we will point out some of the common challenges that often arise over the course of a deployment study.

USER EXPERIENCE WITH A SYSTEM OVER TIME

Karapanos [4] has previously suggested there are three phases that users experience with an interactive system: orientation, incorporation, and identification. Different aspects of the system matter to the users in each of these phases. During orientation, users are still familiarizing themselves with the system and thus the learnability and usability of the system matter the most then. After people are familiar with the basic functionalities of the system, they can then determine how to incorporate and use the technology in their lives. Long-term usability issues that affect the continued use and adoption of the system will become apparent during this phase. Finally during identification, the product has taken on specific value and meaning to the users, and they will develop an attachment with the system.

Ideally, if the user experience is comprised of these distinct phases, researchers and designers can potentially identify phases that they must reach to collect specific pieces of information about a system. However, although the three phases described by Karapanos can guide our understanding of what happens generally, the user experience continuously builds upon itself in practice. In particular, after users become familiar with a system (orientation), they are in a constant cycle of usage (incorporation) and assessment of when, why, and how the system fits into their lives and their practices (identification). They transition from being novice users of the system to knowledgeable or experienced users and then potentially experts. This is because as they use a system, they will learn more about the tool. They will develop an understanding of the effort and cost that is needed to use that system as well as the benefits and values that they receive from it. This will likely then affect how people perceive and value the tool and cause many to reassess how they might use and depend on the system in the future. For example, many studies have shown that users will initially wear fitness tracking devices such as the Fitbit regularly to help them learn how many steps they have taken and use that information to motivate them to reach an ideal daily goal. However, after some time, users will begin to develop an understanding of how many steps are taken when they walk different routes [1]. Because people's lives are often driven by routines, after they have learned this information, they no longer need to use the device in the same way.

LENGTH OF A DEPLOYMENT STUDY

Because the user experience in practice does not break into distinct phases, a key challenge with conducting a deployment study is determining the proper length. What is assumed to happen over the course of a deployment study is that users are interacting with the system. More accurately, they will have some number of opportunities to interact with and use the system. It is this **number of interaction opportunities** that is valuable to consider in the study design. A large number of opportunities to interact with the system is typically needed early to allow users to become familiar enough with the system so that they begin to adopt some uses of the technology in their lives. Thus, if users are potentially exposed to only a small number of opportunities to interact with the system on a daily basis, then arguably the study would require many days just to uncover the initial learnability issues and possibly longer to learn how users adopt and use the system or how it performs in real practice.

As well, along with the number of interaction opportunities, researchers and designers must also take into consideration the following issues to determine the appropriate length of the study:

- How many **different use contexts** would be experienced within a certain period of time?
- What is the **frequency at which people are actually expected to use the system?**
- What is its **"shelf-life?"**

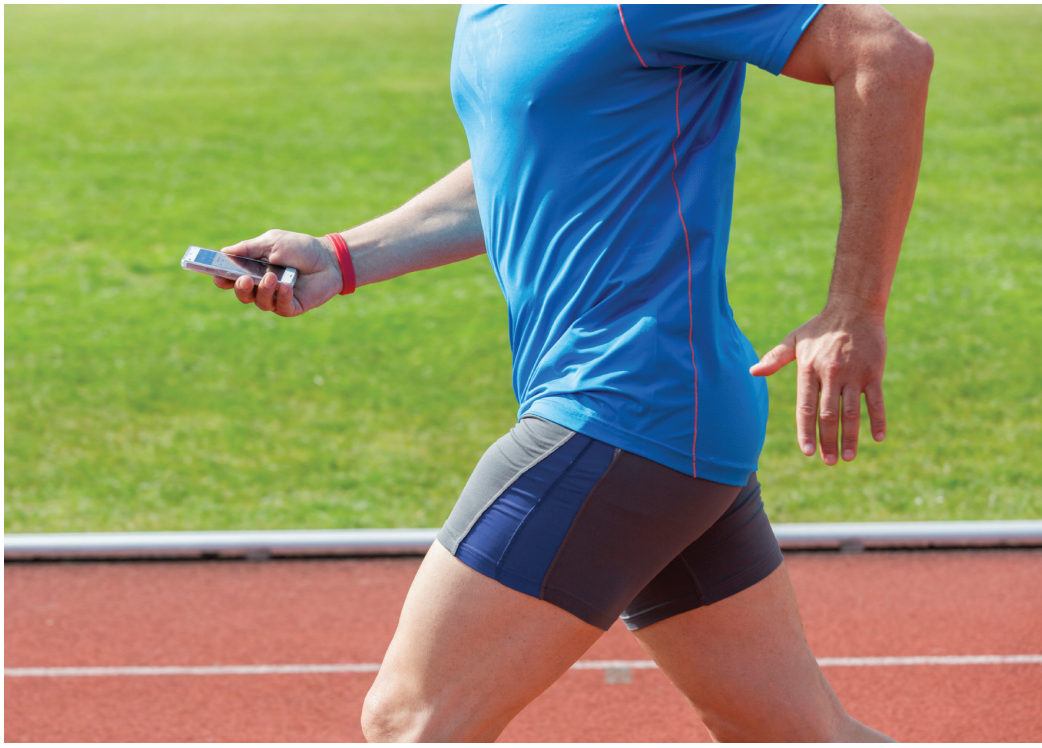
As we previously mentioned, an important reason for conducting a deployment study is that it allows researchers and designers to learn how a system would be used in real life. What are the typical situations in which people will use the system or not? To answer such a question requires that the study be sufficiently long that it captures regularity in the users' lives. During this period, it would then be possible for the different usage contexts to repeat, allowing researchers and designers to determine if and how a system is used in these particular situations. One common mistake often made is not taking into consideration how these contexts play out in their users' lives. Take for example a deployment study that is conducted for three days only. This assumes that any three days in the users' lives are comparable to one another. In reality, however, people's schedules on the weekends are very different from their schedules during the week. Even during the week, one's routine on a Monday might not be similar to her routine for the next day. Thus, if researchers and designers want to understand how the system would be used at any time, they might need to conduct the study to include

all days of the week. Furthermore, they might need to conduct the study for more than one week to include enough repetition of specific contexts that are unique to particular days of the week.

Users might interact with some interactive systems frequently on a regular basis (for example, a system which helps to coach a dementia patient complete activities of daily living). On the other hand, there are some systems that would be used infrequently or only when needed (for example, a system that helps users locate lost objects). Thus, how often people might use a system should be taken into consideration when determining the length of the study as well. Researchers and designers must conduct a long enough deployment study to collect a large number of actual uses of their system.

Finally, how long a system continues to be usable or useful to users before it may ultimately be abandoned is a factor that can also influence the length of a deployment study. We refer to this factor as the system's intended shelf-life (drawing analogy to the shelf life of grocery items or how long people may keep those items in their pantry before they can no longer use those products). Some video games, for example, have what is referred to as "hours of gameplay" which describes how long a typical player takes to complete the game. While the user is still trying to complete a game, there might be heavy usage. However, typically, what happens after a person has finished with a video game is that game may not be played again at the same regularity (and likely not played in the same way) after its shelf-life period has elapsed.

Although we discuss the importance of using the above factors to determine an appropriate length for a deployment study, we note this does not mean that data collected from a deployment study that is shorter or longer than this length is invalid. Though the data can still potentially be valid, conducting a study that is shorter than the ideal length means that the data collected by researchers and designers may not be what they want to learn. On the other hand, conducting a study that is longer than the ideal length can cost researchers and designers additional time and resources that would not otherwise have needed to be spent.



Photo, istockphoto.com

BECAUSE THE USER EXPERIENCE IN PRACTICE DOES NOT BREAK INTO DISTINCT PHASES, A KEY CHALLENGE WITH CONDUCTING A DEPLOYMENT STUDY IS DETERMINING THE PROPER LENGTH

DATA FROM A DEPLOYMENT STUDY

When the user is new to a system, she typically must learn how to use it. Such a user often exhibits novice user behaviors. Typical of novice behaviors with a system are the mistakes and errors that an inexperienced user will make when using a system. This is because the user is still learning how to operate the technology. Furthermore, the user may experiment with using the system in many different situations to see how the system reacts. Some of these situations may not be appropriate use contexts. This helps the user learn when the system can or cannot be used. Thus, data collected at this stage of the deployment study may contain information about how users initially learn and develop an understanding about the system. It can also potentially reveal what features of the system initially motivate the users to try the system. Note that during this time, however, the way that users interact with the system may not necessarily represent how they may use the system later when they have a more complete understanding of the system. Furthermore, when a system is new to the users, they may be excited about the technology and may interact with the system simply because it is new. This is known as a *novelty effect*, which is the initial change to user behavior when new technology is introduced. During the novelty period, users will attempt to test a system in situations in which they would not actually use it in the future. Thus, data collected is not often trusted as an accurate reflection of how the system would be used eventually.

After some time, the user develops a familiarity with the system and a greater understanding of how to use the system. At this point, the user has then transitioned into being a knowledgeable or experienced user. An indication that the user is no longer a novice is when she no longer makes a large number of mistakes and errors with respect to when and how she uses the system. A knowledgeable or experienced user will typically have developed usage strategies that allow her to be efficient when interacting with the system in different situations. Data collected from knowledgeable or experienced users will provide insight about how a system might be used on a normal basis. Data collected can also be used to determine how well the system performs and helps the average users during real activities. It is important to note that sometimes usage patterns will continue to change over time. Thus, extending the length of a deployment study at this time will allow researchers and designers to learn about the wide range of contexts when and why people use the system as well as when and why they do not. Additionally, it will provide an understanding of the values and perceptions that users have attached to the technology, along with why and how those change. This includes reasons such as why users adopt a specific usage of the system, if they continue to use the system over time, and potentially why some users eventually abandon it (if the study is long enough).

Through continued use of a system, some people may eventually become expert users. Indications that a user has become an expert include an extensive knowledge of the system and the ability to use it in ways that the average user typically does not. Data collected with such users can show how well the system performs when it is used by people who are proficient with it. It will provide an understanding of when, why, and how experts use the system.

Although we have discussed the data that can be collected to be dependent on the length of a study, the reverse is also applicable. That is, the information about a system that researchers and designers want to learn can also be used to determine the length of the study.

SUMMARY

In this article, we have discussed how to use the approach of deploying a system with real users and allowing them to interact with it how they choose to enable researchers and designers to collect information about their system in real use. We describe what type of information can be collected from a deployment study and discuss how the length of a study should be dependent on:

- the number of interaction opportunities that users have with a system during that period of time
- the different use contexts that can be sampled within that period
- the frequency at which people are expected to use the system that is being evaluated
- the system's intended shelf-life
- the information that researchers and designers wish to learn about the system.

For example, the length of the study should be selected to provide users with ample opportunities to interact with the system. If the system is intended to be infrequently interacted with, then the study should be sufficiently long enough to ensure that the desired number of usage instances is being sampled.

While we are able to describe these considerations at a general level, more research is needed about deployment studies before we can recommend a specific study design based on these considerations. For example, studies of response rates to different survey approaches have allowed researchers to determine how large of a population must be targeted to obtain a desired number of responses [3]. Similar research is needed about deployment studies to help provide answers to questions such as:

- on average, what is the uptake rate of the opportunities to interact with a system by novice users vs. experience users, *etc.*?
- how long should a study be given a particular interaction frequency value intended for a system?
- how does the length of a study affect the quality of feedback provided by the users?

Similar to how the stages of behavior change in the Transtheoretical Model can be identified using validated survey questions [2], research is also needed on ways to determine which stage of the user experience a person is in as a way of assessing if a deployment study has been conducted long enough. Additionally, in some context, it might be possible to determine how much time is needed to understand whether a system has any effect on the users. For example, sleep researchers have determined that a two-week window can provide an accurate picture of people's daily sleep habits [5]. Thus, researchers could also study and provide recommendations for deployment study windows required to examine specific effects or to collect specific types of feedback from users.

Until research on the deployment study approach matures more, our best advice is to be aware that particular types of information can be collected and reported at a different stages of the evaluation (and more specifically once users reach a particular experience level with the system). Thus, it is important to understand and determine if there are indications that a particular user experience level or evaluation stage has been reached before collecting that type of information and reporting it. To be on the safe side, conducting a deployment study longer than the ideal length can help to ensure researchers and designers are able to learn what they want to know about their system. ■

REFERENCES

[1] Fritz, T., Huang, E. M., Murphy, G. C., & Zimmermann, T. (2014, April). Persuasive technology in the real world: a study of long-term use of activity sensing devices for fitness. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 487-496). ACM.

[2] Glanz, K., Rimer, B. K., & Viswanath, K. (Eds.). (2008). Health behavior and health education: theory, research, and practice. John Wiley & Sons.

[3] Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A comparison of web and mail survey response rates. *Public opinion quarterly*, 68(1), 94-101.

[4] Karapanos, E. (2013). User experience over time. In *Modeling Users' Experiences with Interactive Systems* (pp. 57-83). Springer Berlin Heidelberg.

[5] Monk, T. H., Reynolds, C. F., Kupfer, D. J., Buysse, D. J., Coble, P. A., Hayes, A. J., Machen, M.A., Petrie, S.R. & Ritenour, A. M. (1994). The Pittsburgh sleep diary. Journal of sleep research, 3(2), 111-120.