Case Study: Longitudinal Comparative Analysis for Analyzing User Behavior

Jhilmil Jain
Senior UX Researcher
Google, Inc.
1600 Amphitheater Pkwy
Mountain View, CA 94043 USA
jhilmil.jain@gmail.com

Susan Boyce
Principal UX Lead
Microsoft
1310 Villa Street
Mountain View, CA 94041
suboyce@microsoft.com

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract
In this case study we describe a four-step process for eliciting and analyzing user behavior with products over an extended period of time. We used this methodology for conducting a comparative study of two mobile applications over a period of seven months with 17 participants. To focus the discussion, we are concentrating on the methodology rather than the results of the study.

Keywords
Qualitative longitudinal data, interview, diary, sketching, retrospective reconstruction, survey

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

Introduction
As user experience issues become more central to HCI, the value of longitudinal research—collecting user data over time—is increasingly recognized. UX researchers understand the importance of observing extended use of products and systems, and seek to improve methodology and develop best practices for longitudinal research.

Traditional user research and evaluation methods tend to focus on 'first-time' experiences with products [3], which trends the results more towards discoverability

or learnability problems, rather than usability concerns that may persist over time. This case study extends current thinking by providing a four-phase methodology that has proven effective for longitudinal data collection and analysis.

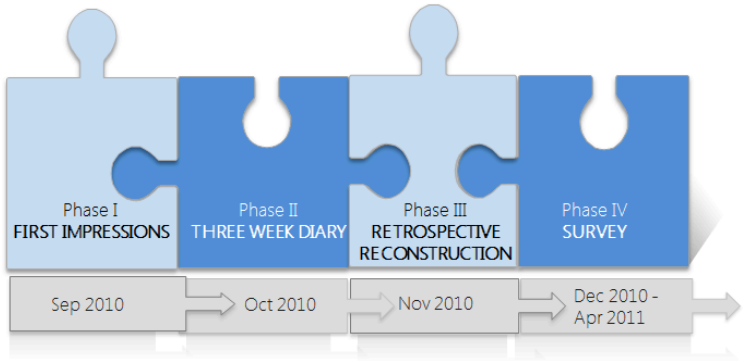

figure 1. four step process for longitudinal data elicitation

In this study, we wanted to uncover first impressions of two mobile applications (we will refer to them as Application 1, and Application 2) and then map how these impressions changed with sustained use. The applications were commercially available to be downloaded from the app store, and performed functions like search and recommendation.

Phase 1: First Impressions

Our goal was to determine the first impressions for both the applications and to study if and how behavior changes over time. Thus a standard usability test helped level set the observation.

This was a lab study where each session was 45 mins long. Participants were asked to download and interact with both applications in the lab. None of the participants were familiar with the applications prior to the study, and therefore they were able to truly provide their first impressions.

The result of the usability test was that almost all participants (16/17) preferred Application 2, and expressed the desire to use it in the next phases of the study.

Phase 2: Diary Study

In phase 2, participants were asked to keep a diary of their interactions with both applications. Phase 2 lasted for 3 weeks, and participants were instructed as follows:

- week 1: use both apps. Complete a total of 4 tasks each day (2 tasks using each application). Tasks could be different for each application, in fact we encouraged this.

- week 2: choose an app and use only this for the whole week. Complete at least 2 tasks each day.

- week 3: continue with the app selected in week 2, or switch if you were unhappy with the choice made in week 2. Complete as many tasks as you like this week.

We did not prescribe the tasks ahead of time; rather we wanted participants to use the applications for tasks that would occur in their daily life. Using this approach,

we were able to identify aspects for which participants preferred one application over the other.

Participants sent us their diary entries via email. They were provided the following structure (see fig 2) to record their interaction with the apps. They were also encouraged to capture screen shots of their interactions with the applications.

Based on a postmortem discussion with participants at the end of Phase 2, *providing a structure for the dairy entries* set expectation of the type of feedback we were expecting. In the first few days, we *provided feedback to each participant* via email to coach them on how to write a good diary entry; and *sent examples of "good" and "weak" entries*. Both of these techniques dramatically improved the quality of the diaries. We also sent *frequent reminders* throughout the 3 weeks.

Phase 3: Retrospective Reconstruction

In phase 3, we conducted an in-person interview with each participant in the lab. The goal was to use the diary artifacts to get a sense of the apps they continued using, tasks they conduct etc.

We started the interview by using the repertory grid technique to elicit differences between the 2 apps. Participants were provided screenshots of the 2 apps and then asked to "think of a property or quality that makes the apps alike or different". The goal was to assess people's mental models of the two apps by encouraging them to talk about the two side by side.

Participants were then asked to sketch their experiences with the apps using the day reconstruction technique. "Constructive theory" model was used to reconstruct their experiences [8]. Using the constructive approach, participants were asked to recall episodic information in chronological order (days/weeks) that will help them reconstruct emotion experiences with the apps. The diary entries (and screen shots) captured in the past 3 weeks was used as "contextual cues" to help them jog their memories. The goal was to elicit temporal context of the recalled experiences.

1. BACKGROUND

Spend about 30% of your writing effort here. Help us recreate the scene with lots of details help, e.g.,

- what information did you need? what were you doing? why?
- which app did you pick? Why?

2. MY INTERACTION WITH THE APP

Spend 40% of your writing effort here. Almost a play-by-play of what happened, e.g., did you type, use voice?

- what did you say or type?
- how did the app respond?
- did you get what you needed? Did you have to do multiple searches?

3. MY OBSERVATIONS

Spend 30% of your writing effort here. Go deep. share your reflections on that interaction, e.g.,

- how did it go? how did it make you feel?
- anything surprising or unexpected?
- If the app failed to get you the answer, what else did you use?

figure 2. structure of the diary entry

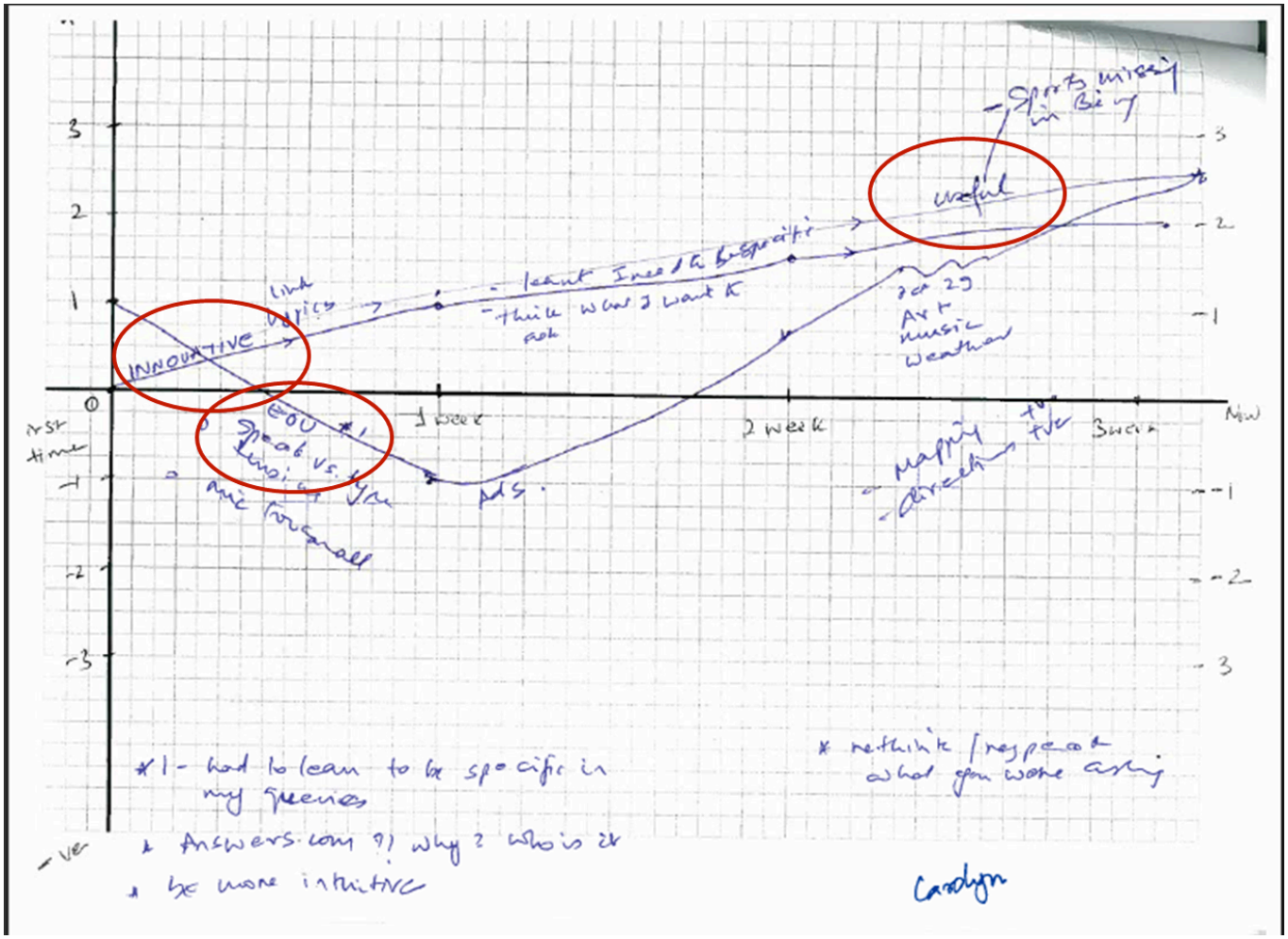

figure 3: sample retrospective reconstruction sketch – participants asked to sketch 3 curves: innovativeness, ease of use, usefulness. The Y-axis was the rating scale from -3 to +3. X-axis was the timeline first impression, week 1, week 2, week 3

For each reported episodic incident, they were asked to rate their experience on the following 3 dimensions, and plot it on the graph (see figure 3):

[1] *Usefulness*: the ability of the app to provide the necessary functions for a given task. Related words – practical, meaningful
[2] *Ease-of-use*: the ability of the app to provide the functions in an easy and efficient way. Related words – simple, clear
[3] *Innovativeness:* the ability of the app to excite the user through its novelty. Related words: exciting, creative, magical, delightful

The goal was to break down the user experience on three dimensions: usefulness, usability and innovativeness to see a correlation between them (we will not be discussing the correlation between the factors in this case study).

Phase 4: Follow on survey

After another four months of use, we asked the participants to fill a short email survey to determine which app they have continued to use and why? We again asked them rate each application on usefulness, ease of use and innovativeness using the same scale. Note: in the four month period, the survey was sent twice, so we asked for their feedback at 3 months and then at 7 months of use.

Results

Over the course of this study we analyzed over 30 hours of recorded interviews, 500 pages of diaries and screen shots submitted by the participants, and the answers to the final surveys.

The results of the longitudinal study were astonishing. Over the period of 3 weeks, approximately 60% of the participants converted from favoring Application B to Application A. By the end of the study, 25% were using Application A, 25% were using Application B, 17% were using both, and the rest of the participants had abandoned both apps.

The participants' ultimate impressions of the applications differed markedly from their first impressions, lending further evidence that longitudinal study, paired with more traditional usability testing is essential in evaluating product usability and usefulness.

Risks associated with longitudinal research

Having conducted multiple longitudinal studies, we find two key issues arise when you propose such an approach:

1. *Push back from management*: We hear things like "the process takes too long (or too expensive), can't you just do a usability study". To address this issue, we first started with a typical usability test to get data out to the product teams instead of waiting until the end of the study to report on the results. Additionally, each week during the diary study, we send informal readouts to the various teams to highlight emerging trends to keep them engaged.

2. *Retention of participants:* While recruiting participants, we had an agreement that they would be paid only after they complete the entire study. If they chose to drop in between

the study they would not be paid since we would have incomplete data.

Previous related work

Dr. Jain has been involved in conducting a series of five events to build a body of knowledge about longitudinal research practice at CHI and UPA conferences.

After the CHI 2007 SIG [4], we established a wiki on longitudinal research [6]. The goals of the wiki are to share best practices, case studies, and lessons learned about longitudinal data collection and analysis. At CHI 2008 [5], we had organized a panel where researchers from industry and academia gave their viewpoints and case studies. Then, to provide a better venue for in-depth discussion, we conducted a workshop at CHI 2009 [2], where participants discussed the open issues raised at the SIG and panel, as well as their individual goals for longitudinal research. Prior to the CHI 2009 workshop, the authors also conducted a workshop at the 2008 Usability Professionals' Association conference [1].

We generated alternative definitions of "longitudinal research," prioritized over 30 questions of interest, and began developing best practices.

At CHI 2010, we had a lively SIG [7] where the need for a methodological treatment for longitudinal studies by the CHI community was echoed. In particular, practitioners expressed a gap in techniques/tools that they could use for comparative data analysis of both quantitative and qualitative data. Thus we felt the need to explore a structured approach to do the same.

Acknowledgements

We thank all Speech@Microsoft design studio members that helped with the longitudinal study data collection and analysis, and all the participants for their time and enthusiasm with the project.

References

[1] Courage C., Rosenbaum, S., and Jain. J. *Exploring Best Practices in Longitudinal Usability Studies*. UPA Workshop.https://www.usabilityprofessionals.org/upa_conference/program/2008/activity.php?id=9220

[2] Courage C., Jain. J, and Rosenbaum, S. *Best Practices in Longitudinal Research*, Ext. Abstracts CHI 2009. ACM,New York,NY.

[3] Mendoza, V., and Novick, D. G. *Usability over time*. In ACM 23rd International Conference on Computer Documentation, ACM Press (2005), 151-158.

[4] Vaughan, M. and Courage, C. *Capturing longitudinal usability: what really affects user performance over time?.* Ext. Abstracts CHI 2007, ACM Press (2007), 2261-2264.

[5] Vaughan, M., Courage, C., Rosenbaum, S., Jain, J. Hammontree, M., Beale, R., and Welsh, D. *Longitudinal usability data collection: art versus science?.* Ext. Abstracts CHI 2008, ACM Press (2008), 2149 – 2152.

[6] Wiki on Longitudinal Research: http://longitudinalusability.wikispaces.com/

[7] J. Jain, S. Rosenbaum, and C. Courage, Best practices in longitudinal research, New York, New York, USA: ACM Press, 2010

[8] Karapanos, E., Martens, J.-B., Hassenzahl, M. (2010) On the Retrospective Assessment of Users' Experiences Over Time: Memory or Actuality?**.** *CHI'10 extended abstracts on Human factors in computing systems. Atlanta*, ACM Press.

Authors

Jhilmil Jain is a Sr. UX Researcher at Google. Prior to that she was a Sr. UX Strategist at Microsoft, and before that she was a Sr. UX Lead at HP Labs. At Microsoft, she led user research and usability efforts for the speech@microsoft product group. At HP Labs she led UX efforts for multiple business incubations. This case study details the work that was done while she was at Microsoft. She has several publications and patents in information visualization, user research, multimodal interaction modeling, personal information management systems, and experimental evaluation. She has served as the program chair for CHIMIT 2009; on the program committees of various conferences such as CHI, HCII, and UPA; on the editorial board for the International Journal of Handheld Computing Research; on the review boards for two books "Handheld Computing for Mobile Commerce" and "The Psychology of Facebook"; and is currently serving a third term as the UX community chair for CHI 2012. She is a member of ACM, UPA, Phi Kappa Phi, and Upsilon Pi Epsilon.

Susan Boyce is a Principal User Experience lead at Microsoft. Susan has spent more than 20 years researching and designing speech recognition technology. She holds a Ph.D. in Cognitive Psychology and has worked at Bell Labs, AT&T Labs, several startups and most recently Tellme, prior to Microsoft. She's published numerous articles, chapters and given talks in the general area of designing speech recognition systems.