On the Retrospective Assessment of Users' Experiences Over Time: Memory or Actuality?

Evangelos Karapanos
Eindhoven University of Technology
P.O. Box 513, 5600 MB, Eindhoven
The Netherlands
E.Karapanos@tue.nl

Jean-Bernard Martens
Eindhoven University of Technology
P.O. Box 513, 5600 MB,
Eindhoven
The Netherlands
J.B.O.S.Martens@tue.nl

Marc Hassenzahl
Folkwang University
Universitätsstraße 12, 45117 Essen
Germany
marc.hassenzahl@folkwang-hochschule.de

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract

An alternative paradigm to longitudinal studies of user experience is proposed. We illustrate this paradigm through a number of recent tool-based methods. We conclude by raising a number of challenges that we need to address in order to establish this paradigm as a fruitful alternative to longitudinal studies.

Keywords

User experience evaluation, longitudinal methods, retrospective techniques, day reconstruction, experience sampling

ACM Classification Keywords

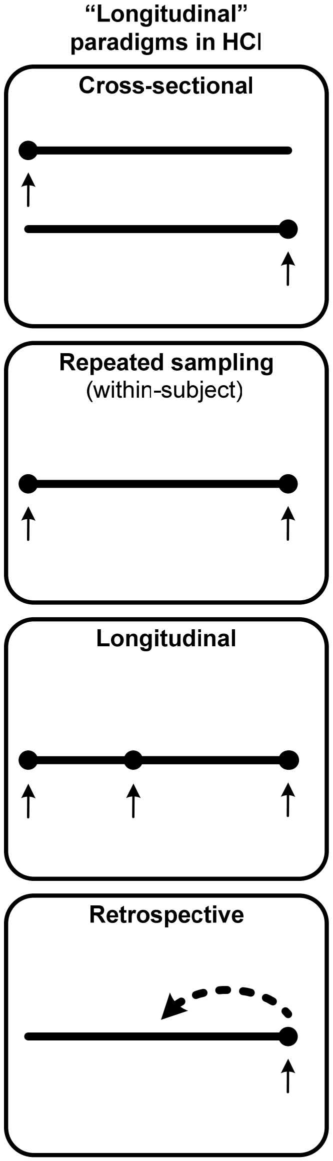
H5.2. User Interfaces: Evaluation/methodology.

General Terms

Human Factors, Measurement, Theory

Introduction

Human-Computer Interaction (HCI) and User Experience in consequence have traditionally been interested in users' initial interactions with products. A growing



interest is however attributed to the study of users' experiences over prolonged use [3-5, 11, 13, 15].

We [12] distinguished three dominant methodological paradigms in understanding the change in users' experience and behavior over time.

Cross-sectional approaches are the most popular in the HCI domain [e.g. 18]. Such studies distinguish user groups of different levels of expertise, e.g. novice and expert users, or different lengths of ownership of a product. Such approaches are limited as one may fail to control for external variation and may falsely attribute variation across the different user groups to the manipulated variable. Prümper et al. [18] already highlighted this problem, by showing that different definitions of novice and expert users lead to varying results.

Within-subject repeated sampling (pre-post) designs study the same participants at two points in time. For instance, Kjeldskov et al. [15] studied the same seven nurses, using a healthcare system, right after the system was introduced to the organization and 15 months later, while we [11] studied how 10 individuals formed overall evaluative judgments of a novel pointing device, during the first week of use as well as after four weeks of using the product. While these studies inquire into the same participants over an extended period of time, one may not readily infer time effects as these might be due to random contextual variation, given that we have only two measurements.

Longitudinal designs take more than two measurements and, thus, enable greater insight into the exact form of change. For instance, Minge [16] elicited judgments of perceived usability, innovativeness and the overall attractiveness of computer-based simulations of a digital audio player at three distinct points: a) after participants had seen but not interacted with the product, b) after 2 minutes of interaction and c) after 15 minutes of interaction. In [13], we followed 6 individuals after the purchase of a single product over the course of 5 weeks. One week before the purchase of the product, participants started reporting their expectations. After product purchase, during each day, participants were asked to narrate the three most impactful experiences of the day.

While longitudinal studies are considered as the gold standard in studying changes in users' behavior and experiences over time, they are increasingly laborious when one needs to generalize over large populations of users and products.

A fourth paradigm, aiming at providing a lightweight alternative to longitudinal studies, relies on the retrospective elicitation of users' experiences from memory. Participants may be asked, within a single contact, to recall the most salient experiences they had within a given time period, and provide estimated temporal details regarding the recalled experiences. In the remainder of the paper we review existing approaches to the retrospective assessment of temporal dynamics of experience and outline some of the challenges that we need to address in order to establish this paradigm as a fruitful alternative to longitudinal studies. Through this position paper, we attempt to generate some discussion on this avenue of research.

A review of retrospective methods

One of the most popular retrospective techniques is the Day Reconstruction Method (DRM) proposed by Daniel

Kahneman and colleagues [10] as an alternative to the established Experience Sampling Method (ESM) [8]. DRM asks participants to reconstruct, in forward chronological order, all experiences that took place in the previous day. Each experience is thus reconstructed within a temporal context. Kahneman et al. [10] showed that DRM provides a surprisingly good approximation to Experience Sampling data, while providing the benefits of a retrospective method. While DRM has not been proposed as a longitudinal method, it may well be applied as such. In [13], we employed DRM in an ethnographic study of users' experiences with iPhone during the first 4 weeks of ownership. Khan [14] proposed a tool that combines ESM and DRM in a longitudinal setting.

Can reconstruction be extended to longer periods of time (e.g. full time of ownership of a product)? Von Willamowitz et al. [20] proposed a structured interview technique named CORPUS (Change Oriented analysis of the Relation between Product and User). CORPUS starts by asking participants to compare their current opinion on a given product quality (e.g. ease-of-use) to the one they had right after purchasing the product. If change has occurred, participants are asked to assess the direction and shape of change (e.g., accelerated improvement, steady deterioration). Finally, participants are asked to elaborate on the reasons that induced these changes in the form of short narratives, the so-called "change incidents".

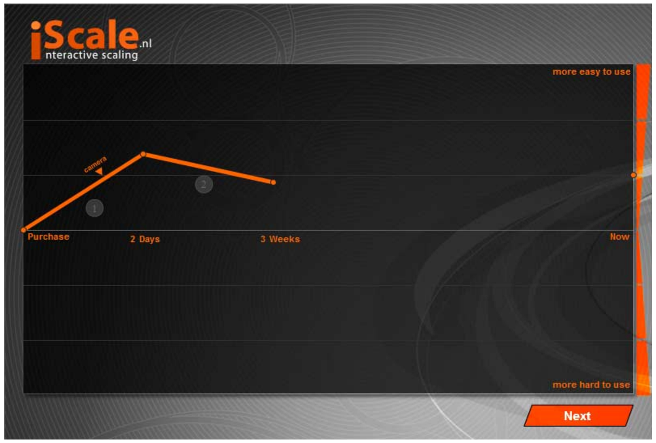


Figure 1. In **iScale** participants are asked to "sketch" how their opinion has changed from the moment of purchase till the present. Each sketch is annotated by the respective time period and participants are asked to narrate one or more experiences that induced this change in their perception of the respective product quality.

One may wonder about the extent to which these data suffer from retrospection biases. One could argue, however, that this issue, the veridicality of reconstructed from memory experiences, is of minimal importance as "these memories (1) will guide future behavior of the individual and (2) will be communicated to others" (Karapanos et al. [12], Norman [17]).

In [12] we argued that while the validity of these experience reports may not be as crucial, their reliability is, in the sense that participants should at least report the same experiences when asked repeatedly to report their most impactful experiences with a product. In other words, what we remember might be different from what we experienced; however, as long as these memories are consistent over multiple recalls, they provide valuable information.

We proposed *iScale* [12], a survey tool that aims at assisting users in reconstructing their experiences with a product (see figure 1). iScale employs sketching in imposing a specific procedure in the reconstruction of one's experiences. We proposed and empirically validated two different versions of iScale, the *Constructive* and the *Value-Account* iScale, each motivated by a distinct theory on how people reconstruct emotional experiences from memory [2, 19]. The Constructive iScale tool imposes a chronological order in the reconstruction of one's experiences. It assumes that "emotional experience can neither be stored nor retrieved" [19, p. 935], but instead is re-constructed on the basis of the recalled contextual details. Moreover, it assumes that chronological reconstruction results in recalling more contextual details surrounding the experienced events [1] and consequently to a more reliable recall of experiential information. The Value-Account iScale tool explicitly distinguishes the elicitation of the two kinds of information: value-charged (e.g. emotional) and contextual details. It assumes that value-charged information can be recalled without recalling concrete contextual details of an experienced event due to the existence of a specific memory structure, called Value-Account, that stores the frequency and intensity of one's responses to stimuli [2]. Overall, the constructive iScale tool was found to outperform the Value-Account iScale tool and to offer a significant improvement in the amount, the richness and the test-retest reliability of recalled information when compared to *free recall*, an approach that does not involve sketching in the recall process [12]. Further, iScale was found to have a number of limitations, but also benefits, over its analog version, *free-hand sketching*.

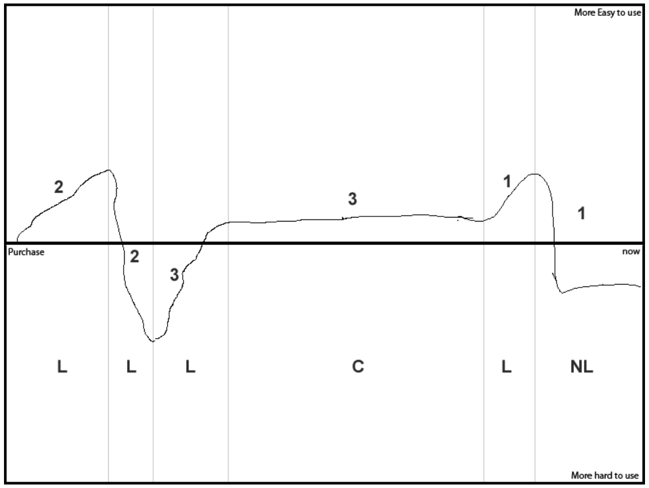


Figure 2. An example of **Free-Hand Sketching** (the analog version of iScale, see [12]). Identified segments are indicated by vertical lines. Each segment is coded for the type of experience report (1: Reporting a discrete experience, 2: Reporting on attitude, reasoning through experience, 3: Reporting on attitude with no further reasoning) and type of sketch (C: Constant, L: Linear, NL: Non-Linear, D: Discontinuous).

Techniques like *CORPUS* and *iScale* provide two kinds of data: a) self-reports of personal experiences that induced the changes in the respective product quality, over time and b) a recalled pattern of change on the perceived product quality. Both techniques emphasize qualitative data; sketches are seen as a way to assist participants in reconstructing their experiences with a product.

Analytic Scale is a lightweight technique for characterizing the temporal development of a judgment analytically. It relies on Value Account Theory [2] that assumes that people may recall an overall emotional assessment of an experience without recalling the exact details of the experienced event. Participants are thus asked in an analytic fashion to characterize the pattern of change (if any). Participants are first asked to judge whether this was improving, deteriorating, or stable. In a second step, they are asked to further distinguish between different patterns, for example, did it improve in the beginning or at the end of the respective period, or was it stable throughout the whole period or did it follow some fluctuation ending approximately at the same value?

Analytic Scale may firstly be employed in eliciting an emotional assessment of an experience (e.g. "how did you feel while using the product?"). This affective information is assumed to be directly accessible through the hypothetical Value-Account memory. Secondly, it may be employed in characterizing the temporal development of overall evaluations (e.g. goodness [6]) or quality judgments (e.g. perceived ease of use) of a product. In these later cases, only affect that is partially attributed to the product is used to reconstruct the temporal development of the judgment [7].

Analytic Scale may be employed in characterizing the temporal development on a macro-scale as in the case of CORPUS and iScale (e.g. 6 months), but also on a micro-scale (e.g. single interaction session). This latter information may complement traditional experiential evaluations in which participants are asked to summarize their experience in a single judgment. Despite their practical value, such reductive evaluations have been shown to reflect biased representations of experiences (e.g. peak-and-end phenomenon [9]). The question is thus: can the assessment of the temporal development of emotion or judgment provide complementary information to the overall amplitude of an experience reflected in a single summative judgment of intensity?

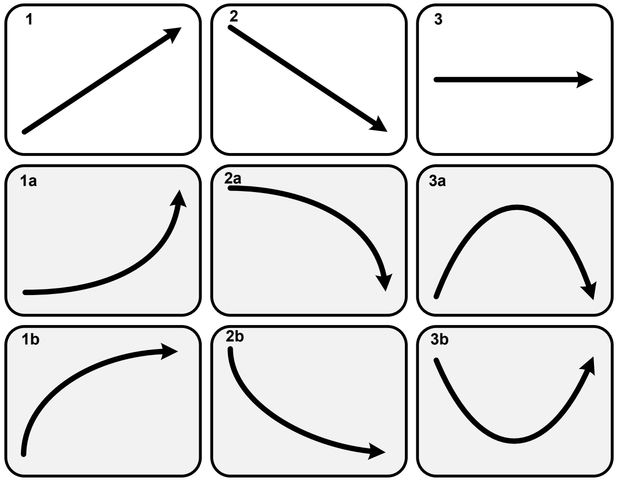


Figure 3. ***Analytic Scale*** asks participants to characterize the temporal development of a judgment analytically. Participants are first asked to judge whether this was improving, deteriorating, or stable. In a second step, they are asked to further distinguish between the three shapes, e.g. improved in the beginning, at the end of the respective period, or linearly (parts 1/1a/1b).

Challenges to retrospective assessment: Memory or actuality?

Perhaps the most crucial question is: how do these memories differ from actual experiences, and, if so, for which contexts are memories more important than actuality?

Kahneman [10] aimed at demonstrating close approximation of retrospective data elicited through the Day Reconstruction Method to experiential data elicited through the Experience Sampling Method. On the contrary, we [12] argued for reconstructing experiences spanning greater lengths of time. They argued that these memories may vary substantially from the actual experiences. Reliability (i.e. consistency over multiple recall trials) is more important than veridicality (i.e. consistency between the memory and the experience), they argued.

iScale was found to provide a substantial improvement in the test-retest reliability of participants' reconstruction process [12]. It thus provides meaningful information. But, what does this information represent, and in what ways is it different from the actual experiences? Are there systematic biases in these recalled experiences? For instance, we found user-perceived time to relate to the actual time through a power-law. Participants were more inclined to recall experiences that took place in the first month of use (75% of all recalled experiences related to the first month while 95% related to the first six months) and participants' sketches reflected a logarithmic scale of time. These questions are yet to be addressed and are crucial in order to establish retrospective techniques as an alternative to longitudinal methods in HCI.

References

[1] Anderson, S.J. and Conway, M.A., Investigating the structure of autobiographical memories. Learning. Memory, 1993. **19**(5): p. 1178-1196.

[2] Betsch, T., Plessner, H., Schwieren, C., and Gutig, R., I like it but I don't know why: A value-account approach to implicit attitude formation. Personality and Social Psychology Bulletin, 2001. **27**(2): p. 242.

[3] Courage, C., Jain, J., and Rosenbaum, S., Best practices in longitudinal research, in *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*. 2009, ACM: Boston, MA, USA.

[4] Fenko, A., Schifferstein, H.N.J., and Hekkert, P., Shifts in sensory dominance between various stages of user–product interactions. Applied Ergonomics, 2010. **41**(1): p. 34-40.

[5] Gerken, J., Bak, P., and Reiterer, H. Longitudinal evaluation methods in human-computer studies and visual analytics. in *Visualization 2007: IEEE Workshop on Metrics for the Evaluation of Visual Analytics*. 2007.

[6] Hassenzahl, M., The interplay of beauty, goodness, and usability in interactive products. Human-Computer Interaction, 2004. **19**(4): p. 319-349.

[7] Hassenzahl, M. and Ullrich, D., To do or not to do: Differences in user experience and retrospective judgements depending on the presence or absence of instrumental goals. Interacting with Computers, 2007. **19**: p. 429-437.

[8] Hektner, J.M., Schmidt, J.A., and Csikszentmihalyi, M., Experience sampling method: Measuring the quality of everyday life. 2007: Sage Publications Inc.

[9] Kahneman, D., Fredrickson, B.L., Schreiber, C.A., and Redelmeier, D.A., When more pain is preferred to less: Adding a better end. Psychological Science, 1993: p. 401-405.

[10] Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., and Stone, A.A., A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. 2004, American Association for the Advancement of Science. p. 1776-1780.

[11] Karapanos, E., Hassenzahl, M., and Martens, J.-B., User experience over time, in *CHI '08 extended abstracts on Human factors in computing systems*. 2008, ACM: Florence, Italy.

[12] Karapanos, E., Martens, J.B., and Hassenzahl, M., Reconstructing Experiences through Sketching. arXiv:0912.5343, 2009.

[13] Karapanos, E., Zimmerman, J., Forlizzi, J., and Martens, J.-B., User Experience Over Time: An initial framework, in *Proceedings of the Twenty-Seventh Annual SIGCHI Conference on Human Factors in Computing Systems - CHI '09*. 2009, ACM: Boston.

[14] Khan, V.J., Markopoulos, P., Eggen, B., Ijsselsteijn, W., and de Ruyter, B. Reconexp: a way to reduce the data loss of the experiencing sampling method. in *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services* 2008: ACM.

[15] Kjeldskov, J., Skov, M.B., and Stage, J., A longitudinal study of usability in health care: Does time heal? International Journal of Medical Informatics, 2008.

[16] Minge, M., Dynamics of User Experience, in *Proceedings of the Nordichi '08 Workshop "Research Goals and Strategies for Studying User Experience and Emotion"*. 2008.

[17] Norman, D.A., THE WAY I SEE IT. Memory is more important than actuality. interactions, 2009. **16**(2): p. 24-26.

[18] Prümper, J., Zapf, D., Brodbeck, F.C., and Frese, M., Some surprising differences between novice and expert errors in computerized office work. Behaviour & Information Technology, 1992. **11**(6): p. 319-328.

[19] Robinson, M.D. and Clore, G.L., Belief and feeling: Evidence for an accessibility model of emotional self-report. Psychological Bulletin, 2002. **128**(6): p. 934-960.

[20] von Wilamowitz Moellendorff, M., Hassenzahl, M., and Platz, A., Dynamics of user experience: How the perceived quality of mobile phones changes over time, in *User Experience - Towards a unified view, Workshop at the 4th Nordic Conference on Human-Computer Interaction*. 2006.