View from a Distance: Comparing Online and Retrospective UX-Evaluations

Ina Wechsung
TU Berlin, Telekom Innovation Laboratories
Ernst-Reuter-Platz 7
10587 Berlin
ina.wechsung@telekom.de

Kathrin Jepsen
Telekom Innovation Laboratories
Ernst-Reuter-Platz 7
10587 Berlin
kathrin.jepsen@telekom.de

Felix Burkhardt
Telekom Innovation Laboratories
Ernst-Reuter-Platz 7
10587 Berlin
felix.burkhardt@telekom.de

Annerose Köhler
Telekom Innovation Laboratories
Ernst-Reuter-Platz 7
Annerose.Koehler@t-systems.com

Robert Schleicher
TU Berlin, Telekom Innovation Laboratories
Ernst-Reuter-Platz 7
10587 Berlin
robert.schleicher@telekom.de

MobileHCI'12, September 21–24, 2012, San Francisco, CA, USA.
ACM 978-1-4503-1443-5/12/09.

Abstract
This paper reports the results of an explorative field study investigating if and how remembered UX evaluations differ from evaluations collected during usage. Results show that while quantitative ratings are similar, qualitative data differs: Comments assessed during usage were less detailed but contained more affective evaluations compared to the retrospective remarks.

Author Keywords
User experience; field study; methodology; evaluation

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces

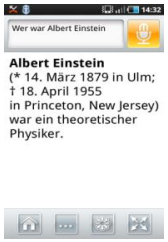
Introduction
User experience (UX) is widely understood as a dynamic construct changing over time [1]. Accordingly, evaluation comprising only short periods of usage cannot assess this temporal aspect adequately. But as pointed out by a recent review of UX methods, longitudinal studies are rather rare [2]. Furthermore evaluations are often conducted subsequent to usage and not during usage. This is insofar critical as findings from cognitive psychology indicate that retrospective evaluations of experiences may not be the sum of the experienced

AskWiki

With AskWiki users can access Wikipedia content via speech control with a specific question. In contrast to the official Wikipedia app by the Wikimedia Foundation, AskWiki does not present the whole article but the precise answer to a specific question. E.g. the question "What is the capital of California?" will result in the answer "Capital of California is Sacramento". If the users want additional information, they can request more (including pictures).

(a)

(b)



(c)

AskWiki GUI:
(a) start screen
(b) search screen
(c) result screen

moments but represent the moment of the highest intensity (e.g. negative vs. positive) and the moment of the ending [3]. This effect known as the peak-end-rule implies that retrospective evaluations may provide only limited information about the actual experience. On the other hand, findings from emotion research suggest that for multi-episode events spanning over several days, the retrospective global evaluation can be predicted best with average of the single evaluations [4]. The author concludes that the gap between experience and remembered experience is rather small for multi-episode events. The current paper presents a study aiming to investigate this potential gap between actual experience and subsequent report in the context of product evaluation over a middle-term usage period of one week.

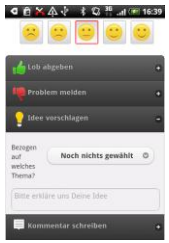
Related work

Although the peak-end-rule was initially researched in medical contexts [3], its validity could also be shown for product evaluations [5]. In a study by [5], the participants were asked to perform several tasks with software used for setting up hearing instruments. Mental workload was assessed after each task while perceived usability was measured after all tasks were completed. The results showed a recency-effect: the perceived usability was predicted best with the mental workload rating at the end of the episode. However the time frame of the investigated episodes was rather short (M= 2 h.). Accordingly it is unclear if the predictions of the peak-end rule are also applicable with extended time-frames. To our knowledge the few longitudinal studies have so far not explicitly addressed the predictions of the peak-end-rule or the gap between actual and remembered experiences. Karapanos et al. [6] for example studied perceived product quality of a TV set-top box over a four week-period taking measurements after week one and again after week four. Hence, their results do not comprise continuous measurements of the experiences over time but the perceived quality at the specific time points. The main finding was that in the beginning pragmatic aspects determine satisfaction with the product while after four weeks hedonic qualities dominate the perceived goodness of a product. In a study by Kujala et al. [7] remembered experiences of facebook and mobile phone usage were assessed, but not the actual experience immediately after interaction. It was shown, that changes in long-term UX are related to the hedonic qualities of a product, rather than to its pragmatic qualities. Also Kujala et al. [8] and von Wilamowitz-Moellendorff [9] et al. measured the remembered experiences of mobile phone usage over time, but again did not assess the actual experience. One of the few studies including online measurements over a longer time frame was also done by Karapanos and colleagues [10]: They started with data collection one week before product purchase and ended data collection after four weeks of usage. The product investigated was again a mobile phone. Before the purchase, the participants' expectations were collected, and after the purchase participants were asked to post their experiences with the phone one a daily basis. Moreover they should also rate the overall and specific product quality each day. With this very thorough study the authors were addressing the change of user experience over time as well as the product qualities determining positive experiences. Still they did not aim to investigate the potential gap between actual and remembered experience specifically. So the degree of concordance between actual and remembered experiences for multi-episode events is still unknown. If the remembered and experi-

Feedback tool

Opening the feedback tool via the menu

Start screen of the feedback tool



Accessing the sub-categories

Choosing a sub-category

Explaining the ratings with text input

For more information on the feedback tool see: http://reqorder.com/

ences would largely correspond to the actual experiences, time-consuming assessment of continuous data could, at least for some evaluations, be replaced with the retrospective elicitation of experiences from memory as proposed with iScale [11], Corpus [9] or the UX Curve [8]. Thus we wanted to investigate, if and to what extent evaluations measured during usage differ from remembered experiences.

METHOD

Participants

We invited 20 participants to take part in our study. All of them owned an Android smartphone. One user could not participate in the wrap-up workshop and was excluded. Three participants were accidentally assigned to the same ID number and were thus also excluded. Participants were aged between 19 - 59 years (M=31y., 3 f., 13 m.). In return, they received 80 Euro as compensation. To further motivate the users we offered two additional 20 Euro vouchers, one for the best idea and one for providing the most qualitative feedback.

Tested Application & Device

The application evaluated was an Android app called AskWiki [12] (cf. Sidebar AskWiki). As the app was installed on the users' own devices, several different smartphones were used (cf. Figure 1).

Measures

During the field trial users were asked to give daily feedback via an additional feedback app which was implemented in the AskWiki app (cf. Sidebar Feedback tool). The in-app feedback comprised an overall rating using an adapted version of the faces scale by [13]. The scale shows five faces ranging from "very sad" to "very happy". Although the scale was initially developed for assessing pain intensity, it has been successfully employed for measuring UX during PC usage [14].

Additionally participants should explain their ratings in form of praises or problems regarding AskWiki. For feedback which was neither praise nor a problem they could choose an additional open category for miscellaneous comments called "other". Furthermore they could suggest new ideas for AskWiki. For each of these categories (praise, problem, other, idea) four subcategories were offered. The subcategories were "speech recognition", "response time", "search result", "handling & usability" and again an open one called "other". Participants were asked to indicate to which aspect (subcategory) of AskWiki the comment referred to. The categories were derived in an internal expert workshop prior to the study. As AskWiki's functional scope is relatively narrow we expected the subcategories to cover the range of potential feedback. Another reason for offering subcategories was to reduce the time-consuming annotation of the resulting qualitative data. However we checked the data after the collection and had to re-categorize some comments. Most adjustments were necessary in the category "other"; many comments contained both, a praise and a problem, e.g. "the speech recognition was working well, but still I did not get a search result". If this was the case, the comment was counted in both categories. For the re-categorization the inter-coder reliability was Cohen's κ =.850. We also checked the plausibility of the valence (positive/praise, negative/problem, ambivalent) of the comments; here Cohen's kappa was κ=.861. Disagreement was solved with discussion.

After the field trial we collected the retrospective assessments. Participants should give us an overall rating

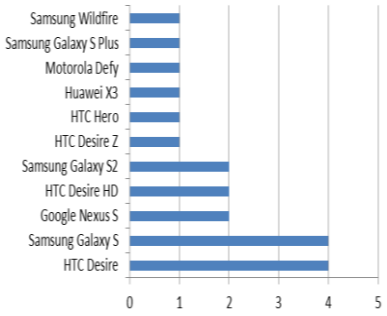


Figure 1. Frequency of devices used in the study

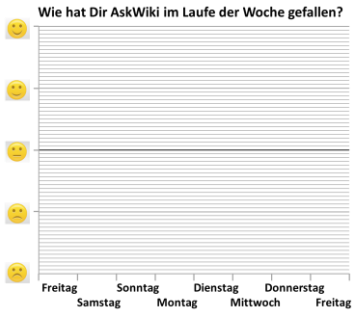


Figure 2. UX-Curve Template

of AskWiki on the faces scale in combination with the UX-Curve proposed by [8] as a paper-and-pencil survey The five faces were presented at the y-axis of the curve template; on the x-axis the days of the field trial were shown (cf. Figure 2). Participants were asked to draw a curve describing how much they liked AskWiki during the course of last the week. Furthermore they should recall the experiences which led to change in the curve and annotate the curve accordingly. The experiences annotated in the curve were categorized employing the subcategories already used in the feedback tool. All categorization was conducted by two independent intercoders. For the categories Cohen's kappa was κ= .891, for the valence ratings Cohen's kappa was κ= .913. Again disagreement was solved by discussion.

Procedure

Participants were invited to a kick-off workshop, were the app was installed on their devices. Then the instructions for the field trial and the in-app feedback tool were explained. The instructions asked the participants to be honest and to post their feedback daily. AskWiki itself was only briefly explained in order to simulate the natural situation of a Market download. The users' were just told that AskWiki offers precise answers to precise questions. Before they left, they received a booklet with the instructions of the field trial and the explanation of the feedback tool and the categories.

The kick-off workshop was followed by a period of eight days during which the online feedback was gathered. The trial started on a Friday afternoon after the kick-off workshop and ended on the following Friday with the wrap-up workshop. Here the retrospective measures were collected.

Results

Quantitative data - Overall ratings

At first we inspected the agreement between the overall ratings measured during usage and after the field trials. Therefore smiley ratings assessed with the in-app feedback tool and the smiley ratings with the UX-Curve were plotted against each other for each day. We only included participants who provided feedback for at least four days, which was half of the trial's duration. That left 11 participants for further analysis. As depicted in Figure 3, remembered ratings and ratings collected during usage intra-individually clearly differed. While for the measures collected during usage seven curves (1, 3, 6, 7, 13, 14, 19) showed better ratings in the end than in beginning, after usage only in four curves (5, 7, 10, 13) the ratings were improving. To further analyse the ratings we used the mixed models function in SPSS. We chose this method as we did not want to exclude subjects who had failed to provide data only for one or two days. Mixed linear models yield results comparable to a classical repeated measure ANOVA, but can deal with missing data [15] .Although the individual curves implied large variance, analysis showed that ratings given during usage did not significantly differ from ratings after usage, $F(1, 68.51)=3.75$, $p=.057$. Also between the different days no differences were found, $F(1, 84.38)=0.17$, $p=.990$ (see Figure 4).

Qualitative Data - Comments

During the field trial users posted 118 (excluding ideas) comments, most of them reporting a problem (N=73). The category with the most feedback was search results (N=44), followed by speech recognition (N=30), response time (N=18) and usability (N=16). The open category was only used for reporting system crashes (N=4) or problems with the feedback tool (N=3).

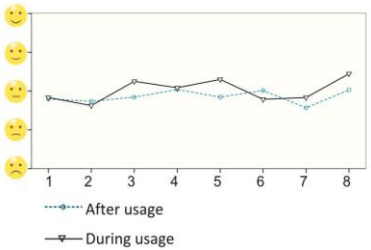


Figure 4. Mean ratings on Smiley Scale (y-Axis) collected during usage and after usage for each day of the field trial (x-axis) overall participants

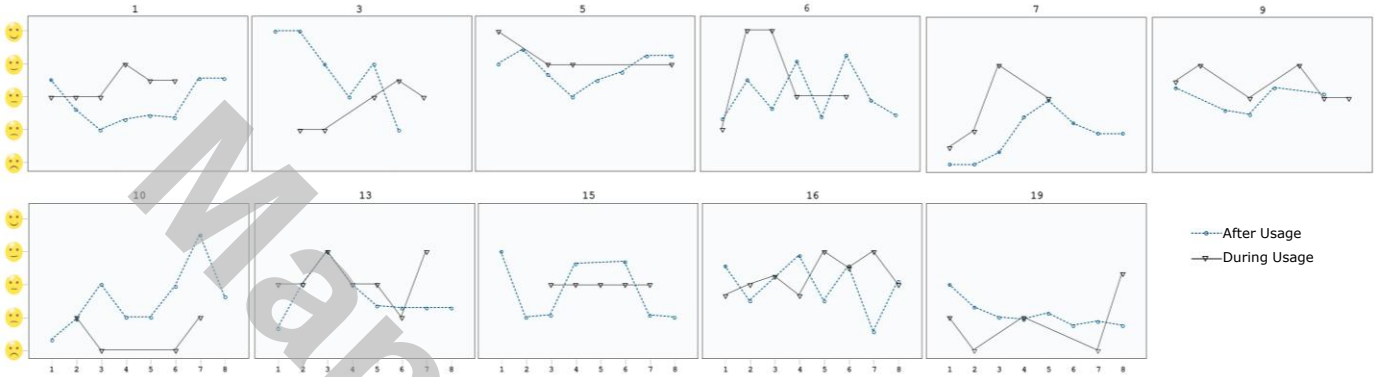


Figure 3. Mean ratings on Smiley Scale (y-Axis) collected during and after usage for each day of the field trial (x-axis) by participant

Three further comments could not be categorized (cf. Figure 5). The comments assessed after usage (N=104), the annotations of the curve, showed similar patterns: Most feedback was concerned with search results (N=36), again followed by speech recognition (N=23). Nine comments referred to response time, six referred to usability and five to system crashes (cf. Figure 6). Regarding those categories a χ^2- test indicated no significant difference between the online and the remembered comments.

For the annotations on the curve, 21 of the comments did not fit in annotation scheme. Those comments were too general and did not refer to a specific aspect of the app ("worked normal"). However 14 were affective judgment e.g. "I was positive surprised" or "disillusion". We then checked again the comments assessed during usage, only one of the comments was "purely" affective. The other comments were rather unemotional, reporting specific problem or nice features, e.g. "the search takes very long" or "very detailed information" but the users did not explicitly state how the problem/praise affected their mood or attitude towards AskWiki. Please note that the instruction during usage and after usage differed only slightly: During usage they were asked "How much do you like AskWiki all in all?"; after usage the question was "How much did you like AskWiki in the course of the last week?".

Discussion

The results of this explorative study indicate that the gap between remembered and online measurements of UX is rather small. These findings are in line with the previous results of [4]. To sum up the results: the quantitative ratings did not show significant differences, still for the qualitative data results show, that during usage, the feedback is more specific but contains less affective evaluations of the app. Considering the UX model proposed by [1], which differentiates between the hedonic and the pragmatic dimension of quality perceptions, comments during usage were mostly regarding the pragmatic dimensions, while ratings after usage also included hedonic aspects. Accordingly, if the overall attractiveness of, or the emotional reactions to-

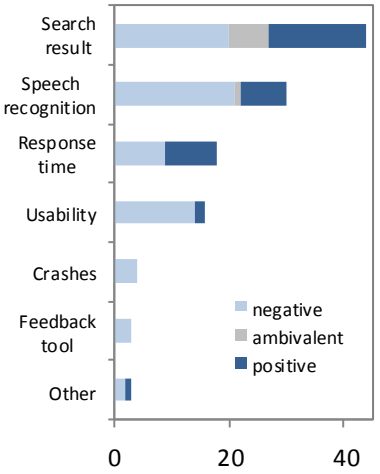


Figure 5. Comments collected during usage

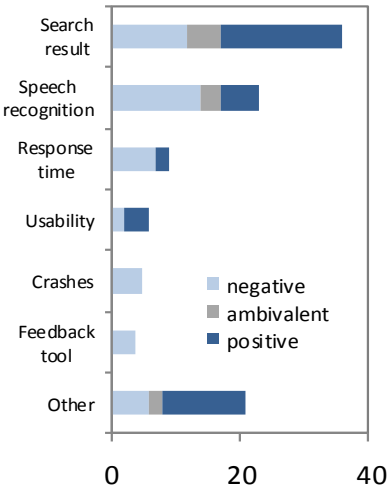


Figure 6. Comments collected after usage

wards, a device or service are in the focus of a study, it may be sufficient to conduct what Karapanos [11] labeled as "lightweight" longitudinal studies, which rely on participants' remembered UX. Still, for the developers of AskWiki, the detailed daily feedback (e.g. distorted search results) provided the most valuable insights in terms of improving the app's usability. Related to this, it may be the case that participants remembered their experiences so well because entering the feedback during the trial lead to a deeper processing and thus to a better memorability of their experiences in our study. Another side note of our study is that users gave more qualitative (N=164, including ideas) than quantitative feedback (N=104, all users) during the trial. This may however be due to the additional voucher we offered during the field study for the most qualitative feedback – still we expected the opposite as quantitative feedback is easier and faster to provide.

References

[1] Hassenzahl, M., Diefenbach, S. and Göritz, A. Needs, affect, and interactive products - Facets of user experience. *Interact. Comput.* 22, 5 (2010), 353-362.

[2] Bargas-Avila, J. A., Hornbæk, K. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proc. CHI 2011*, ACM (2011), 2689-2698.

[3] 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*, 4 (1993), 401-405.

[4] Miron-Shatz, T. Evaluating Multiepisode Events: Boundary Conditions for the Peak-End Rule. *Emotion*, 9, 2 (2009), 206-213.

[5] Hassenzahl, M., Sandweg N. From mental effort to perceived usability: transforming experiences into summary assessments. In *Proc. CHI 2004*, ACM (2004), 1283-1286.

[6] Karapanos, E., Hassenzahl, M., Martens, J.-B. User experience over time. In *Proc. CHI 2008*, ACM (2008), 3561-3566

[7] Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K. and Sinnelä, A. Identifying Hedonic Factors in Long-Term User Experience. In *Proc. DPPI 11* (2011), 137-144.

[8] Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., Sinnelä, A. UX Curve: A method for evaluating long-term user experience. *Interacting with Computers*, 23, 5 (2011), 473-483

[9] von Wilamowitz-Moellendorff, M., Hassenzahl, M., Platz, A., Dynamics of user experience: how the perceived quality of mobile phones changes over time. In*: NordiCHI Workshop User Experience - Towards a unified view*, (2006), 74-78.

[10] Karapanos E., Zimmerman J., Forlizzi J., Martens J.-B. User Experience Over Time: An Initial Framework, In *Proc. CHI 2009*, ACM (2009), 729-738.

[11] Karapanos, E., Martens, J.-B., Hassenzahl, M. On the Retrospective Assessment of Users' Experiences Over Time: Memory or Actuality? In *Proc. CHI 2010*, ACM (2010), 2689-2698.

[12] Burkhardt, F., Zhou, J., "AskWiki": Shallow Semantic Processing to QueryWikipedia. In Proc. *EUSIPCO 2012 (2012)*.

[13] Andrews, F. M., Whitney, S. B. *Social Indicators of Well-Being:America's Perception of Life Quality*. Plenum, New York, NY, USA, 1976.

[14] Schleicher, R., Trösterer, S. The 'Joy-of-Use'-Button: Recording Pleasant Moments While Using a PC, In *Proc. INTERACT 2009*, Springer (2009), 630-633.

[15] Quenée, H. van den Bergh, H. On multi-level modeling of data from repeated measures designs. *Speech Communication*, 43 (2004), 103-121.