Usefulness of Long-Term User Experience Evaluation to Product Development: Practitioners' Views from Three Case Studies

Jari Varsaluoma
Tampere University of Technology
Korkeakoulunkatu 6, P.O. Box 589, 33101
Tampere, Finland
jari.varsaluoma@tut.fi
+358 40 8490856

Farrukh Sahar
Tampere University of Technology
Korkeakoulunkatu 6, P.O. Box 589, 33101
Tampere, Finland
farrukh.sahar@tut.fi

ABSTRACT

Understanding the temporal aspects of user experience (UX) has received increasing attention in the HCI community. However, little empirical evidence is available on how practitioners in product development companies evaluate the usefulness or actually use long-term UX evaluation data in their work. In this study, we explore how practitioners (e.g., managers, designers and UX specialists) evaluate the usefulness of long-term UX evaluation results to their own work. Three case studies were conducted with longitudinal and retrospective methods in a company developing interactive sports products. Our findings suggest that long-term UX evaluation provides results that are perceived as interesting, relevant and useful by practitioners. Potential uses for the results were e.g., verifying practitioners' expectations, planning future work, understanding changes in UX, the development of future products, and updating current software products. Future research should focus on how to provide long-term UX evaluation results in more efficient manner to benefit product development.

Author Keywords

Usefulness; evaluation; long-term; longitudinal; user experience; usability; product development; case study.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Today, several companies developing interactive products have adopted user studies as a regular part of their product development processes. Traditional user research methods often focus on the first experiences and learnability problems that novice users have with interactive products. However, previous research suggests that conventional usability testing methods may not reveal the problems that can cause frustration for more experienced users over time [11]. Indeed, there has been an increasing interest in HCI field towards the temporal aspects of usability and user experience (UX) [4, 7, 15, 16].

There are no exact definitions for terms *long-term UX* and *longitudinal research* in HCI literature. However, several of the proposed UX models consider the temporal aspects of UX [e.g. 10]. Also, an emerging definition states that longitudinal research looks beyond the initial UX (or learning experience) [4]. Longitudinal research "is ideal for studying how and when users transition from novice to expert, as well as addressing issues such as abandonment or adoption rates, learnability, comfort with technology, productivity, and evolution of user perceptions" [4]. In this paper, *long-term UX evaluation* refers to longitudinal and retrospective studies that focus on understanding the change in product UX over time.

Motivation and benefits to conduct long-term studies have been addressed by the HCI research community [4, 7, 15, 16]. However, there is lack of empirical research on how practitioners in companies utilize results from long-term UX evaluations in their work and how *useful* this information is from practitioners' perspective. We argue that providing useful UX evaluation results that can support product development is a key factor in motivating stakeholders to invest in conducting long-term studies in future.

In this paper, we explore the *usefulness of long-term UX evaluation results for practical work* over three case studies in one company. The questions that motivated this research were:

- What kind of long-term UX evaluation results are the most useful for practitioners (e.g., managers, designers, UX specialists), who participate in the development of interactive products?

NordiCHI '14, October 26 - 30 2014, Helsinki, Finland
Copyright 2014 ACM 978-1-4503-2542-4/14/10…$15.00.
http://dx.doi.org/10.1145/2639189.2641218

- For what purposes are the reported long-term UX evaluation results seen as useful?
- How do practitioners actually use the long-term UX evaluation results in their own work?

During the case studies presented in this paper, information was collected on users' experiences with products and how UX relates to other aspects, such as customer loyalty. Long-term studies can result in a vast amount of information that can be beneficial to practitioners in different positions, e.g. management, marketing and design. Therefore, we were interested to explore how managers, in addition to UX specialists and designers, would use the reported long-term UX evaluation results. Due the exploratory nature of this study, we let the practitioners freely describe, what use (if any) they had for the reported results.

By presenting new empirical research results, this study can help building the body of knowledge for long-term research in HCI. The issues highlighted in this paper can contribute to the ongoing discussions and motivate the future research of long-term UX evaluation practices for both industry and academia.

First, we present an overview of the current long-term UX research practice in HCI, followed by discussion on measuring the usefulness of long-term UX evaluation results. The research process chapter presents the three case studies and the personnel surveys for measuring the usefulness of the long-term UX evaluation results. The results chapter describes the findings from the personnel surveys. In discussion, the main findings are reviewed and their meaning discussed. Finally, the research is summarized in the conclusions, with the limitations of the study and implications of the findings for future research.

BACKGROUND

Long-Term UX Research Practice in HCI

In [17], three perspectives for HCI studies were presented based on the time period the study covers. Typical usability tests are a *micro* perspective studies (one to two hours), while longer-term studies are divided into a *meso* perspective (e.g., 5 weeks) and *macro* perspective studies (from years to the whole product lifecycle). While macro perspective studies are rare in HCI, the number of published meso perspective studies has been increasing since 2006, judging by the number of workshops and other events around the topic [7].

Long-term studies are not dependent on any specific method, and using a combination of quantitative and qualitative methods is encouraged [4]. Retrospective methods such as CORPUS [17], iScale [6] and UX Curve [9] can be cost-effective alternatives to repeated measurement methods, such as the Experience-Sampling Method (ESM) [2] and the Day Reconstruction Method (DRM) [5]. Retrospective studies rely on users' memories of experiences and are prone to biases. However, memories can guide customers' future behavior and what experiences will be reported to others [12]. Therefore memories of product use can be relevant information for product development purposes. Lastly, data logging methods (e.g., usage logs) provide an interesting viewpoint for observing changes in product use over time [4, 7].

Longitudinal studies are useful for studying change over time, as they include two or more observations or measurements with the same users [17]. However, in quantitative longitudinal studies a minimum of three measurements is advised to differentiate true change from measurement error [14]. The length of previous longitudinal studies in HCI varies from few weeks to three years [4]. In order to track the change over time, some of the dimensions (e.g., tasks, users, measures, or products) have to stay constant over the study period. As the same participants use the studied product over time, they will get more experienced with the product. Therefore, no longitudinal survey samples the exactly same users twice [13]. This should be considered especially when the learning process itself is of interest, e.g. how long it takes and why for new users to learn to use a product efficiently?

How to decide the timing and frequency for measuring long-term UX? Considering longitudinal studies in general, if no theoretical guidance is available for deciding the measurement times, Ployhart and Ward [14] propose to 1) consider "natural" measurement occasions for the studied phenomenon, 2) conduct interviews or observations with subject matter experts, and 3) review literature that studied similar phenomena. Few studies in HCI literature discuss the most beneficial measurement times with interactive products. In a longitudinal study by Kujala and Miron-Shatz [8], DRM [5] and questionnaires were used to study 22 users' experiences with new mobile phone models. After the first week with DRM, more retrospective measurements were conducted on the 6th day, after 2.5 months and after 5 months of product usage. Surprisingly, some basic usability problems were reported still after 2.5 months. Another study [13] used a cross-sectional approach to study differences between novice and expert users regarding frustration episodes. Although the sample size was small, results suggest that studies where applications are used beyond a year may not be beneficial, as the most observable differences occur within three to six months from the beginning of use.

In practice, the number and times of UX measurements can depend on several factors, including the product itself (e.g. use frequency, the estimated length of the product learning period and product life cycle), users' characteristics (e.g. previous experience with similar products), available research resources, measured factors, and stakeholders' demand for receiving actionable results. Finding a balance between the length of a single survey and the number of measurements is important, since each measurement requires effort from the participants and participant drop-out is common for longitudinal studies [4, 14]. Overall, it seems that more empirical research is required as the work towards building a rich body of knowledge for long-term UX research in HCI continues.

Usefulness of Long-Term UX Evaluation Results to Product Development

As user research, be it long-term or short-term, is conducted in a product development company, the probable goal is to provide useful information to be used in specific phases of the product development process. To our knowledge, measuring the *usefulness of long-term UX evaluation results to work practice* has received little attention in HCI literature. Usefulness has been measured before regarding the use of different HCI methods. In [1], the perceived usefulness of different HCI methods by HCI practitioners was measured using a pen-and-paper questionnaire and a web survey. Participants rated the usefulness of the provided HCI methods for different phases of the development process (start, mid, and end phase) using a rating scale of 1 (Not at all useful) to 5 (Very useful). In addition, the participants were asked what methods they had actually used in the different product development phases.

In the current study, the evaluation of the *usefulness* of the long-term evaluation results was supported with additional measurements that we considered meaningful. These related factors included: 1) what is considered *interesting* in the results, 2) what is *relevant* (similar to importance) in the results for each practitioner's work, 3) the *novelty value* of the results, 4) *likeability to utilize* the results, and 5) the *actual use* of the results in practice. Our hypothesis was that information rated as interesting, novel or relevant to the practitioners' own work would also have more potential of being useful. However, it is possible that information that is considered e.g., relevant to one's work, can be considered uninteresting, or vice versa. Furthermore, although the reported likelihood to utilize the results in future might relate to the usefulness of the reported information, an observation or measurement of actual use of the results is required to properly evaluate their usefulness.

METHOD

Between the years 2011 and 2013, three case studies evaluating long-term UX of products were conducted with one Scandinavian company developing interactive digital sports equipment. The studies were a part of a joint research project between a university and the company.

The focus of this paper is in how the practitioners in the company evaluated the usefulness of the long-term UX evaluation results. Detailed results of the case studies are not in the scope of this paper and therefore only the type of the reported results is presented. Next, we describe the case studies (briefly), the personnel surveys, and how they were conducted.

Case studies

Table 1 summarizes the case studies and their research methods. Web surveys were mainly used, since all the studies were international. Both qualitative and quantitative questions were used to collect data on users' experiences with the products. All the participants were contacted via the company's customer database and were chosen based on a screening survey. One of the authors participated closely in the design of the studies. In total there were five sessions (DC and SWb were reported in two parts) during the three case studies where the results were presented to the company personnel.

Case Study 1: Diving Computer (DC)

The first case study DC evaluated the UX of a diving computer and its associated software after the first months of usage. Another objective was to study how a new software update would affect the UX. Furthermore, the study acted as a pilot for using the iScale tool [6] in a remote study.

Two retrospective measurements were carried out using a web survey with the Attrakdiff questionnaire and iScale. Attrakdiff provides quantitative data describing user's perceptions towards the evaluated product [3]. 33 users, all male, answered the first survey (part 1). The time of product usage varied between the participants from one month to five months. 21 of the participants continued to the second survey after a software update was released. The second survey was sent to the participants after each of them had used the product for six months.

Both results presentations (DC part 1 & 2) of the study included: 1) customer background information e.g., previous experiences from the brand and similar products, 2) expectations from the product before purchase (in retrospect) and how the expectations were met, 3) the

Case study	Studied product	Number of respondents	Measured product usage period	UX evaluation methods
1. DC	Diving computer	Part 1: 33 Part 2: 21	Part 1: Varied from 1 to 5 months Part 2: 6 months	Retrospective: Web survey with Attrakdiff [3] + iScale [6]
2. SWa	Sports watch A	25	From 3 to 6 months	Longitudinal/Retrospective: Weekly web survey
3. SWb	Sports watch B	Part 1: 111 Part 2: 104	Part 1: From 1-2 to 4-5 months Part 2: From 1-2 to 7-8 months	Longitudinal/Retrospective: Monthly web survey with Attrakdiff [3]

Table 1. Case study summary. Results from the case studies DC and SWb were reported to the company in two parts.

attractiveness of the diving computer and its associated software, 4) satisfaction with the product and why, 5) the importance of the product for oneself and why, 6) willingness to recommend the product to friends and why, 7) an abridged version of Attrakdiff, and 8) summary of the iScale curve shapes with positive and negative experiences.

Case Study 2: Sports Watch A (SWa)

The goal of the second study SWa was to understand how the UX of different product components associated with a sports watch can affect the evaluation of the product's overall UX. Furthermore, changes in the UX after the initial learning period were studied. The studied product consisted of the sports watch as the main unit, two sensor units and a web service.

The study was carried out as a weekly web survey, following a repeated measurement design. 25 participants (4 female) were chosen, each with over two months of use experience with the product. The data collection phase lasted for two months and included eight weekly surveys per participant.

The results presentation included: 1) customer background information, 2) the number of different sports activities where the product had been used, 3) overall positive and negative feelings with the product over time 4) willingness to recommend the product to a friend over time, 5) the number of reported positive and negative experiences over time for each product component, 6) the summaries of positive and negative experiences for each product component, 7) experience quotes from users, and 8) component-specific design ideas based on the experiences.

Case Study 3: Sports Watch B (SWb)

The studied product in the third case study SWb consisted of a sports watch, a sensor unit, installable software, and a web service. The research goal was twofold: first, to understand the customer journey since the beginning of use, including users' expectations and their fulfilment. Second, to learn how the UX changes over time and what factors affect these changes (e.g. software updates).

The study consisted of six monthly web surveys. The final report included results from 104 participants (7 female). The study covered the experiences with the product from the first and second month until the eighth and ninth month of usage, depending on the date of the product purchase.

A preliminary report (SWb part 1) was created to present the main results from the first three surveys (with still 111 participants), including: 1) customers' background information, e.g., previous experiences with similar products, relationship with the brand, 2) expectations before the product purchase (in retrospect), 3) how easy it was to take the product into use and need for support, 4) challenges when starting product usage, 5) product use frequency over the first three months, 6) satisfaction with each product component over time, 7) willingness to recommend the brand to a friend over time and why, and 8) summary of positive and negative experiences that the users reported with the product.

The final report (SWb part 2) included: 1) customers' background information, 2) how easy it was to take the product into use and need for support, 3) product use frequency over six months, 4) satisfaction with each product component over time, 5) willingness to recommend the brand to a friend over time and why, 6) Attrakdiff measurements over time, 7) the most important product qualities over time (based on Attrakdiff), 8) expectations before product purchase (in retrospect), 9) how expectations were fulfilled, 10) quantitative analysis results, e.g., how emotions relate to the willingness to recommend, satisfaction with the product, and Attrakdiff measurements, 11) summary of positive and negative experiences that the users reported with the product over time, 12) new design ideas based on the users' suggestions, 13) summary of the experiences that users reported on the 6th (last) survey in more detail, and 14) the conclusions of the study, including the first steps in the customer journey, how expectations were fulfilled, and changes in UX over time.

Personnel Surveys

Procedure

Before each of the five results presentation sessions, a contact person from the company informed stakeholders about the upcoming presentation. Project team members presented the results in a live meeting with a Microsoft PowerPoint. Paper surveys were used to gather feedback from the session participants, who were free to answer the survey during or after the presentation. The presentation slides and case study data files were delivered to the company contact person after each presentation.

In order to verify if practitioners had actually used the research results from the case study SWb in their work, a follow-up survey was administered in 2014, nine months after the last SWb results reporting session. During this time, the company had launched a new model of the sports watch that had been studied in the case study SWb. An email invitation to answer a web survey was sent to the participants who had attended either of the case study SWb results sessions. Reports from the case study SWb had been available in the company's intranet for the last nine months. Participants were asked to skim through the reports before answering the survey.

Survey questions

Table 2 presents the questions for each of the personnel surveys. New questions were added in SWa and SWb part 1. SWb part 2 contained more results than any of the previous presentations and we decided to modify the survey to follow the structure of the presentation. For each results section in the presentation there was a separate page in the survey, with three new questions: 1) "How useful these results are for you? (1=Not at all useful, 5=Very useful)", 2) "What is the novelty value of these results for you? (1=No novelty value, 5=High novelty value)", and 3) "Any comments, feedback or thoughts from these specific results?" Each page included pictures of the presentation slides as a memory aid. Since the usefulness was asked separately for each results section, the question about relevant information was excluded. However, the question about "what was interesting in the results" was kept as a summary question (see Table 2).

In the follow-up survey, the participants were asked if they had read or used either of the case study SWb reports in their work. If the report had not been used, the participant was asked why not. Otherwise, the participant was asked that what kind of information he or she had been looking for in the report and for what purpose. Furthermore, the respondent was asked how useful the information in the reports had been in the participant's own work, on a scale 1 to 7, where 1 = not at all useful and 7 = very useful. The same scale was used at the end of the survey to ask how useful the long-term user studies are in general from the participant's point of view, continued with an open-ended question: "Why? Please clarify your answer".

Participants

30 individuals from the company answered at least in one of the five personnel surveys (52 responses in total). 20 (67%) answered only in a single survey. Two participants answered in all the five surveys. Figure 1 presents a summary of the personnel survey participants. Each response was categorized into one of the four categories based on participant's title and work tasks. The categorization was done in cooperation with two employees from the company. "Manager, high-level" category included titles such as Business Unit Director, Design Manager and Program Manager. "Manager" category included e.g., Product Manager, Product Concept Manager and Team Manager. "Designer/UX Specialist" category included e.g., UI Designer, Interaction Designer and UX specialist.

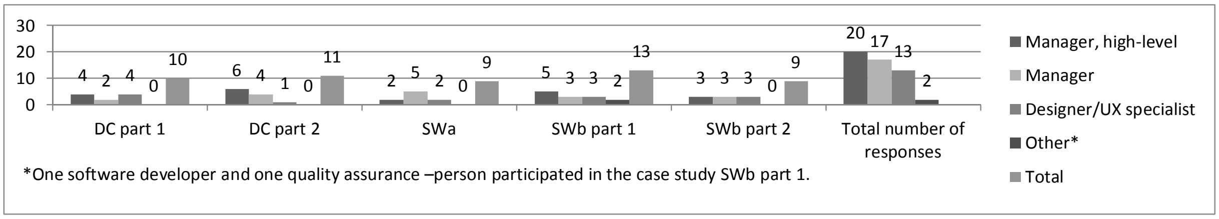
Unfortunately, despite two reminders over one month, only five responses were received to the follow-up survey: two "high-level managers", two "managers" and one "other".

Analysis

The responses to the open-ended questions were content analyzed by one of the authors. If a single answer entailed several different aspects, each one was coded as an individual item. Similar items between responses were linked into appropriate categories that were named to describe the items in them. Data for each question from each survey were first analyzed separately. After this, similar categories from all five surveys were combined and the categorization descriptions updated as necessary. If the same participant had similar responses to the same question in different surveys, items from each response were added up as separate items for their categories. Due the small number of participants, no statistical tests were conducted to compare the follow-up survey results with the previous surveys.

Question	DC part 1 & 2	SWa	SWb part 1	SWb part 2
"What kind of user information would be the most beneficial for you?"			x	x
"Was there something especially interesting in these results? What?"	x	x	x	x
"Was there something relevant to your own work in these results? In what way?"	x	x	x	
"Where would you use the relevant results? (e.g., specific phases of product development)"	x	x	x	
"How likely will you utilize the presented results in your own work? (Not at all likely 1-7 Very likely)"	x	x	x	x
"What was missing in the results or would have been more useful for you?"		x	x	x

Table 2. The questions (five open and one Likert-scale) used in personnel surveys during the case studies.



*One software developer and one quality assurance –person participated in the case study SWb part 1.

Figure 1. Personnel categorization and the number of responses to the personnel surveys in the case studies. In total, 52 responses were collected from 30 separate participants.

RESULTS

1) What kind of information was considered interesting in the long-term UX evaluation results? Figure 2 presents a summary of what the personnel found interesting in the long-term study results. After summarizing all the five surveys, there were in total 14 different categories with 70 items related to what was considered interesting. Nine items were unique. From 52 possible responses, on seven occasions (14%) a participant left this question unanswered.

The three largest categories with 8 responses (15%) were related to: 1) comparing the results with the participant's own expectations, 2) the positive and negative experiences that users reported, and 3) how UX changed (or did not change) over the measurement period. User satisfaction with the product, UX of specific components, and factors that affected UX were also found as interesting topics (5 responses each).

It seems that practitioners had initial expectations for the results based on their subjective knowledge: "Expected results. Good that initial 'hunch' of UX predicted results were aligned with actual results" (ID8, Designer/UX), or previous research: "Very consistent with previous research findings and our subjective understanding about the topic" (ID15, Manager, high-level). Two high-level managers commented, on different case studies, that most of the findings were already known to them: "Interesting, yes, but not much new info, most of this is already known by us." (ID25).

2) What kind of information in the long-term UX evaluation results was considered relevant to the practitioners' own work? Figure 3 summarizes what the participants found the most relevant for their own work in the reported results. The analysis resulted in 12 categories with 56 items (11 unique). This question was not included in the fifth personnel survey (SWb part 2). From the 43 possible responses, two were blank (5%). The same participants had left the previous question about interesting results blank.

The most repeated responses (8 in both categories, 19%) were: 1) all the results were relevant and 2) the results show where to focus next, e.g., improving a specific product component/feature or promoting specific aspects of the product in future. Furthermore, the positive and negative experiences with the product and user feedback in general were found relevant (7, 16%). It was noted that a single participant gave similar comments in three different surveys for categories "Taking the product into use" (ID8, Designer/UX) and "Long-term UX" (ID2, Manager).

Two participants did not seem to make any distinction between what was interesting or relevant content in case the SWa presentation, as they answered the later question: "things mentioned above" (ID16, Manager, high-level) and "see previous" (ID17, Manager).

3) Where the information that was considered relevant could be used? Figure 4 illustrates where the participants could have used the research results from the first four presentations. The qualitative analysis resulted in 12 categories with 52 items, from which eight were unique. Two participants did not respond to this question.

Majority of the responses (12, 28%) related to proposing, concepting and defining new products: "Good points for defining product specifications and requirements before actual development project" (ID20, Manager). Two of the second largest categories with seven responses (14%) indicated that the findings could be used e.g., in updating the current product, and as an input to future product development to avoid some of the reported problems. For example, one designer commented: "We can use these results because software development is still going on daily" (DC part 1, ID10). For case SWb part 1, there were comments related to current software and future products: "Considering the content of the next software releases…" (ID26, Manager, high-level), and "Found problems will be very likely to be fixed in future products" (ID23, Manager, high-level).

4) What kind of information was considered novel and/or useful in the results of case study SWb part 2? Participants' evaluation of the novelty and usefulness of the result section are provided in Table 3. The most useful results seem to have been the satisfaction scores, the detailed experiences with the product and the conclusions of the study. All the results, apart from the quantitative analysis, were rated above average in usefulness, with mean 3.9 or higher (scale being from 1 to 5). Although the least useful, the quantitative results were seen as the most novel results. It seems that the presentation time and content were not enough to communicate the quantitative analysis results to the audience properly: "This part was a bit difficult to comprehend and would have needed a little bit more practical explanation. Still, the content was interesting", (ID9, Designer/UX).

Attrakdiff results were seen as the least novel, but the differences with other result sections were small. In overall, the novelty of the results was rated slightly above average. However, the standard deviations of the novelty values were slightly higher than those in the usefulness scores. This indicates that there were more differences among the participants in what was considered as novel information when compared with what was seen as useful information.

5) How likely the participants were going to use the results in their own work? Table 4 presents the mean values and standard deviations for each survey. There were seven missing responses, six of them for the case study DC part 2. All the mean scores were above average, suggesting that majority of the participants were planning to use some of the results in their work. Case SWa had the lowest score and the highest standard deviation, indicating that some of the participants did not see clear usage for the research results. One participant who gave the lowest rating (3) also commented that the product "is not in the core of my responsibilities" (ID17, Manager). Other low score (4) came from a manager (ID20), who stated that they had received similar feedback from other sources and that "most of the detailed findings being taken into account already".

6) Was there some information missing in the results? After the analysis of 25 responses it seems that the practitioners had been looking for more quotes from users and raw data in the case study SWa presentation (7 responses). Also, the most important usability problems and more insight of problem severity were missed by four participants for case SWb part 1. Furthermore, three participants asked in case SWa, that how many separate users had reported similar comments, since this had not been evident in the presentation. The rest of the comments/questions were unique and related to e.g., comparing different user groups, recommendations on how to improve/maintain UX over time, how to increase the recommendation rate, top 5 positive feedback of the UI, and more detailed data about the product components that received negative feedback. Interestingly, one manager (ID25) noted that while there

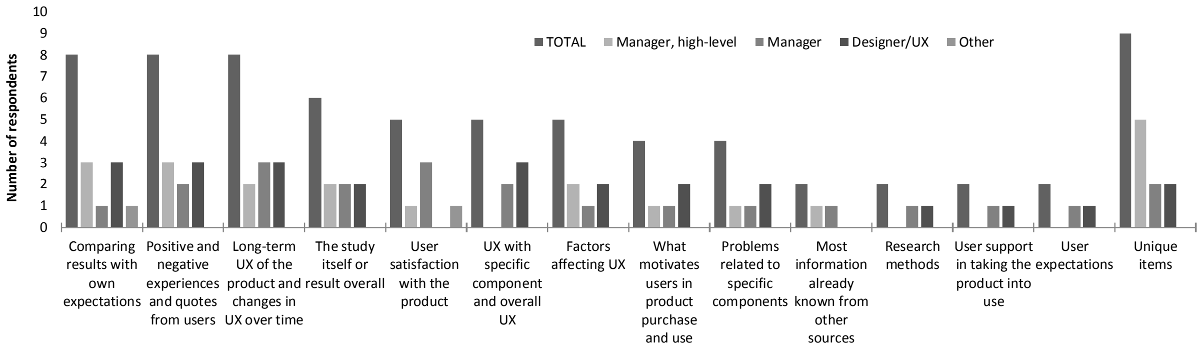


Figure 2. Responses to the question "Was there something especially interesting in these results? What?" Asked in all the five personnel surveys (n=52).

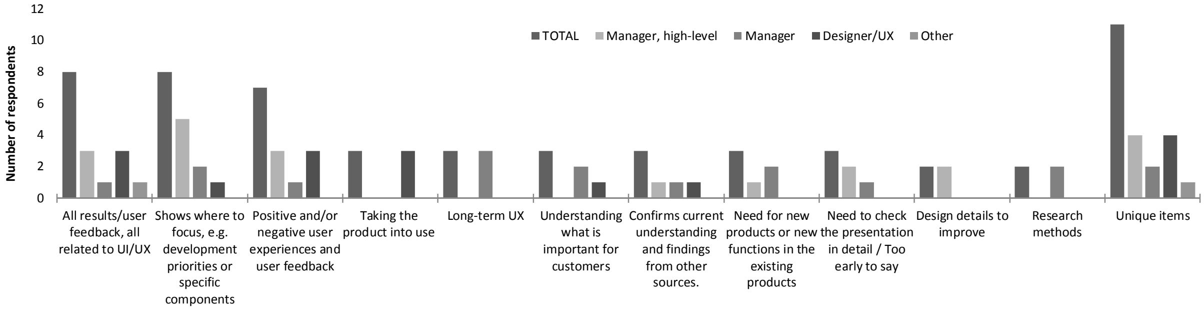


Figure 3. Responses to the question "Was there something relevant to your own work in these results? In what way?" Asked in the first four personnel surveys (n=43).

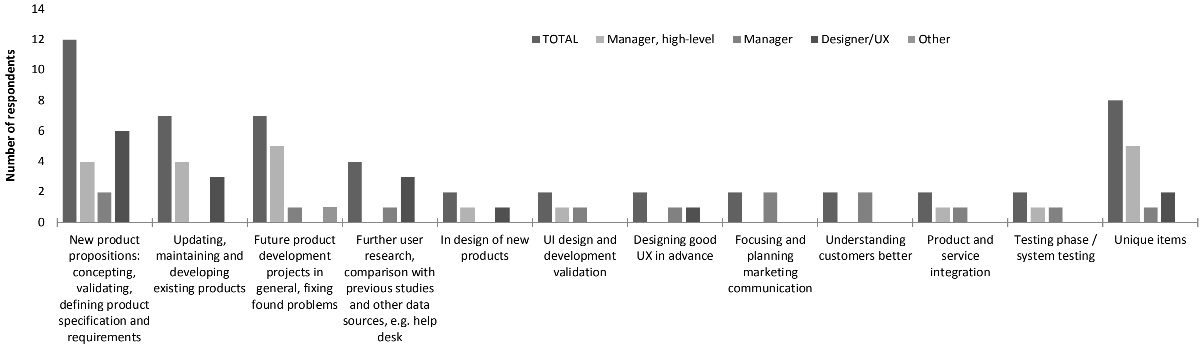


Figure 4. Responses to the question "Where would you use the relevant results?" Asked in the first four personnel surveys (n=43).

Result presentation sections for case study SWb part 2	Usefulness Mean (SD)	Novelty Mean (SD)
1. Attrakdiff measurements over time and the most important product qualities	3.9 (0.6)	**3.1** (0.8)
2. Satisfaction with each product component over time	**4.1** (0.6)	3.2 (1.2)
3. Willingness to recommend the brand to a friend over time and why	4.0 (0.7)	3.2 (1.1)
4. Expectations before the product purchase (in retrospect) and how they were fulfilled	4.0 (0.5)	3.4 (0.7)*
5. Quantitative analysis results, e.g., how emotions relate to the willingness to recommend, satisfaction with the product, and Attrakdiff measurements.	**2.9** (1.1)	**4.0** (1.0)
6. Summary of positive and negative experiences that users reported over time and design ideas based on the users' suggestions	3.9 (0.9)	3.3 (1.0)
7. A detailed summary of the experiences that users reported in the 6th (last) survey	**4.1** (0.6)	3.4 (1.1)
8. Conclusions of the study, including the first steps in the customer journey, fulfillment of expectations, and changes in UX over time	**4.1** (0.8)	3.3 (1.0)

* One answer missing, therefore n=8 for this question.

Table 3. Responses to the questions "How useful these results are for you? (1=Not at all useful, 5=Very useful)" and "What is the novelty value of these results for you? (1=No novelty value, 5=High novelty value)" Asked in case study SWb part 2 (n=9).

	DC part 1	DC part 2	SWa	SWb part 1	SWb part 2	Total
Mean (Std Dev)	6.60 (0.52)	5.83 (0.75)	**4.89** (1.45)	6.08 (1.04)	**6.63** (0.52)	6.02 (1.12)

Table 4. Responses to the question "How likely will you utilize the presented results in your own work? (Not at all likely 1-7 Very likely)". Asked in all the five personnel surveys (45 responses, after seven missing values).

was little new information in the results of case study SWb part 1, the results may also be a bit "outdated" due the rapid development cycle nowadays.

7) What kind of information from the case study SWb results did the practitioners use in their work during the next nine months? Two high-level managers, two managers and one quality assurance person answered the follow-up survey. One manager and the quality assurance person had not used either of the case study SWb reports, while the other respondents had read and/or utilized both of the reports from one to three times. The manager's (ID20) reasons for not reading the reports or using the information were that 1) there has not been a project where to use the information, and 2) lack of time. The quality assurance person (ID21), who had not participated in the part 2 presentation, replied that "I was not aware / I forgot that this document even existed" (SWb part 2) and that the presentation already gave the needed information (SWb part 1).

Regarding both reports of the case study SWb, three of the participants had been looking for information about: 1) "consumer long-term usage" and "using the findings for future work" (ID2, Manager), 2) "who buys the product and why", "who are our users and what they are experiencing", and "what kind of products we should make" (ID3, Manager, high-level), and 3) "enhancement ideas", "feature priorities", "usability pros and cons", and "as motivational feedback to development team to help them understand how important different UX aspects are" (ID6, Manager, high-level).

8) How useful the results of the case study SWb had been (after nine months) and how useful long-term user studies are seen in general? On a scale 1 (not at all useful) to 7 (very useful), the mean rating for the usefulness of part 1 results was 5 (SD 1.2) and for part 2 results 4.8 (SD 1.3). One manager (ID2) gave the rating 7 while other four participants rated the usefulness to 4 and 5. When looking at the previous surveys, all the participants, apart from the quality assurance person, had been very likely to utilize the results in their own work, as they gave the rating 6 or 7. Only one participant (ID20) who gave high ratings in the likeliness to utilize the results (6 and 7 in both SWb surveys) had not used them in his or her own work. The same participant also stated that while the information was important and useful, they receive "quite a lot of feedback continuously, that are often around the same topics as the study."

The mean rating was 5 (SD 1.4) for the question: "In general, how useful do you see the information from long-term user studies for your own work?" (scale 1-7). The answers were nearly identical to the ratings of the usefulness of the case SWb research results. Some of the reasons why long-term user studies were seen useful, were: 1) "they give us insights on a longer term usage of our products which would be difficult for us to do internally at this level" (ID2), 2) "to see effect on software update" (ID6), and 3) "to learn how experience changes with learning and after 'honey moon'" (ID6).

DISCUSSION

Our results indicate that a majority of practitioners, both managers and designers/UX specialists, found the long-term UX evaluation results interesting and relevant for their work. Also, the mean ratings for usefulness and likeliness to utilize the results were high. However, the number of separate categories and unique items in what was considered interesting and relevant suggests that in order to serve the needs of different practitioners in the company, the long-term studies should be versatile in what they measure. Since long-term product evaluations may require a substantial amount of time and resources, the early involvement of stakeholders and careful scoping of the long-term study is recommended [4].

It is interesting to note that while nearly all the results in case study SWb part 2 were considered highly useful, their mean novelty values were lower. A possible explanation for this could be stated in the open comments: similar findings had been received from other sources, and the results were mostly in line with the practitioners' own expectations or current understanding about the topic. This notion underlines one of the challenges with long-term studies: receiving the research results can take too long for them be as beneficial as they could be. As one manager (ID25) commented, the results can be too "outdated" for today's rapid development cycles. Also the fact that more practitioners came to listen to the case SWb preliminary report (part 1) than the final report suggests that the earlier results have more value for practice. One proposed solution is that the ongoing results are published periodically during the study [4]. Preliminary reports could be provided only of the measurements that are fast to analyze and sought after by stakeholders, therefore having better changes to still influence the design and development of the next product version. Alternatively, more systematic utilization of the other data sources that may provide similar information, such as customer care, could be developed.

All the products evaluated in the case studies were already available in the market. Therefore, it is no surprise that the results were mainly planned to be used in proposing, concepting and developing future products. However, for software products, the long-term results were still relevant as they could be used in the upcoming software updates. This raises an interesting question regarding the long-term UX of software: as software updates may alter the product (e.g., user interface), how does this affect the UX over time and how comparable the measurements are as one more dimension (the product, in addition to the learning user) changes? Although software updates add complexity to the evaluation as users may update their products at different times, the feedback regarding updates could be of major importance for stakeholders. Therefore, when planning long-term UX evaluation of software products, the estimated dates for update releases can be beneficial measuring points.

Due the small number of responses to the follow-up survey, it is difficult to make conclusions of the actual use of long-term results. However, it seems that lack of time and opportunities to utilize the information, being content with learning about the results in the first place, or simply not being aware of the available information can be reasons for not utilizing the results. Furthermore, the perceived usefulness of the results seems to decrease over time as three out of five participants rated the usefulness lower in the follow-up survey. Still, these results are not generalizable as the sample size was small. Also, no designers/UX specialists participated in the follow-up survey. This highlights the challenge of high drop-out rates in longitudinal studies, especially in industrial setting, where employees change or even the company can change its owner in the middle of the study [14].

The reasons why long-term studies were seen useful in general seem to echo some of the previously discussed findings from this study. Long-term studies can help understanding how the UX changes over time through learning and how software updates affect the UX. Also, the insight of longer-term usage that the six-month SWb study offered was something that would be challenging to achieve with internal resources. This hints that studies of this extent are not common in the company involved. However, if similar information is available through other feedback channels, even less systematically collected (e.g. via customer care), the perceived usefulness of long-term UX evaluation results seems to diminish. Still, in our case the results of carefully planned long-term studies seemed to have value for practitioners by confirming their own expectations and subjective understanding of the topic.

Apart from the products studied in this paper, the development of other product types might benefit even more from understanding how, when and why UX changes over time. Possible examples could be: 1) practitioners in an online gaming company are interested to know how and why the motivation to play their games changes over time, 2) designers (and customers) want to measure how fast a new employee will learn to use a complex factory monitoring system efficiently, and 3) marketing team of an educational software company needs proof that using their software has positive effect on students' test results over time. Since evaluation takes time and product development needs user feedback as soon as possible, long-term studies may be most beneficial for companies that develop updatable software (e.g. web services, mobile applications) or interactive products based on previous product versions (e.g. mobile phones, cars, domestic appliances).

CONCLUSIONS

This study was set out to provide more empirical evidence in how practitioners in companies evaluate the usefulness of long-term UX evaluation. The question was studied through three long-term case studies in a company developing interactive sports products. The results of this study suggest that managers and developers perceive long-term UX evaluation generally interesting, relevant and useful. Practitioners found the results relevant for 1) comparing the results with previous knowledge, 2) understanding the change in UX over time, 3) focusing future work, 4) concepting and development of future products, and 5) updating current software products. However, challenges remain related to the time and resources needed for conducting long-term studies: 1) research results may arrive too late to benefit ongoing product development and 2) other sources, such as customer care, may provide similar information, which decreases the usefulness of the results of long-term studies.

The main limitations of this study were that only one company was involved and the sample size of product development practitioners was small, especially when measuring the actual use of the results from case study SWb. Also, no responses were received to the follow-up survey from designers or UX specialists, who should be the most obvious people to utilize UX evaluation results. Furthermore, practitioners' feedback regarding the usefulness of the results could have been different if they had spent more time inspecting the evaluation reports before answering. However, in reality, busy managers and designers might not have time to inspect lengthy research reports and therefore live presentations may sometimes be the only channel to deliver research results.

This study highlights some of the benefits and challenges related to long-term UX evaluation in practical product development work. The empirical findings can inform HCI practitioners and contribute to future research on how long-term UX evaluations are conducted in industry. In future, more extensive research with different product development companies and their practitioners is required to determine how long-term UX evaluation results are used in practice, especially by designers and UX specialists. Also, little is known on how to actually design memorable and positive long-term user experiences [10]. Another interesting topic would be the ways of speeding up the process for providing actionable results from long-term UX studies.

REFERENCES

1. Bark, I., Følstad, A., and Gulliksen, J. Use and usefulness of HCI methods: results from an exploratory study among Nordic HCI practitioners. *People and Computers XIX — The Bigger Picture*, (2006), 201-217.
2. Csikszentmihalyi, M., and Larson, R. Validity and Reliability of the Experience-Sampling Method. *Journal of Nervous and Mental Disease*, 175 (1987), 526-537
3. Hassenzahl, M., Burmester, M., and Koller F. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität In: J. Ziegler & G. Szwillus (Hrsg.), *Mensch & Computer 2003. Interaktion in Bewegung*, (2003) 187-196.
4. Jain, J., Rosenbaum, S., and Courage, C. Best practices in longitudinal research. *Ext. Abstracts CHI '10*, (2010), 3167-3170.
5. 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. *Science 306*, 5702 (2004), 1776-1780.
6. Karapanos, E., Martens, J.-B., and Hassenzahl, M. Reconstructing experiences with iScale. *Int. J. Human-Computer Studies 70*, 11 (2012), 849-865.
7. Karapanos, E., Jain, J., and Hassenzahl, M. Theories, methods and case studies of longitudinal HCI research. *Ext. Abstracts CHI'12*, (2012), 2727-2730.
8. Kujala, S. and Miron-Shatz, T. Emotions, experiences and usability in real-life mobile phone use. *Proc. CHI '13*, (2013), 1061-1070.
9. Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., and Sinnelä, A. UX Curve: A method for evaluating long-term user experience. *Interacting with Computers 23*, 5 (2011), 473-483.
10. Kujala, S., Vogel, M., Pohlmeyer, A., and Obrist, M. Lost in Time: The Meaning of Temporal Aspects in User Experience. Ext. Abstracts CHI '13, (2013), 559-564.
11. Mendoza, V. and Novick, D. Usability over time. *Proc. SIGDOC '05*, (2005), 151-158.
12. Norman, D.A. Memory is More Important than Actuality. *Interactions,* March + April (2009), 24-26.
13. Novick, D. and Santaella, B. Short-term methodology for long-term usability. *Proc. SIGDOC '12*, (2012), 205-211.
14. Ployhart, R. E., and Ward, A.-K. The "Quick Start Guide" for Conducting and Publishing Longitudinal Research. J. Bus. Psychol. 26, 4 (2011), 413-422. doi:10.1007/s10869-011-9209-6
15. Vaughan, M. and Courage, C. SIG: capturing longitudinal usability: what really affects user performance over time? *Ext. Abstracts CHI '07*, (2007), 2149-2152.
16. Vaughan, M., Courage, C., and Rosenbaum, S. Longitudinal usability data collection: art versus science? *Ext. Abstracts CHI '08*, (2008), 2261-2264.
17. von Wilamowitz-Moellendorff, M., Hassenzahl, M., and Platz, A. (2006). 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*, 74-78.