Towards Understanding How Agile Teams Predict User Experience

Towards Understanding How Agile Teams Predict User Experience

Kati Kuusinen, Heli Väätäjä, Tommi Mikkonen, Kaisa Väänänen
Tampere University of Technology, Tampere Finland
{kati.kuusinen, heli.vaataja, tommi.mikkonen, kaisa.vaananen}@tut.fi

Abstract. In this chapter, we compare UX assessments of users and agile team members to learn to what extent developers can predict how users experience (UX) the product the developers are working on, and where user involvement is truly required. We compared UX assessments of agile team members (N = 26) and users (N = 29) of six enterprise applications with statistical tests. Moreover, we analyzed the data with principal component analysis to reveal the main dimensions of UX for enterprise software. Our results confirm prior research findings that agile team members can put themselves in the users' position when evaluating instrumental aspects of UX of the software they are working on. However, it seems that developers cannot evaluate non-instrumental quality. Therefore, direct user involvement from participation to evaluation or other means to support user empathy in development process is needed. We recommend additional means, such as personas to help agile team members empathize with the users and their needs for non-instrumental qualities of the enterprise software.

Keywords: UX evaluation, agile software development, enterprise software.

1 Introduction

Building on advances in software technology, rapid and continuous development approaches have become a viable option for numerous end-user applications. With such infrastructure, developers can expose new features to randomized experiments in real-life context, where data regarding actual users' preferences can be collected and analyzed with statistical hypothesis testing. However, executing such tests requires a substantial number of real users, which can be a problem in enterprise software development, which is targeted for work-related use. Moreover, such tests reveal only user behavior with the system, leaving the developers unaware of users' subjective experiences.

The interest in gathering real-life user data reflects the differences between users, who perceive the software via user interfaces, and developers, who know the software from inside out. Gathering data from real-life use can be regarded as a way to address claims that developers do not truly understand users, and that users do not really understand what they eventually want [1]. These problems have been partially solved with rapid iteration cycles promoted by agile software development approaches. Still, while at best, such approaches advocate a paradigm shift from front-heavy planning and design to short development cycles, where user feedback is constantly collected,

delay is introduced when getting feedback from end users as well as when analyzing the feedback.

UX work has traditionally followed the user-centered design process defined in [2], and mechanisms for integrating UX work in agile development frameworks remain largely unestablished. The most widely used approaches include a design upfront phase and (often unsuccessful) attempts to maintain the pace of development iterations with user testing [3,4]. To truly include UX work in agile development, lightweight methods are needed to evaluate UX as a part of iterative development.

Given an improved understanding regarding how agile teams and users assess UX, developers themselves may handle some aspects of UX, at least to a certain degree, thus lightening the workload of UX specialists (UXS). To address developers' ability to predict UX, quantitative measurements are needed to measure and compare UX as assessed by development team members and users. Moreover, to allow frequent evaluation of UX in agile projects, simple evaluation frameworks that minimize work are needed.

We aim to make UX work more rapid in enterprise software development. By enterprise software we refer to applications that are intended for work purposes and are primarily developed to meet organizational rather than user needs; by UX work we mean activities, such as research, design, development, and evaluation that aim at developing software that is usable, fulfills user needs, and provides desired interaction qualities. Our research has three practically oriented goals:

1. To enable collecting rapid user feedback to support iterations that synchronize UX and software development work.
2. To place the focus of limited UXS resources on issues that software developers are not able to handle by themselves.
3. To enable setting clear, meaningful UX goals to focus on big picture and to unify design effort based on real user preferences.

To meet these goals, we study to what extent developers are able to understand UX so that some of the validation steps with real end users could be eliminated. The goal is to understand if some of the UX validation could be performed as a part of the software creation, and, if so, what are the things that truly need experimentation with actual end users.

To this end, we compare assessments of UX between users and team members (developers, product owners (PO), and UXSs). We asked team members to assess the software from two perspectives: as themselves and when trying to put themselves in the users' place. We conducted a survey in six agile development projects from five companies working on enterprise software. We surveyed 26 team members—including developers, UXSs, and POs—and 29 end users considering their perception of UX in the software that was produced in each project. We measured UX using a scale with 16 items from UX dimensions identified in [5,6]. Our results suggest that developers are able to understand the practical quality (such as usefulness) of the developed system, but understanding hedonic qualities (such as pleasure) seems to need support to help agile team members empathize with the users. In addition, our results contribute towards understanding the main UX dimensions for enterprise software.

The rest of this Chapter is structured as follows. Section 2 introduces work related to mechanisms of measuring UX and studies regarding the differences in how users

and development teams perceive UX. Section 3 describes our research methods. Section 4 presents results of the principal component analysis and related varying assessments of UX. Section 5 discusses the validity and limitations of this research. Section 6 discusses the main contributions and the implications of our results, and finally, Section 7 draws some final conclusions.

2 Background and Related Work

The study presented in this Chapter is based on an earlier study [7] of the same projects with the same participants from agile teams, in which we studied how the participants contributed towards UX work. In that study, we found that UXSs (UX specialists) collaborated the most with developers during demo sessions, when discussing the UI design and when determining how to implement design details. Developers did not participate in user studies or tests, or in clarifying end user definitions or target user groups. Thus, developers' understanding of users remained shallow and many of them wished to be more involved in user communication. Those findings motivated us to continue our research with these projects with a further study, reported in this Chapter.

2.1 Concept of UX

UX is subjective, context-dependent, and dynamic [8]. It is affected by *users'* expectations, needs, and motivation, *systems'* characteristics, such as purpose and functionality, and the *context of use* including physical, organizational, and psychological aspects [9]. The standard definition of user experience (UX) is as follows: a "*person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service*" [2].

According to Law et al. [10], in academic research, the most commonly utilized frameworks for UX are the *hedonic-pragmatic model* [11]) and *sense-making experience* [12]. The hedonic-pragmatic model divides user experience into *hedonic* or the non-utilitarian dimension and *pragmatic* or the instrumental dimension [11]. Hassenzahl [11] further divides the hedonic into two subdimensions of identification and stimulation, while the instrumental contains mostly items related to usability and usefulness. Usability is often seen as a necessary precondition for good UX [13,14].

Väänänen-Vainio-Mattila et al. [15] discuss the differences in the conception of UX between academic UX research and industrial UX development. They conclude that while the research concentrates mostly on hedonic aspects and emotions, companies concentrate more on functionality and usability issues [15]. Moreover, although early HCI studies concentrated almost exclusively on task- and work-related usability issues and achievement of behavioral goals [9], UX research has mainly concentrated on consumers and leisure systems (see e.g., [16]) for categorization of publications applying the hedonic). Thus, it is unclear what shapes UX of enterprise software or work-related tools: what are its dimensions and is it different from UX of leisure systems?

2.2 UX Evaluation in Agile Development

Vermeeren et al. [17] identified 96 different UX evaluation methods originating both from academia and industry. The methods included lab, field, and online data gathering activities, such as surveys, focus groups, expert-based methods, controlled observations, and contextual inquiries. Most of the methods were intended to be used with functional prototypes or with working products. Regarding online evaluation methods, Vermeeren et al. [17] conclude that whilst they can be lightweight, cheap, and fast, some of them are problematic because they require laborious analysis, which can decrease their practical feasibility.

In industrial agile development, ensuring the desired UX of implemented features is often addressed with user tests [18]. According to Da Silva et al. [18], user testing is one of the most commonly used practices in agile UCD work, and it is equally conducted on low-fidelity prototypes and on working software. In the most traditional form, user tests are conducted by recruiting users to arranged test sessions where users perform planned use cases or scenarios while a researcher observes them [19]. Arranging and interpreting these sessions require time and resources [19,20,21]. Ardito et al. [22] found in their survey conducted in Danish software development organizations that the most common obstacles regarding usability evaluation was the lack of resources and suitable methods. Lárusdóttir et al. [23] state that integrating traditional user testing into agile context is challenging, and thus companies tend to perform evaluations informally with only few users, gathering qualitative data during unplanned sessions.

In contrast to the traditional model discussed above, user tests can also be conducted remotely either synchronously (with a human moderator) or asynchronously (with a software moderator) [20]. Asynchronous user tests can save considerable time compared to traditional laboratory tests [24] and help to find a number of usability issues, especially when predefined tasks are given for users to conduct [25]. However, according to a recent literature review, automated user tests still seem to be rarely used in agile software development: utilizing "some kind of automated tool" was reported in 10% of the included papers [26] (it should be noted that [26] included also studies conducted in academic context in their review). Also, despite the perceived popularity of the user testing method, remote testing was mentioned in only one of the publications included in the systematic review of [18] . Another remote evaluation method is to publish the feature in a beta group or on the market and collect data of real users' actual use with methods, such as application performance management (APM) and real user monitoring (RUM) [27]. These methods can provide more realistic usage data from a larger amount of users but are mainly aimed for aftermarket evaluation [27].

Finally, randomized experiments with control and treatment groups consisting of real users (e.g., A/B testing) can be utilized for evaluating new features. This, however, requires a large user base. In addition, remote methods lack many qualitative aspects that can be perceived while observing the user, such as user's emotional state, level of satisfaction, or the reasoning behind user's choices [21,28]. Thus, remote evaluation should be accompanied with subjective UX surveying.

2.3 Measuring Dimensions of UX

A systematic review of UX measurements in HCI [29] categorized the measured dimensions of UX. Generic UX was found to be the most commonly measured UX dimension (41%). Other commonly measured dimensions were affect or emotion (24%), enjoyment or fun (17%), aesthetics or appeal (15%), and engagement or flow (12%). Motivation (8%), enchantment (6%), and frustration (5%) were also reported. Only 14% of the analyzed papers in this review measured hedonic quality [29]; they used Hassenzahl's [10] AttrakDiff or AttrakDiff2 scale or a self-modified version of it [29]. In addition, 20% of studies that used questionnaires to assess UX used AttrakDiff or AttrakDiff2, whereas 51% used self-developed questionnaires.

A more recent review of UX measurement reporting scale use found that AttrakDiff was the most used scale [16]. Of the reviewed papers, 58% used it or its adaptations, while the second most used group of scales, namely scales from consumer research, was utilized only in 8% of the included papers. Despite the wide usage of AttrakDiff, Diefenbach et al. [16] claim that it has issues with inter-correlations between the subscales; it does not separate between the UX dimensions clearly enough. Thus, they also conclude that the hedonic itself requires a clearer concept [16].

Other well-known scales include SAM (Self-Assessment Manikin) by Bradley et al. [30] for measuring emotion, a scale by Lavie and Tractinsky [31] for measuring visual aesthetics, the HED/UT scale [32], Pleasures of Play Scale [33], the Subjective Mental Effort Questionnaire (SMEQ) [34], the Flow State Scale (FSS) [35], Attrak-Work [36], Emocards [37], Pleasure-Arousal-Dominance (PAD) [38], and Subjective Usability Measurement Inventory (SUMI) [39].

UX-related measure scales that are utilized for evaluation of enterprise software mainly measure usefulness, productivity, performance, and ease of use. The Technology Acceptance Model (TAM) by Davis [40] predicts users' intention to use through perceived usefulness and perceived ease of use. Technology Satisfaction Model (TSM) is an alteration of TAM, where the intention of use is replaced with user satisfaction, since the use of enterprise software often is mandatory for the user [41]. In addition to perceived usefulness and perceived ease of use, [41] included perceived loss of control and perceived market performance in their scale. Finally, Task-Technology Fit [42] measures the impact of individual performance via effectiveness, productivity, and the system's ability to increase the productivity of the user. Thus, to the best of our knowledge, there are no validated scales available for specifically assessing hedonic quality of work-related software.

As Lindgaard and Kirakowski [43] point out, creating rating scales is tricky. Still, a considerable amount of UX researchers decide to utilize none of the validated scales but create their own scale: authors of 51% of analyzed papers in [29] and 27% in [16] utilized self-developed scales or single items of established scales. Based on our own experiences with rating scales, we assume that current validated scales do not properly assess researchers' needs. While research on dimensions of UX and measuring those has been conducted, it is still unclear how (and with which items) the dimensions actually are (and should be) measured. In addition, most of the validated scales are originally intended for consumer products. Consequently, there is a lack of evidence regarding how well existing scales fit to work-related contexts.

2.4 Different Roles' Perceptions of UX

Few studies have investigated how different stakeholder groups construe UX, i.e., what kind of personal constructs or perceptions they have about UX. Hertzum et al. [44] conducted a study with 48 participants from China, Denmark and India to study the effects of both the nationality and the stakeholder group. The study looked at the constructs of developers and users with the repertory grid interview technique. Concerning the nationality, no significant differences were found. For the two stakeholder groups, there were differences of the UX constructs. While users associate ease of use with leisure time systems and difficulty of use with work-related systems, developers do not have this distinction in their constructs. Furthermore, users conceive usefulness as related to frustration but separate from ease of use, whereas developers perceive ease of use, usefulness, and fun as related. Both users and developers have several constructs that are not visible in the dominant usability definitions at the time of this study, e.g., [2], such as fun and security.

In a study of 24 Chinese, Danish, and Indian usability professionals, Hertzum and Clemmensen [45] used repertory grid interviews to study usability professionals' constructs of usability. In this study, it was found out that goal-oriented performance is central for usability professionals, whereas their perceptions have less emphasis in experiential aspects of UX. Also in this study, the definition of usability [2] was found to be more limited than the constructs of the usability professionals, whose perceptions were broader especially in the experiential aspects of UX. In line with Hertzum et al. [44], usability was found to be construed similarly across the three nationalities of usability professionals studied.

Clemmensen et al. [46] studied the personal constructs of 72 usability professionals, developers, and end users with the repertory grid technique. Their finding was that usability professionals focus more on emotional aspects of UX, whereas users' perceptions of system use is more focused around the utility. Furthermore, usability professionals focus more on subjective aspects of UX than developers. This is in line with the usability professionals attempt to have empathy with the end users and to understand their viewpoint [47].

Sundberg [5,6] carried out research on the importance of UX factors in metals and engineering industry to support new product development. She compared the views of developers and users of industrial products on the most important UX related factors in three supplier cases. The three cases were three supplier companies, each with two of their customer companies. Both developers and users assessed pragmatic aspects more important than experiential (hedonic) aspects. Differing from this work, our research looks into how agile team members and users assess UX of enterprise systems in selected cases, how UX is construed by different groups, and the capability of agile team members to assess user experience in the role of users in order to assess when user involvement is needed in agile development activities.

3 Method

We conducted a survey study to examine how users and agile team members assess UX of enterprise systems created by the agile teams. The aim was to gain understanding regarding how users and team members assess UX and to reveal the main constructs through which they construe the UX of enterprise systems. In the survey, we asked the team members to give evaluation first as themselves (team measurement TO, *team* member evaluating in *own* role) and then as they think a member of a particular user group would answer (team measurement TU, *team* member evaluating in role of *user*). Users answered in a separate survey regarding their experience with the system (*user* measurement US).

3.1 Forming the Survey

We selected the UX measurement items based on a data-driven analysis (of a systematic review) of items utilized in previous UX measurement scales, which was used to create a measurement tool for UX in work contexts in the metals and engineering industry [5,6]. We utilized data from a preliminary analysis of Sundberg's [5] study to form our scale by selecting such UX items that cover all the main UX dimensions identified by Sundberg [5] and are relevant in the context of enterprise software. The items we used in the survey are presented in Table 1. The selection process in more detail was as follows. We selected items from both instrumental and non-instrumental quality categories. We aimed at covering all the main dimensions of UX identified by Sundberg [5,6]; we selected items from all categories containing more than one group of items. We selected items based on their frequency of occurrence found in [5,6]. When possible, we selected at least two items per group for internal validity and to increase measurement accuracy. Altogether, we selected eight items from both instrumental and non-instrumental categories.

We adopted the phrasing of question from AttrakDiff [10]as follows:

- In measurements TO (team member evaluating in own role) and US (evaluation by user): "*With the help of the word-pairs, please enter what YOU PERSONALLY consider the most appropriate description for the software.*" The measurement scale was a seven-point semantic differential.
- In measurement TU (team member evaluating in role of user): "*With the help of the word-pairs, please enter what you think USERS consider the most appropriate description for the software.*" The measurement scale was a seven-point semantic differential.

Table 1 Items (word-pairs) selected for the scale and associated UX dimensions [5, 6].

	Category	**Item left**	**Item right**
1	Overall system quality	Bad	Good
2	Overall system quality	Useless	Useful

3	Productivity	Hard to learn	Easy to learn
4	Productivity	Slow to use	Fast to use
5	Interaction quality	Difficult to use	Easy to use
6	System reliability	Unreliable	Reliable
7	Appeal	Undesirable	Desirable
8	Appeal	Not recommendable	Recommendable
9	Identification	Unconvincing	Convincing
10	Stimulation	Suppresses creativity	Promotes creativity
11	Affective quality	Discouraging	Motivating
12	Affective quality	Dull	Fun
13	Aesthetic quality	Unaesthetic	Aesthetic
14	Aesthetic quality	Amateurish	Professional
15	Aesthetic quality	Unpresentable	Presentable
16	Aesthetic quality	Conservative	Innovative

We specified user roles for the measurement TU to ensure that teams were responding with the particular user group that participated in the user survey in mind (i.e., not the customer). Additionally, we asked in separate questions the overall UX and the ability of the software to fulfill user needs as follows:

- Overall UX: In measurements TO and US: "*How would **you** rate the overall user experience of the software?*" and in measurement TU: "*How do you think **users** would rate the overall user experience of the software?*", both on a seven-point scale from" bad" to" good."
- Need fulfillment: In measurement US: "*How well does the software respond to **your** needs?*" and in measurement TU: "*How well does the software respond to **users'** needs*?". Both were on a seven-point scale from "*not at all*" to "*completely.*"

The question addressing overall UX was used as a reference question for the scale, and the word-pair scores were compared to it in the analysis. We also asked team members to *list one to three most important and least important UX goals for the developed software* from a predefined list of UX items measured in the survey. In addition, we asked the respondents to report their role as users or team members and the version of the system being evaluated. Users also reported their length of experience in using the system.

Table 2 Participating companies, project teams and their development practices. Legend: Scrum is an agile methodology presented in [48]. Kanban board is a tool for lean development introduced in [49]. Continuous development is discussed in [50].

P	Company description	Team size	Team practices
P1	An engineering and technology company with around 20 000 employees worldwide. Utilized both waterfall and Scrum practices. Several small distributed UX teams and UXSs	11, of which 8 developers located in Russia, 1 PO and 1 part-time UXS co-located in Finland	Scrum project. PO communicated with users, UXS drafted high-level design. PO selected the design that was communicated to developers. Developers decided about UX design details.
P2	An IT service company with 100-500 employees in Finland. UXSs working in project teams	6, of which 4 developers, 1 PO, 1 UXS, all co-located in Finland	Kanban project. UXS worked closely with developers. UXS's tasks were chunked and presented on the Kanban board. The UXS had partially also the PO role.
P3	An IT-service company with 100-500 employees in Europe. Utilized Scrum. A centralized UX team in one site and distributed UXSs in others	5, of which 2 developers, 1 PO, 1 UXS, all co-located in Finland	Scrum project. UXS tried to work one sprint ahead. Most of the UX budget was spent already during (heavy) design upfront, and there was less change for iteration during development.
P4 P5	An IT service company with around 20 000 employees worldwide The company mainly utilized customer-defined processes. It had a centralized UX team on one site and numerous distributed UXSs on several sites.	P4: 7 of which 2 developers in China, 2 developers and 1 UXS in Finland in location A, 1 developer and 1 PO in Finland in location B, the whole team working part-time for the project. P5: 4 of which 2 developers in Finland in location A, 1 PO in Finland in location B, and 1 UXS in Latvia	Both projects applied methods from agile frameworks and were moving towards continuous development. Projects had prioritized backlog, Kanban board and continuous integration in use. Demo sessions were arranged on demand. PO was responsible of communicating with users. In P4 the UXS made UX design whereas in P5 a developer made the majority of UX design work and the UXS was more a graphic designer.
P6	A mobile technology company with 100-500 employees worldwide. Utilized agile practices and customer processes. A centralized UX team	2, of which 1 developer and 1 PO co-located in Finland. Possibility to consult a UXS in another location in Finland	Free-form agile development. PO communicated with the UXS who made the UX design.

3.2 Description of Participants, Participating Projects, and Evaluated Software

Participants included both team members of projects developing enterprise software and users of the software under development. We selected development projects with following constraints:

- The project utilizes agile methods. The basic criterion was that the PO considers the project agile.
- The project has a release cycle of six months or less. For each new release in each project, UX was measured with no existing UX data made available for the team.
- The outcome of the project is enterprise software that will be used by several people.
- The outcome has a graphical user interface that requires design work.
- UX design work is ongoing or starting soon.
- Team members are willing to participate (not only the contact person).

We recruited the participant projects (Table 2) by participating in company events (e.g., fairs), from our previous business contacts and by snowball sampling. Participants in a user role were recruited by our contact persons in the projects.

Participants. Survey participants were agile team members from six software development projects from five companies and users of each system being developed. Our sample consisted of users (N = 29) and team members (N = 26) including software developers, UXSs, and POs (Table 3). User participants and project details are described in 3. Education in HCI was self-rated as: none, some self-learning or training, some studies (a compulsory course or similar), more than a couple of courses but less than a minor subject, minor subject, major subject.

Table 3. Participants from agile teams (R = range, M = mean, SD = standard deviation, HCI = human-computer interaction).

Role	**Developers (N = 17)**	**Product Owners (N = 6)**	**UX Specialists (N = 3)**
Mean age (years)	M = 31. (SD 5.)	M = 35 (SD 3)	M = 40 (SD 8)
Educational background	Information technology	Information technology	Information technology, society and culture, or industrial design
Education in HCI	None to major subject. The majority had some either self-learning or some courses.	Some self-learning or some courses	Some courses or major subject

Development experience (years)	R: 0-20, M:8, SD: 5	R: 2-9, M: 7, SD: 2	R: 0-20 M: 9 SD: 10
UX design work experience (years)	R: 0-10, M: 2, SD: 4	R: 0-1, M: 0, SD: 0	R: 5-20, M: 11, SD: 7
Project management experience (years)	R: 0-5, M: 1, SD: 1	R: 0-6, M: 4, SD: 2	R: 0-5, M: 2, SD = 2
Agile work experience (years)	R: 0-8, M: 4, SD: 2	R: 0-7, M:5, SD: 2	R: 5-9, M: 6, SD: 2

The majority (N = 19, 73%) of the participants in agile teams were from Finland. Five participants were from Russia, one was from Latvia, and one from China. In total, there were 40 team members working for the projects, of which 26 responded to our survey, resulting in a response rate of 65%.

As for the users, we did not ask where they were from in the survey. They were aged between 28 to 58 years (M = 42, SD = 9 years), and their roles are listed in Table 4. User response rate is unknown, since in some projects, invitation links to participate were put on an intranet or mailed to user organizations to be further distributed. However, we attached the survey with instructions on qualifications to participate and asked the users to report their role as users and the length of experience in using the system to evaluate participants' eligibility.

Table 4 Participating projects and their outcomes. Participated users and their roles. Number of participants per role and the time respondents had been using the system prior to the evaluation.

P	Developed system	User role	N of team members (26)	N of users (29)	Users' length of experience of using the developed system
P1	License generator	Sales engineers	6	3	'Tried it once or twice' to 'Used it several times'
P2	Communal online service for officers and citizens	Communal inspectors	4	4	'Tried it once or twice' to 'Used it a few times'
P3	Information system for nursery schools	Nursery school teachers	3	3	'Used it regularly for over a month but less than a year'

P4	Customer process monitoring feature	Service managers	6	2	'Used it regularly for over a month but less than a year'
P5	Launchpad and single sign-on for web applications	Employees of a IT services company	5	14	'Tried it once or twice' to 'Used it regularly for over a month but less than a year'
P6	Tool for software testing	Testers and developers	2	3	'Watched somebody using or demonstrating the system' to 'Tried it once or twice'

3.3 Analysis

We utilized the following quantitative analysis methods for the data:

Normality test. We utilized Shapiro-Wilk test for normality of the distribution; the data was non-normal.

Item counts. We counted occurrences of mentioned items to assess projects' most and least important UX goals.

Descriptive statistics. We calculated means and standard deviations to summarize the sample. As the number of participants and the evaluated software varied case by case, we utilized means when analyzing equality between responses of users and team members.

Tests for equity and difference. We chose to use nonparametric tests in our analysis. Our data consisted of both related and independent samples. We collected paired samples from team members in measurements TO (team member evaluating in own role) and TU (team member evaluating in role of user). For this data, we ran Wilcoxon signed rank test. Our null hypothesis was "*the median difference between measurements TO and TU is zero*", or, in practice, "*there is no difference between measurements TO and TU*". The test was run separately for each UX item. We utilized a Mann-Whitney U test for equality of means to compare responses of agile teams and users. We analyzed the difference between the following:

1.) Users' responses (measurement US) and team members responding as themselves (team measurement TO) and
2.) Users' responses and team members responding as they think users would respond (team measurement TU).

Our null hypothesis was "*the distribution of [UX item] is the same across categories of respondent type (user or team member)*."

We determined correlations pairwise for each UX item variable in all the measurements between the UX item and 1) overall UX score and 2) the need fulfillment score using Pearson product-moment correlation. We calculated similarity matrices and included values with significance level $p < .01$. We utilized a critical

value table and included cases as follows: measurements TO and TU (N = 26, df = 24): r > .496, and in measurement US (N = 29, df = 27): r > .471.

Principal component analysis (PCA). PCA is a multivariate statistical method that is used for extracting the important information from data and compressing the data set size by discarding other information, thus analyzing the structure of the data [51]. Principal components are obtained as linear combinations of the original variables and each component has the largest possible variance under the constraint that it must be orthogonal to the preceding components [51]. We conducted PCA with SPSS to detect structure in the data and to reduce the correlated observed variables to a smaller set of UX items. We used Varimax with Kaiser normalization as the rotation method. The amount of extracted principal components was selected based on eigenvalue (>1) and coefficients with absolute value less than 0.5 were suppressed in the analysis.

Scale reliability/internal consistency. We calculated Cronbach's Alpha coefficients for created principal components to measure internal consistency of the items loaded to the component. We interpret the alpha according to Nunnally [52] and use 0.70 as the threshold of acceptable consistency. Generally, a correlation coefficient of 0.7–0.9 indicates *high correlation*, whereas 0.5–0.7 indicates *moderate correlation*.

4 Results

We begin by presenting results of the principal component analysis and continue by presenting results of the assessments of agile team members and users.

4.1 Principal component analysis (PCA)

The 16 measured items loaded into four components in PCA (Table 5). Item scores in Table 5 indicate the strength of correlation between the item and the component. The first four principal components account for 69% of the variation (Figure 1). Table 6 presents the internal consistency of each component, indicating the extent to which items in the component measure the same dimension of UX.

Table 5 Rotated component matrix presents significant component loadings of PCA. Rotation was converged in 9 iterations using Varimax with Kaiser Normalization using SPSS. The data consists of measurements TO and US, N = 55.

Item	Component			
	1	2	3	4
Motivating – Discouraging	.81			
Fun – Dull	.79			
Promotes creativity – Suppresses creativity	.77			
Presentable – Unpresentable	.61			
Aesthetic – Unaesthetic	.57			
Innovative – Conservative	.56			

Easy to use – Difficult to use		.80		
Easy to learn – Hard to learn		.77		
Fast to use – Slow to use		.74		
Desirable – Undesirable		.53		
Good – Bad		.64	.52	
Useful – Useless			.71	
Recommendable – Not Recommendable			.58	
Professional – Amateurish				.85
Convincing – Unconvincing				.67
Reliable – Unreliable				.53

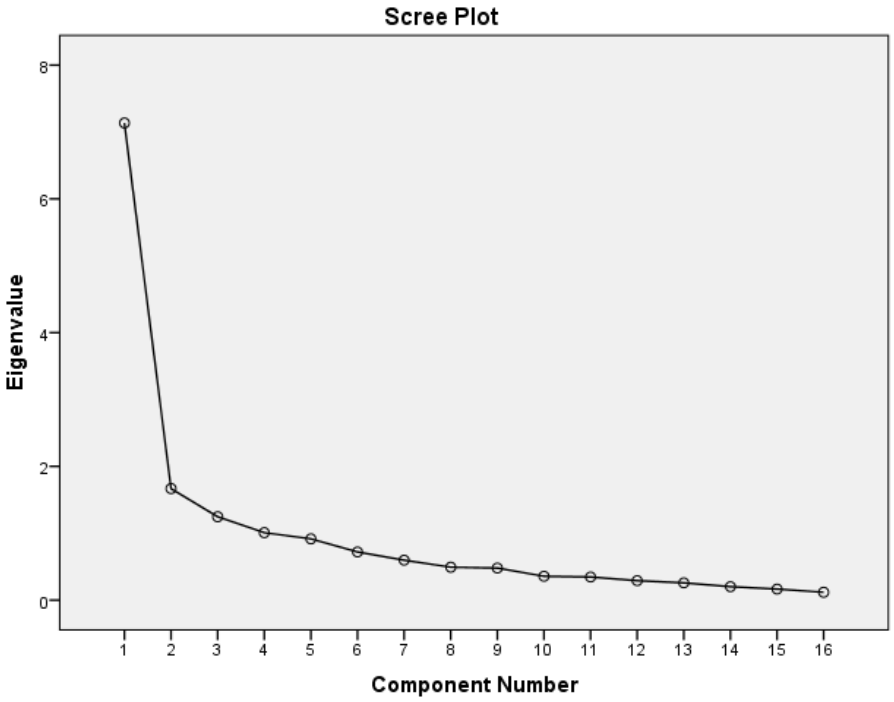

Figure 1 Scree plot for the variables. Cumulative percentage of variance for the first four components is 69. The first principal component explains 45% of the variance, the second 10%, third 8%, and fourth 6% of the variance.

Table 6 Internal consistency of principal components

Component name	Cronbach's Alpha	N of items in component
Motivation	.87 (good)	6
Productivity	.81 (good)	5
Usefulness	.75 (acceptable)	3

Professionalism	.69 (questionable)	3

Based on the strongest correlations of each component, we named the generated components as follows: 1. *Motivation*, 2. *Usability and willingness to use*, 3. *Usefulness*, 4. *Professionalism.* Items in each component vary accordingly.

The first component (motivation) explains the system's ability to motivate user via positive affect. It consisted of the following components*: motivating, fun, promotes creativity, presentable, aesthetic,* and *innovative*. It contains items from categories of affective and aesthetic quality and stimulation defined during the review. This component holds many items related to traditional hedonic quality, and it is also in line with *stimulation* defined by Hassenzahl [53].

The second principal component (usability and willingness to use) measures usability. It is thus connected with the user's willingness to use the system. The following items loaded to the second component: *easy to use, easy to learn, fast to use*, and *desirable*. In addition, item *good* partially loaded to this component. Based on the presence of components *desirable* and *good* with traditional usability metrics, this component can be interpreted that if the perceived usability of the system is low, users in general are not willing to use the system. The second component contains items from productivity, interaction quality, appeal, and overall system quality categories defined in [5].

The third component (usefulness) measures the scope of the system; how well does it fit to its purpose and is it useful? It is correlated with overall satisfaction and recommendability.

The fourth component (professionalism) seems to relate to work-related use itself and to the system's appropriateness to professional use. It contains items of *professional*, *convincing*, and *reliable*. The component can also be associated with the plausibility of the system's ability to complete required tasks.

These results from our work-related sample indicate that in work contexts, the dimensions of UX might not be the same as in leisure systems, and UX items might measure different aspects in work-related and leisure systems. For instance *professional* has been connected with aesthetic quality in leisure systems—the system *looks* professional instead of amateurish. In our study, it was connected with items *convincing* and *reliable*. Still, the basic dimensions of hedonic and pragmatic quality were clearly present in our study. The first principal component explained the majority of traditional hedonic UX aspects, whereas the second one explained the majority of traditional instrumental qualities of UX.

4.2 Estimating and Predicting UX

In this section, we present results of the empirical study considering the way users and team members assessed UX.

4.2.1 Users' Evaluation on Projects' UX Goals

We asked team members to list one to three of their most and least important UX goals for the project and then compared those goals with users' assessments. In all projects, team members emphasized the importance of pragmatic aspects of UX. The three most often mentioned UX goals were the following: *easy to use* (of the 26 participants, 18 mentioned this), *easy to learn* (13 mentions) and *fast to use* (13 mentions). Each of these three goals was mentioned in all six projects by at least one team member. *Fun* (16 mentions) and *promoting creativity* (13 mentions) were named as the least important UX goals in every project. This result was expected since pragmatic aspects, productivity in particular, are often emphasized in enterprise system development [54]. Similarly, importance of pragmatic aspects was emphasized in a study carried out in metals and engineering industry [5,6].

With this data from developers, we then analyzed how users evaluated those items that teams considered the most and least important UX goals compared to other UX items. Users did not give higher assessments for these dimensions compared to other dimensions; *fast to use* was in fact among the lowest scored items. Users gave the highest evaluations for the following dimensions: *good* (6.1), *useful* (6.1), and *recommendable (6.1*) while the lowest were the following: *fun* (4.5), *promotes creativity* (5.0), aesthetic (5.1), and *fast to use* (5.1). The mean of users' overall UX evaluation was 5.7, while the mean over all the UX dimensions was 5.6.

4.2.2 Differences between Measurements

When evaluating the UX of the outcome, team members were more critical when they were asked to evaluate as they think a member of a particular user group would evaluate (measurement TU) compared to when the team members responded as themselves (measurement TO). The mean evaluations were systematically lower in measurement TU compared to measurement TO. We compared mean values of each item separately per project and found that in 60% of the cases, the mean value in measurement TO was higher than in measurement TU, while the value of measurement TU was higher in only 14% of the cases. All the roles (developers, POs, and UXS) systematically gave lower assessments in measurement TU compared to measurement TO. However, when comparing team members' assessments (TO and TU) to users' assessments (US), only for UXSs and POs did putting themselves in the users' role improve their UX assessments compared to users (measurement TU was closer to measurement US for UXSs). We consider this finding interesting and worth further studies.

There was a statistically significant difference between team members' and users' responses on six UX items when team members were asked to respond as they thought users would respond (comparison of measurements TU and US) (**Error! Reference source not found.**). The equity of distribution across users' (US) and team members' responses was greater when team members were asked to respond as themselves (measurement TO). In the latter case (comparison of measurements TO and US), the null hypothesis remained for all items.

The distribution of cases where user evaluation was higher than team evaluation and vice versa was relatively even when comparing measurement US with measurement TO (US is higher in 47% and lower in 43% of the cases, **Error! Reference source not found.**). However, when comparing measurement US with measurement TU, cases where user evaluation was higher than team evaluation were overly represented. User evaluation was higher in 65% of the cases and lower in 28% of the cases.

Table 7 Distribution of differences between users' (measurement US) and team members' (measurements TO and TU) mean evaluations grouped by the direction of the difference. The mean difference between measurement US and measurement TO or TU is presented in brackets.

	Measurement TO	**Measurement TU**
US is higher (mean difference)	47% (0.6)	65% (0.7)
US is lower (mean difference)	43% (0.6)	28% (0.6)
US and TO or TU are equal	10%	7%

Table 8 Results of tests of equity between user and team responses when team members were asked to respond as they think users would. Test statistics grouping variable is respondent type (user or team member). Rejection of the null hypothesis ($p < .05$) is indicated by emboldening the value.

UX item	**Mann-Whitney U**	**Z**	**Asymp. Sig. (2-tailed)**
Easy to learn – Hard to learn	229.0	-2.67	**< .01**
Fast to use – Slow to use	362.5	-.260	.80
Easy to use – Difficult to use	271.0	-1.87	.06
Reliable – Unreliable	328.5	-.856	.39
Desirable – Undesirable	255.5	-2.136	**<.05**
Recommendable – Not Recommendable	258.5	-2.17	**<.05**
Good – Bad	217.5	-2.88	**<.005**
Useful – Useless	330.5	-.855	.39
Motivating – Discouraging	231.5	-2.60	**<.01**
Fun – Dull	332.5	-.77	**<.05**
Aesthetic – Unaesthetic	321.0	-.98	.46
Professional – Amateurish	334.0	-.779	.44
Convincing – Unconvincing	277.5	-1.80	.07
Presentable – Unpresentable	316.0	-1.08	.28
Promotes creativity – Suppresses creativity	307.5	-1.21	.23
Innovative – Conservative	287.5	-1.56	.12

Based on the above, developers were overly critical with their responses in measurement TU, whereas developers' evaluations corresponded with users' evaluations better when they were not trying to predict the user assessment. In contrast, both POs' and UXSs' assessments were closer to users' assessments when they put

themselves in the users' place. On average, developers assessed UX items 0.3 points lower than users when assessing as themselves (measurement TO) and 0.5 points lower than users when they tried to predict users' assessment (measurement TU) (on a seven-point scale). POs' assessments in measurement TO were 0.2 points higher than users' (US) and in measurement TU 0.1 lower than users' (US), on average. UXSs assessments were on average 0.2 points higher than users' in measurement TO and 0.1 points lower than users' in measurement TU. We consider POs' and UXSs' assessments quite accurate with users' assessments while developers' assessments differed from those of users'. Given that in the participating projects UXSs and POs handled communication with users while developers' understanding of users and their needs remained shallow (Kuusinen 2015), we conclude that trying to empathize with users seems to be unsuccessful with lacking understanding of the user. This finding is in line with [55]. However, our sample included responses only from three UXSs and six POs, and thus we want to be cautious with our conclusions.

In general, team members' evaluations varied more between measurements TO and TU for items measuring non-instrumental quality. We compared team members' responses between measurement TO and TU with Wilcoxon test using the following null hypothesis: "the median of differences between measurement TO and TU for each UX item separately is zero"; that is there is no difference between measurement TO and TU item-wise. The null hypothesis was rejected for the following items:

- good ($Z = -2.83$, $p < .005$)
- motivating, ($Z = -2.94$, $p < .005$)
- fun ($Z = -2.50$, $p < .05$), and
- innovative ($Z = -2.18$, $p < .05$).

Thus, team members changed their evaluation more for abovementioned items.

4.2.3 Assessments of Overall UX and Need Fulfillment

Table 9 Significant correlations (Pearson's r, $p < .1$) between overall UX evaluation scores and measured UX items per measurement. N = 26 in measurements TO and TU and N = 29 in measurement US. Item name is in italics when correlation was found only in measurement US.

Measurement TO		**Measurement TU**		**Measurement US**	
Item	R	Item	R	Item	R
Good	.73	Easy to use	.60	*Presentable*	.71
Desirable	.66	Useful	.59	Innovative	.71
Innovative	.61	Easy to learn	.56	Convincing	.68
Recommendable	.51	Convincing	.55	Easy to use	.68
		Good	.537	Good	.63

	Professional	.52	*Aesthetic*	.62
	Innovative	.50	*Reliable*	.61
			Desirable	.60

The survey asked two questions about overall UX and the scope of software (as responding to needs). To assess overall UX and need fulfillment, we compared evaluations of measured UX dimensions to the evaluation of overall UX with Pearson's product-moment correlation. These results are presented in Table 9.

The following correlations were found only in measurement US: *presentable, aesthetic*, and *reliable*. *Desirable* was found in measurement US but not in measurement TU, and *convincing* was found in measurement US but not in measurement TO. Of the correlated items in measurement US, only "*easy to use*" measures instrumental quality. Thus, non-instrumental aspects correlated with the overall UX assessment clearly more than pragmatic ones. None of the items measuring instrumental quality correlated with the overall UX assessment in measurement TO (team members as themselves). In general, Pearson's r value grew smaller in measurement TU compared to measurement TO, which might indicate that the team members were less confident with their responses in measurement TU.

The following items (Table 10) had a strong and statistically significant correlation with the users' assessment of how well the system fulfills their needs: *Recommendable, useful, motivating, aesthetic, convincing, presentable,* and *innovative*. They all belong to hedonic UX dimensions except *useful*, which is considered to measure the overall quality of the system.

Table 10. Strong and significant correlations of measured items with the user assessment of the system's ability to fulfill user needs.

Item pair	Pearson's r value	2-tailed significance (p)
Presentable – Unpresentable	.78	< .001
Innovative – Conservative	.72	< .001
Useful – Useless	.61	< .001
Recommendable – Not recommendable	.60	< .001
Motivating – Discouraging	.59	< .001
Aesthetic – Unaesthetic	.58	< .001
Convincing – Unconvincing	.54	< .01

5 Limitations

Threats to external validity: We have only studied a restricted set of companies all operating in Finland, which threatens population validity. The number of studied companies was limited to five, and as the data was collected from development projects, the sample is clustered; projects, their outcomes, and users are unique and thus not directly comparable. We utilized the same team population in another study before, which subjects the study to multiple-treatment interference. As the sequence of measurements TO and TU was fixed, the study is prone to order bias.

The data was small (55 participants) for PCA; it would be beneficial to double the number of participants. We based our sampling on [56], where the writers argue for smaller sample sizes, even for samples of 20. Therefore, we consider our sample size sufficient, but also admit that a larger size would have been beneficial. For instance, Gorsuch [57] argues there should always be at least one hundred participants even for a small number of variables. Comrey and Lee [58] consider that having 100 participants is sufficient but poor and a good sample size would be 500 participants.

Threats to internal validity: Selection bias always exists when comparing groups. In this particular setting, utilizing randomized groups was impossible. Measurements TO and TU might be affected by learning effect, as participants answered the same questions twice (as themselves and as they think users would answer). We did not select the user participants by ourselves, and thus we are unaware of the possible level of implementation bias. Although we guided the contact persons in selecting user participants, some of them might have selected, for instance, users that they knew who were positive towards the software. Moreover, we did not control for a user answering the survey twice.

Using semantic differentials is prone to several types of evaluation bias. Those include the following: Central tendency bias occurs when respondents tend to favor the middle levels of a scale [59]. This was also observed in our study. Position bias concerns the order of evaluated items; users tend to treat the middle items differently than those in the beginning and in the end [60]. We did not utilize counterbalancing, which can lead to position bias. PCA is prone to this bias since it can have an impact on the correlations between variables.

6 Discussion

6.1 UX Scale

The scale we utilized showed strong internal consistency in measuring 1 (hedonic qualities of UX) and 2 (instrumental qualities of UX). Internal consistency was acceptable for measuring 3 (scope or overall quality of the system) and questionable for measuring 4 (fitness for professional use). However, internal consistency can be improved by increasing the number of items in the category [62]. It is possible in this case since there were only three items in components 3 and 4. Thus, confirmatory studies should be conducted for further validation of the dimensions of enterprise software UX. Also, different phrasing of items could be tested for improved fitness, and the determined dimensions and their interpretation should be further analyzed.

It seems that some items behave differently when measuring enterprise and leisure software. In leisure software such items as presentable, professional, and innovative (design) have been used for measuring aesthetic quality. For instance, Lavie et al. [31] understand aesthetic in a broad sense, and they divide it to dimensions of *classical* and *expressive aesthetics.* They describe the latter as follows: "*The expressive aesthetics dimension is reflected by the designers' creativity and originality and by the ability to break design conventions*". Especially in enterprise software the UX design

often concentrates on user interaction or UI design. Thus, "breaking design conventions" most probably indicates bad design decisions since design conventions, in style guides for instance, have been created to instruct on developing fluent user interaction [62]. In addition, the phrasing of questions in our study asked the participants to evaluate the system (as a whole) and not just its design or appeal. Thus, the results cannot be directly compared to studies where the appeal of UI designs alone have been evaluated.

6.2 UX as Assessed by Team Members and Users

Based on our findings, it seems likely that developers are able to understand the pros and cons of the developed enterprise software quite well. However, they tend to focus on pragmatic aspects of the system neglecting the non-instrumental ones that in fact seem to be more important to users in terms of their UX. As enterprise software typically provides tools that are used to perform practical tasks, instrumental quality naturally should be sufficient. However, non-instrumental quality contributes to user satisfaction and thus to human productivity, which might be an important organizational goal.

The first principal component (motivation), revealed in our analysis, measured mainly the system's ability to motivate the user while the second one measured usability and is correlated with the user's willingness to use the system. Both these qualities of enterprise software are important for productivity and job satisfaction [63, 64]. Developers seemed to think that users would appreciate especially qualities related to efficiency and productivity. They emphasized instrumental qualities even more when they were asked to assess the system as they think users would. This finding is in line with [45] who found that usability professionals have a tendency towards utilitarian dimensions of usability. Also, Innes [54] argued that developers of ERP systems tend to neglect the hedonic. In our study, in comparing measurements TO and TU (team members assessing UX (TO) in individual roles vs. team members (TU) placing themselves in users' role), it seems that the tendency towards the instrumental was increased when developers were to think how users would assess the UX. However, in users' assessments, the hedonic correlated most with their overall UX evaluation, and in PCA, it was the first component.

Clemmensen et al. [46] did not find many differences between users' and developers' perception on usability. On the other hand, usability professionals construed usability differently from developers and users. Given that only three UXSs participated in our study, we want to be cautious to make generalizations about differences between developers' and UXSs' assessments. However, in our study, UXSs and POs were the best to predict user evaluation of both the pragmatic and the hedonic. Developers tended to be overly negative in their evaluations. Such finding might be explained by the fact that UXSs and POs were the most involved with users and that they thus have the best understanding of user needs and capabilities. However, the frequency of user communication [7] seemed not to improve the ability of POs to predict the UX as assessed by users. Altogether, the small number of POs and UXSs make this finding questionable, and it definitely requires more research.

Developers were more critical towards the UX when they were asked to evaluate the software from the users' point of view compared to their own evaluation and users' evaluation. This finding is interesting considering the common practice among developers to try to think as they believe users would. The result might indicate that if developers do not have a proper understanding of the user, putting oneself in an imaginary user's place seems to lower the ability to predict the actual user evaluation. Again, the small number of UXSs and POs in our study allows only cautious conclusions. However, in our population, putting oneself in the user's place seemed to improve the accuracy of UXSs' and POs' evaluation. This finding provides an interesting opportunity for future work: does, for instance, utilizing personas or exposing developers to users improve the developers' ability to put themselves in the users' place and thus improve developers' ability to predict UX? Another question is if it has an impact whether the UI is designed by the developers themselves or by a UXS. Also, it can have an impact how closely the developers work with the UXS.

Neither users nor teams gave better evaluations to those UX dimensions that the teams considered the most important ones. Teams focused on usability and productivity, whereas affective and aesthetic qualities seemed to better predict the overall UX of the users. Thus, we hypothesize that setting clearer UX goals informed by user preference and shared among the whole project team might improve both the overall UX and the rating of the most important aspects.

6.3 Using the Scale to Focus UX Goals

When designing work related systems, such as enterprise systems, selecting and setting UX goals to support design and development activities is equally important as when designing consumer products [65, 66, 67]. Setting a limited and carefully selected set of high-level UX goals enables focusing the design effort on the most important experiential aspects. A plain list of UX dimensions can be useful when considering the UX goals and setting measurable UX targets for a project. The list itself can act as a constant reminder for developers of the multidimensionality of UX in a similar way the *personas* method is often used. In the personas method, archetypes of users are created based on user data. Descriptions of personas are often hung on walls to remind developers for whom they are developing. In our study, agile teams considered productivity items as the most important UX goals. This finding is in line with Innes [54]. To be able to guide the UX implementation during development, the team needs information on how users perceive the software being developed. In our study, we found that teams considered *fast to use* as one of the most important UX goals, while users considered it as one of the poorest performing dimensions. The team can use this information to focus their improvement work on the experienced speed of use.

It would be interesting to measure if improving a quality with a low evaluation score would improve the overall UX score. On the other hand, another hypothesis could be that improving performance on dimensions with the strongest correlation to the overall UX score would increase the overall UX score. Users might also expect that items related to productivity and efficiency need to be on a sufficient level not to lower the UX. However, after that, other qualities become more important predictors of the perceived UX. Thus, a third hypothesis is that concentrating on items belonging

to the first principal component (motivation) would increase the overall UX assessment of users.

7 Conclusions

We compared UX assessments of members of agile teams and users of the software systems under development. Our results indicate that developers concentrate on instrumental aspects of UX, whereas for users, non-instrumental aspects might be a more important predictor of their perception of overall UX. Moreover, it seems to be difficult for developers to place themselves in the user's position, and thus trying to do so can even be harmful when the team member does not have sufficient understanding of the user. These findings contribute towards understanding development team members' ability to understand UX in order to enable allocating UX tasks between team members and thus focusing the limited UXS resource to those tasks that developers cannot handle by themselves.

Acknowledgments

We thank Timo Partala for us on the data analysis methods. We are grateful to all the study participants. Our research has been supported by TEKES as part of the Cloud Software and Need for Speed research programs of DIGILE (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT and digital business) and by TEKES as part of the User Experience and Usability of Complex Systems (UXUS) research program of FIMECC (Finnish Metals and Engineering Competence Cluster).

References

1. Saiedian, H., Dale, R. (2000). Requirements engineering: making the connection between the software developer and customer. *Information and Software Technology* 42, 6, 419-428.
2. ISO 9241. (1998). Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. Genève, CH: International Organization for Standardisation.
3. Salah, D., Paige, R. and Cairns, P. (2014). A systematic literature review on agile development processes and user centred design integration. *In Proc. of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE'14)*. ACM. Article 5, 10 p.
4. Sy, D. (2007). Adapting usability investigations for Agile user-centered design. *J of Usability Studies* 2, 3, 112-132.
5. Sundberg, H.-R. (2015). The importance of user experience related factors in new product development—Comparing the views of designers and users of industrial products. 23rd *Nordic Academy of Management Conference*, 12-14 August 2015, Copenhagen, Denmark.
6. Sundberg, H.-R. (2015). The role of user experience in a business-to-business context. Doctoral dissertation. Tampere University of Technology. Publication; 1278. http://URN.fi/URN:ISBN:978-952-15-3450-8

7. Kuusinen, K. (2015). Task allocation between UX specialists and developers in agile software development projects. *Proc. Human-Computer Interaction – INTERACT 2015*, LNCS 9298, pp. 27-44, Springer International Publishing
8. Law, E., Roto, V., Hassenzahl, M., Vermeeren, A. and Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. In Proc. CHI'09, ACM, 719-728.
9. Hassenzahl, M. and Tractinsky, N. (2006). User experience - A research agenda. *Behaviour & Information Technology* 25 (2), 91-97.
10. Law, E. L. C., Hassenzahl, M., Karapanos, E., Obrist, M., & Roto, V. (2015). Tracing links between UX frameworks and design practices: dual carriageway. Proc. HCI Korea (pp. 188-195). Hanbit Media, Inc.
11. Hassenzahl, M. (2004). The Interplay of Beauty, Goodness and Usability in Interactive Products. Proc. HCI. Lawrence Erlbaum Associates, 19, 4, 319-349.
12. McCarthy, J., & Wright, P. (2004). Technology as experience. *Interactions*, 11(5), 42-43.
13. Hassenzahl, M. (2008). User experience (UX): towards an experiential perspective on product quality. Proc. 20th International Conference of the Association Francophone d'Interaction Homme-Machine (pp. 11-15). ACM.
14. Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in Human Behavior*, 43, 35-48.
15. Väänänen-Vainio-Mattila, K., Roto, V. & Hassenzahl, M. (2008). Towards practical user experience evaluation methods. EL-C. Law, N. Bevan, G. Christou, M. Springett & M. Lárusdóttir (eds.) *Meaningful Measures: Valid Useful User Experience Measurement (VUUM) (2008)*: 19-22.
16. Diefenbach et al. (2014). Diefenbach, S., Kolb, N., Hassenzahl, M. (2014). The 'Hedonic' in Human-Computer Interaction. Proc. of the 2014 Conference on Designing interactive systems (DIS), pp. 305-314, ACM
17. Vermeeren, A. P., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. Proc. of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (pp. 521-530). ACM.
18. da Silva, T.S., Martin, A., Maurer, F., Silveira, M. (2011) User-centered design and Agile methods: a systematic review. In: Proc. of the International Conference on Agile Methods in Software Development, AGILE 2011, IEEE
19. Holzinger, A. (2005). Usability engineering for software developers. *Communications of the ACM*, 48 (1), pp. 71–74
20. Dray, S., Siegel, D. (2004). Remote possibilities? international usability testing at a distance. *Interactions* 11(2), pp. 10-17. ACM
21. Ivory, M. Y. and Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. Comput. Surv. 33(4), pp. 470-516. ACM
22. Ardito, C., Buono, P., Caivano, D., Costabile, M.F., Lanzilotti, R., Bruun, A., Stage, J., (2011). Usability evaluation: a survey of software development organizations. Proc. of the International Conference on Software Engineering and Knowledge Engineering, SEKE 11. pp. 282–287 Knowledge Systems Institute, Skokie.
23. Lárusdóttir, M. K., Cajander, A, Gulliksen, J. (2013). Informal Feedback Rather Than Performance Measurements – User Centred Evaluation in Scrum Projects. *Behaviour and Information Technology*. 33(11), pp. 1118-1135.
24. Bruun, A., Gull, P., Hofmeister, L., and Stage, J. (2009). Let your users do the testing: a comparison of three remote asynchronous usability testing methods, Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp 1619-1628. ACM
25. Bruun, A., Stage, J. (2012). The effect of task assignments and instruction types on remote asynchronous usability testing Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 2117-2126

26.Salvador, C., Nakasone, A. and Pow-Sang J. A. (2014). A systematic review of usability techniques in agile methodologies. *Proc. 7th Euro American Conference on Telematics and Information Systems (EATIS '14)*. Article 17, 6p. ACM.
27.Panwar, M. (2013) Application Performance Management Emerging Trends. Proc. Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), pp. 178-182.
28.Bastien, J. M. C. (2010). Usability testing: a review of some methodological and technical aspects of the method. *Int. J. of Medical Informatics*. 79(4). pp. e18-e23. Elsevier
29.Bargas-Avila, J., Hornbæk, K. (2011). Old Wine in New Bottles or Novel Challenges? A Critical Analysis of Empirical Studies of User Experience. In: Proc. Annual Conference on Human Factors in Computing Systems. ACM, pp. 2689–2698.
30.Bradley, M.M. Lang, P.J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *J. of Behavior Therapy and Experimental Psychiatry* 25, 49–59.
31.Lavie, T., Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *Int. J. of Human-Computer Studies*, 60, 3, 269-298.
32.Voss, K. E., Spangenberg, E. R., & Grohmann, B. (2003). Measuring the hedonic and utilitarian dimensions of consumer attitude. Journal of marketing research, 40(3), 310-320.
33.Costello, B., and Edmonds, E. (2007). A study in play, pleasure and interaction design. Proc. the 2007 conference on Designing pleasurable products and interfaces (DPPI'07). pp. 76-91. ACM.
34.Zijlstra, R. (1993). Efficiency in Work Behaviour. A Design Approach for Modern Tools. Delft University Press.
35.Jackson, S. Marsh, H. (1996). Development and validation of a scale to measure optimal experience: The Flow State Scale. *J. of Sport and Exercise Psychology*, 18. 17-35.
36.Väätäjä, H., Koponen, T. & Roto, V. (2009). Developing practical tools for user experience evaluation: a case from mobile news journalism. *European Conference on Cognitive Ergonomics (ECCE '09)*. VTT Technical Research Centre of Finland, VTT, Finland. pp. 240-247.
37.Desmet, P., Overbeeke, C., Tax, S. (2001). Designing products with added emotional value; development and application of an approach for research through design. *The Design Journal* 4 (1), 32-47.
38.Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4), 261-292.
39.Kirakowski, J. (1996). The software usability measurement inventory: Background and usage. In P. W. Jordan et al. (Eds.), Usability Evaluation in Industry, Taylor & Francis, 169-178.
40.Davis, F. D. (1989). A technology acceptance model for empirically testing new end-user information systems: Theory and results. *MIS Quarterly*, 13 (3), pp. 319-340
41.Lee, T. M. & Park, C. (2008). Mobile technology usage and B2B market performance under mandatory adoption. *Industrial Marketing Management*, 37 (7), pp. 833–840.
42.Goodhue, D. L. & Thompson, R. L. (1995). Task-Technology Fit and Individual Performance. *MIS Quarterly*, 19 (2), pp. 213-236.
43.Lindgaard, G., & Kirakowski, J. (2013). Introduction to the special issue: The tricky landscape of developing rating scales in HCI. *Interacting with Computers*, 25(4), 271-277.
44.Hertzum, M., Clemmensen, T., Hornbaek, K., Kumar, J., Shi, Q., and Yammiyavar, P. (2011). Personal usability constructs: How people construe usability across nationalities and stakeholder groups. *Int. J. of Human-Computer Interaction*, 27 (8), pp. 729-761.
45.Hertzum, M., Clemmensen, T. (2012). How do usability professionals construe usability? *Int. J. of Human-Computer Studies* 70, 26-42.
46.Clemmensen, T., Hertzum, M., Yang, J., and Chen, Y. (2013). Do usability professionals think about user experience in the same way as users and developers do? Interact 2013, Part II, LNCS 8118, 461-478.

47.Shackel, B. (2009). Usability-Context, Framework, Definition, Design and Evaluation. *Interacting with Computers* 21(5-6), 339–346.
48.Schwaber, K. (2004). Agile project management with Scrum (Microsoft professional), 1st. ed., Microsoft Press.
49.Poppendieck, M. & Poppendieck, T. (2003) Lean Software Development: An Agile Toolkit. Addison-Wesley Professional, 203p.
50.Fitzgerald, B. & Stol, K.-J. (2014) Continuous software engineering and beyond: trends and challenges. In Proc. 1st International Workshop on Rapid Continuous Software Engineering (RCoSE 2014). ACM, New York, NY, USA, pp. 1-9.
51.Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), pp. 433-459. John Wiley & Sons.
52.Nunnally, J. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.
53.Hassenzahl, M. (2005) The Thing and I: Understanding the Relationship Between User and Product. In Blythe, M., Overbeeke, K., Monk, A., and Wright, P. (Eds.). Funology: From Usability to Enjoyment. Kluwer Academic Publishers, 31–42.
54.Innes, J. (2011). Why Enterprises Can't Innovate: Helping Companies Learn Design Thinking. In HCII 2011, LNCS: 6769, 442–448. Springer Berlin / Heidelberg
55.Cockton, G., & Woolrych, A. (2001). Understanding inspection methods: lessons from an assessment of heuristic evaluation. In People and Computers XV—Interaction without Frontiers (pp. 171-191). Springer London.
56.Preacher, K. J., & MacCallum, R. C. (2002). Exploratory Factor Analysis in Behavior Genetics Research: Factor Recovery with Small Sample Sizes. *Behavior Genetics*, 32, 153-161.
57.Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Hillsdale, NJ: Erlbaum
58.Comrey, A. L. and Lee, H. B. A first course in factor analysis. Hillsdale, NJ Erlbaum
59.Yu, J.H., Albaum, G., and Swenson, M. (2003). Is a central tendency error inherent in the use of semantic differential scales in different cultures? *International Journal of Market Research*, 45(2), 213-228.
60.Blunch, N. J. (1984). Position bias in multiple-choice questions. *J. of Marketing Research*, 21, 2, 216-220.
61.Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16(3), pp 297-334. Springer
62.Kuusinen, K. and Mikkonen, T. (2014). On Designing UX for Mobile Enterprise Apps. Proc. Software Engineering and Advanced Applications (SEAA 2014), pp. 221–228.
63.Calisir, F. and Calisir, F. (2004) The relation of interface usability characteristics, perceived usefulness, and perceived ease of use to end-user satisfaction with enterprise resource planning (ERP) systems. *Computers in Human Behavior* 20(4), pp. 505–515. Elsevier
64.Hafeez-Baig, A. and Gururajan, R. (2013). Expectations, Usability, and Job Satisfaction as Determinants for the Perceived Benefits for the Use of Wireless Technology in Healthcare. *Pervasive Health Knowledge Management*, pp 305-316. Springer.
65.Kaasinen, E., Roto, V., Hakulinen, J., Heimonen, T., Jokinen, J.P.P., Karvonen, H., Keskinen, T., Koskinen, H., Lu, Y., Saariluoma, P., Tokkonen, H. and Turunen, M. (2015). Defining User Experience Goals to Guide the Design of Industrial Systems. *Behaviour & Information Technology journal*, Taylor & Francis. DOI: 10.1080/0144929X.2015.1035335
66.Varsaluoma, J., Väätäjä, H., Kaasinen, E., Karvonen, H., Lu, Y. (2015). The fuzzy front end of experience design: Eliciting and communicating experience goals. *Proc. OZCHI 2015*, Brisbane, Australia.
67.Väätäjä, H., Savioja, P., Roto, V., Olsson T. and Varsaluoma, J. (2015). User experience goals as a guiding light in design and development—Early findings. *In INTERACT 2015 Adjunct Proceedings*. Univ. of Bamberg Press (2015), 521-527.