User Experience Evaluation Methods: Current State and Development Needs

Arnold P.O.S. Vermeeren[1], Effie Lai-Chong Law[2], Virpi Roto[3], Marianna Obrist[4], Jettie Hoonhout[5], Kaisa Väänänen-Vainio-Mattila[6,3]

[1]Delft University of Technology
2628 CE Delft, the Netherlands
a.p.o.s.vermeeren@tudelft.nl

[2]University of Leicester
LE1 7RH Leicester, U.K.
elaw@mcs.le.ac.uk

[3]Nokia Research Center
00045 Nokia Group, Finland
firstname.lastname@nokia.com

[4] ICT&S Center, Univ. of Salzburg
5020 Salzburg, Austria
marianna.obrist@sbg.ac.at

[5]Philips Research Laboratories
5656 AE Eindhoven, the Netherlands
jettie.hoonhout@philips.com

[6]Tampere University of technology
33101 Tampere, Finland
kaisa.vaananen-vainio-mattila@tut.fi

ABSTRACT

The recent shift of emphasis to user experience (UX) has rendered it a central focus of product design and evaluation. A multitude of methods for UX design and evaluation exist, but a clear overview of the current state of the available UX evaluation methods is missing. This is partly due to a lack of agreement on the essential characteristics of UX. In this paper, we present the results of our multi-year effort of collecting UX evaluation methods from academia and industry with different approaches such as literature review, workshops, Special Interest Groups sessions and an online survey. We have collected 96 methods and analyzed them, among other criteria, based on the product development phase and the studied period of experience. Our analysis reveals development needs for UX evaluation methods, such as early-stage methods, methods for social and collaborative UX evaluation, establishing practicability and scientific quality, and a deeper understanding of UX.

Author Keywords

User Experience, Evaluation method, Methodological development needs

ACM Classification Keywords

H5.2. User Interfaces (D.2.2, H.1.2, I.3.6): Evaluation/methodology

INTRODUCTION

Although the interest in user experience (UX) in industry and academia is high, there is still a lack of systematic research on how to evaluate and measure UX. Recent guidebooks on UX evaluation are still largely based on basic usability targets [e.g., 26]. A multitude of evaluation methods for usability and to some extent, UX, exist. However, a clear understanding of the current state of UX evaluation methods is yet to be developed. We have identified a need to analyze what UX evaluation methods are currently available, which are missing, and to specify development needs for UX evaluation methods.

What is UX?

According to ISO 9241-110:2010 (clause 2.15), user experience is defined as: *a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service* [9] (for brevity's sake, hereafter the word "product" refers to products, systems, and services). This formal definition is supplemented by other interpretations: User experience explores how a person feels about using a product, i.e., the experiential, affective, meaningful and valuable aspects of product use[1].

UX is generally understood as inherently dynamic, given the ever-changing internal and emotional state of a person and differences in the circumstances during and after an interaction with a product [6, 16]. Therefore, UX should not only be seen as something evaluable after interacting with an object, but also *before* and *during* the interaction. While it is relevant to evaluate short-term experiences, given dynamic changes of user goals and needs related to contextual factors, it is also important to know how (and why) experiences evolve over time. In addition, users' values affect their experiences with products and services, and thus this relationship has to be considered in the design process right from the beginning [15]. These points already make it clear that it is essential to look beyond static aspects and to investigate the temporal aspects of UX – how UX changes over time [e.g., 13, 16]. A thorough understanding of users' experiences, be they positive or negative, a product evokes [5], is at the core of UX evaluation.

NordiCHI 2010, October 16–20, 2010, Reykjavik, Iceland.
Copyright 2010 ACM ISBN: 978-1-60558-934-3...$5.00.

[1] See http://en.wikipedia.org/wiki/User_experience/

What Do We Mean by UX Evaluation Methods?

Before discussing our collection of UX evaluation methods we clarify what we mean by UX evaluation methods, by comparing 'UX' methods to 'usability' methods and by distinguishing 'design' methods from 'evaluation' methods.

Distinguishing UX from Usability Evaluation Methods

The relationship between usability and UX is intertwined. Attempts have been undertaken to demarcate or even dismiss the boundary between them, conceptually and operationally. We take the stance that usability is subsumed by UX. The implication is that UX evaluation entails the augmentation of existing methods for usability evaluation.

Usability tests tend to focus on task performance whereas UX focuses on lived experiences [14]. As UX is subjective [16], objective usability measures such as task execution time and the number of clicks or errors are not sufficient measures for UX: we need to know how the user feels about the system. Although the subjective component of usability (i.e., satisfaction) can be seen as part of UX evaluation, UX addresses a range of other subjective qualities. A user's motivation and expectations play a stronger role in UX than in traditional usability [19].

Distinguishing Evaluation from Design Methods

A sharp distinction between design and evaluation methods is sometimes hard to make. Design methods are often called inspirational or generative methods and aim at bringing inspiration for developers when they create new products and designs [e.g., 3]. We are interested in finding the means to evaluate UX of existing concept ideas, design details, prototypes, or final products. The main focus of evaluation methods is to help in choosing the best design, to ensure that the development is on the right track, or to assess if the final product meets the original UX targets (see e.g., [25]).

MOTIVATION AND STARTING POINT

Whilst the notion of user experience is not entirely new, what can be considered new is the emphasis on its importance over traditional usability. There exist a number of usability evaluation methods (UEMs) and UX evaluation methods with the former being more mature, given years of research efforts in collecting, documenting and categorizing them systematically (e.g., [17] and various websites[2]). In contrast, similar work is yet to be done for UX evaluation methods. Presumably, a clear overview can reveal where the gap lies and inform the future development of these methods. Hence, we are convinced of the need to identify the current state of UX evaluation methods, especially their characteristics and qualities, and are motivated to achieve this aim through various approaches. Specifically, in the last three years, we have collected data on such methods through workshops and Special Interest Group (SIG) sessions at scientific conferences [22, 23, 27], with an online survey, and by looking up the literature.

As the initial step, we searched the literature for overviews of UX evaluation methods. Collections and categorization of UX evaluation methods are rare. Those we found either include other tools and methods besides UX evaluation methods [1, 8, 12] or focus on certain types of evaluation such as assessing momentary emotions [11]. Furthermore, as the research area of UX is evolving rapidly, it is important for us to collect the most recently developed methods from both researchers and practitioners.

Related Method Overviews

Patrick Jordan [12] was among the first ones to list a wider set of methods for designing pleasurable products. His collection consisted of inspirational design methods, methods for evaluating pleasurable aspects of product designs, and examples of multiple-method approaches.

Within the European Union (EU) ENGAGE project a set of UX design and evaluation methods were collected between 2004 and 2006 [1]. The collection can be accessed on the Web[3]. The collected tools and methods were classified into Generative and Evaluative. The Evaluative methods were further categorized in three groups according to what kind of measures the methods focus on: Sensory characteristics, Expression or Meaning, and Emotional reactions.

Another EU activity, HUMAINE, has been collecting and developing design and evaluation methods for affective interactive systems [8]. The HUMAINE website[4] mainly shares information about tools, rather than user study methods for affective systems developers.

Additionally, Isomursu et al. [10] classify some UX evaluation methods that focus on understanding users' emotions aroused while interacting with a system.

Our collection comprises 96 UX evaluation methods. This can well reflect the current state and provide us a solid basis to sustain this research effort in the future. Specifically, based on the descriptions of individual methods, we can derive what kind of methods are scarce or abound, and what their strengths and weaknesses are.

COLLECTING AND ANALYSING UX EVALUATION METHODS

In collecting UX evaluation methods, we were rather liberal about what we mean by 'method'. For example, in some cases UX evaluation formed part of a set of methods or techniques that also evaluated other aspects. In addition, we came across a number of novel UX evaluation methods that are still in their early phases of development. The validity of the findings from many of such methods has not yet been examined. However, since UX is an emerging area and

[2] http://jthom.best.vwh.net/usability/,
http://www.usabilityhome.com/,
http://www.usabilityfirst.com/glossary/cat_66.txt

[3] ENGAGE: http://www.designandemotion.org/society/engage/

[4] HUMAINE: http://emotion-research.net/

innovative methods might lead to some interesting new developments, we decided to include the relatively novel UX evaluation methods that are still in their infancy.

In addition, we included composite methods dealing with UX evaluation, even though some components of the method do not focus on UX evaluation. For example, logging the interaction or monitoring a user's heart rate does not tell how the user feels about the system, but together with the user's authentic comments, e.g. by an interview, they can provide valuable information about a user's feelings in relation to product use. Since several publications report that plain psycho-physiological data are not enough for UX evaluation [e.g., 2, 24], we decided to include in our collection only methods that use these data together with other types of data (such as user's own comments). A multi-method approach allows collection of different types of data, thereby enabling the formation of a big picture of UX [20]. In our collection the class of psycho-physiological methods is listed as one method, instead of as separate methods.

Sources of UX Evaluation Methods

The methods were collected from a variety of sources. The list of 96 methods with their characteristics can be found on the Web at *http://uxems.shorturl.com*. In our analysis, a predefined set of properties of all methods were entered into a template. The template contains data fields for various variables that may characterize a certain method (see Table 1). For methods from the literature, search templates were filled in by one of the six authors of this paper and cross-checked by the others. For the workshops, SIG and survey, templates were filled in by the participants. It should be noted that for some methods not enough information was available to fill in all data fields. The method descriptions thus collected were then analyzed by the authors collaboratively with divergent views being negotiated to reach consensus.

Workshops and SIG Session

In the UXEM'08 workshop participants were asked to write position papers describing UX evaluation methods they have used or developed [27]. Most of the presented methods focused on usability, and only three reported methods were judged to be UX evaluation methods by the organizers and their characteristics were entered into the method description template. In the SIG at CHI'09 [22] participants were asked to describe any choice of method they know or use in practice, using a template with data fields very similar to the ones shown in Table 1. However, it didn't include data fields for the availability of the method and origin of the method, and for some other fields the answer categories were slightly different. Based on the feedback received after the SIG, some categories were improved to increase the clarity of each question (e.g., Period of Experience). The completed templates from the UXEM'08 and SIG sessions were then transposed into the final template form by some of the authors, filling in missing data where possible (e.g., by using references mentioned in the method description). This led to 37 method descriptions. In the UXEM'09 workshop at INTERACT'09 participants were asked to write position papers describing their methods [23]. Moreover, they were asked to describe their method using a template similar to that of Table 1. This yielded 8 additional method descriptions.

Additional Sources

Fourteen methods came from the collection of methods that one of the authors had gathered from the toolsets used by herself and her colleagues in industry.

To further complete our collection of methods we referred back to the existing method collections described earlier on. We identified 15 UX evaluation methods from Jordan's set of methods [12] and included them in our collection. Additionally, 17 methods from the ENGAGE pool of methods [1] were added, as well as 8 methods from the HUMAINE set [8] and 5 methods from Isomursu et al's collection [10]. We looked further up in the literature for possible additional methods. We identified 10 new methods from the UX literature found in the ACM Digital Library.

The final batch of additional 9 UX evaluation methods came from an online survey we conducted. This survey contained questions on the data to be entered into our template. It was publicized through ACM SIGCHI's announcements list, the website of UXNet[5], and the mailing lists consisting of participants to the UXEM'08 workshop and the SIG at CHI'09.

Altogether 123 UX evaluation methods had been collected, which were filtered to eliminate duplicates and irrelevant ones, resulting in the batch of 96 that we have analyzed.

Duplicates and Non-UX Methods

In the process of reaching consensus on method descriptions or categorization of data, some methods that originally seemed to be different were found to be similar or variations of each other. Specifically, various types of Experience Sampling Methods (ESMs) were reported. These methods differed either in what triggered sending experience sampling questions to users (e.g., user actions, context, or time-based) or in the format used for indicating the experience (e.g., free text, scales, and images). The former were considered as distinct methods whereas the latter were seen as variants of the same method.

Some methods on closer inspection were found to be pure usability methods, or not really UX evaluation methods (e.g., pure market research methods, or inspirational rather than evaluative methods).

Data Set and Analysis

The data of the 96 collected methods were analyzed in several steps. First, the dataset was processed

[5] http://uxnet.org/

quantitatively, for identifying interesting patterns in what types of methods are scarce or abound. Then, in content analyses, strengths and weaknesses of the methods were analyzed, identifying needs that should be addressed in future UX evaluation method development. Here the attribute *Period of Experience*, reflecting the dynamic nature of UX, is used for further analysis. Methods deemed uniquely applicable to a specific period (e.g., before, during or after usage) apparently are sensitive to the characteristics of that period. Thus, we differentiated between methods, applying the attribute *Period of Experience*, which consists of five predefined values, viz. (i) Before Usage (prior to interacting with a product/service); (ii) Momentary (snapshot, e.g., emotion); (iii) Single episode in which a user explores design features to address a task goal; (iv) Typical test session (e.g., one hour in which a user performs some tasks; (v) Long-term usage (e.g., interacting with a product/service in everyday life).

A UX evaluation method can be applied to only one or to more of these conditions. It is intriguing to know which UX evaluation methods address only *one* specific type of Period of Experience and what requirements these methods entail. Such requirements can be derived by systematically analyzing strengths and weaknesses of UX evaluation methods (two attributes in free-text format) as described in the template (Table 1, Items 14 and 15).

Categorization of strengths and weaknesses was done in two steps. First, we broke down the text, if not yet in point-form, into independent points and filtered out irrelevant remarks, if any. Second, we iteratively developed a requirement type scheme (Table 2) based on our data, and classified each point for identifying main types of strengths and weaknesses in the application of specific categories of methods.

RESULTS

Our collection of 96[6] methods varied on a number of attributes. First the collection will be discussed based on its attributes. Then, strengths and weaknesses will be analyzed in a content analysis.

Characterizing the Collected Methods

Categories of methods are discussed based on the individual attributes described in the templates (Table 1), including: origin of the method, type of collected data, type of application, information sources, location, period of experience, development phases, and special requirements.

Origin of the Method

Most of the collected methods originate from academia (70%), roughly one-fifth from industry and some from a combined academic and industrial effort (Table 1, Item 13). That the majority of methods is academia-based does not necessarily mean that academia would develop a wider selection of UX evaluation methods than industrial partners who rarely publish the methods they have developed. Although almost half of the participants in our method collection events were from industry, we believe that many industry-based methods remain unrevealed.

Type of Collected Data

About one-third of the methods were reported to provide quantitative data, one-third qualitative data and one-third both (Table 1, Item 9). However, a pure distinction between quantitative and qualitative is sometimes difficult to make (e.g. questionnaires with scales often have a brief follow-up interview to explain findings; in qualitative studies instances are often counted to categorize findings).

Type of Application

Roughly speaking, for each listed type of application an equal number of methods is available (61-88% of the methods per type of application, Table 1, Item 10). Only 22 methods are application-specific in the sense that they can only be used in one or two of the application types. More than two-third of the methods (69 out of 94) are relatively application-independent as they were reported to be suitable for three or more types of application.

Information Sources

The majority of the methods (79) can be used with single users as information sources (Table 1, Item 5). 28 (35%) of these methods can be used in at least one of the early development phases and 15 (19%) are more or less development-phase independent as these can be used in three or more phases (e.g., paired comparison, repertory grid technique, emofaces, exploration test, mindmap, private camera conversation).

Only 13 methods are expert methods of which 6 require users or groups of users in addition to the expert. Seven methods (7%) are expert-based purely (e.g., playability heuristics, property checklists). Four expert methods are relatively development-phase independent and can be used in three or more development stages (i.e., heuristic matrix, perspective-based inspection, expert evaluation, playability heuristics).

Of all methods, 16 (17%) have user groups as a possible source of information (e.g., AttrakWork questionnaire, outdoor play observation scheme, Living Lab, product personality assignment).

Location: Lab, Field or Online?

About half of all methods (46) can only be used in one location: in the lab (21; e.g., facial EMG, controlled observation, TRUE), in the field (24; e.g., immersion, various types of ESM, day reconstruction method, living lab, contextual inquiry) or online (2; AttrakWork, ServUX). Remarkably, amongst the lab-only methods no method can be used with groups of users as information source, and

[6] Note: For some methods it was not possible to enter data about all characteristics. Therefore, depending on the characteristic being discussed, the totals may vary.

CHARACTERISTICS OF UX EVALUATION METHODS			
1. **Name** of UX evaluation method/tool			
2. **Main idea**. Description of the main idea of the method/tool			
3. **General procedure**. Description of the general procedure for applying the method/tool:			
4. **Availability** of the method/tool (n=56)			
Available for free (e.g., published in a journal, on the internet, etc);	66%	Not available (e.g., only internal use/self-developed)	11%
Available under a license;	16%	Not applicable (e.g., unstructured interview)	5%
5. **Information source**. Who provides the UX information that is collected by using the method/tool? (n=96)			
Specific selection of users (1 at a time)	80%	UX experts (no users involved);	14%
Random choice of users (1 at a time)	33%	Pairs of users;	4%
Groups (e.g., focus groups);	17%	Other.	4%
6. **Location** in which the method/tool is used (n=96)			
Lab (researcher's premises);	67%	Online on the Web (n=67)	40%
Field (researcher's choice);	52%	Other	4%
Field (user's own context of use);	44%		
7. **Product development phase**. Which product development phase fits best to use of the method/tool? (n=95)			
Fully functional products	81%	Conceptual design ideas in very early phases of the design process	25%
Functional prototypes	79%	Non-functional prototypes	23%
8. **Period of experience**. What period of experience is studied? (n=95)			
Single behavioural episode with beginning and end (e.g., task or period in which user explores some specific design feature);	63%	Long term (product or service in everyday life)	36%
Typical test session (e.g., one hour of performing tasks) (n=59)	59%	Before usage (n=59)	22%
Momentary (snapshot, e.g., emotion);	45%	Other (n=67)	1 %
9. **Type of collected data** (n=95)			
Quantitative only	39%	Both	30%
Qualitative only	32%		
10. **Applications/designs**. What kind of applications/designs can the method be applied to? (n=94)			
Web services	81%	Hardware designs	66%
Mobile software;	77%	Other (e.g., games)	12%
PC software	76%		
11. **Time requirements**. If you would start to prepare an evaluation now, how many person days will it take to get the results out?			
Minimum (person days, median) (n=61)	1	Maximum (person days; median) (n=36)	7,5
12. **Other requirements**. Conducting the evaluation… (n=93)			
…does not require special equipment (n=92)	67%	…requires a trained researcher	49%
…can be done remotely	51%	…does not require much training;	41%
13. **Origin of the method**. Where was the method/tool developed? (n=77)			
Academia	70%	Both	12%
Industry	18%	Don't know	0%
14. **Strengths** of the method. What are the main strengths of the method/tool?			
15. **Weaknesses** of the method. What are the main weaknesses of the method/tool?			
16. **References describing the method**. Please cite some literature or Web references describing the method/tool.			
17. **References discussing quality issues**. Please cite some literature or Web references discussing validity, reliability or sensitivity (etc.) issues in relation to the method.			
18. **General comments**.			

Table 1. Data fields used in the template for describing UX evaluation methods. The percentages represent the number of methods that scored on each variable.

only two methods can be used for studying long-term usage (i.e., private camera conversation and mental mapping).

For 40 of the methods collected, researchers and practitioners can choose between lab and field. Nineteen methods (20%) are even location independent as they can be used in the lab, in the field as well as for online studies (Table 1, Item 6, *location*; e.g., emotion sampling device, SUMI, paired comparison, intrinsic motivation inventory).

Only 27 methods can be used online. Fourteen of these methods (lab, field, online methods) have single users as their main information source, can only be used in the later two development phases, and can be used for the three periods of experience: momentary, single episode and test session.

Period of Experience

The period of experience (Table 1, Item 8) that can be studied with a method varies. Almost half of all methods (43) can be used to study momentary UX (e.g., various types of ESM, facial EMG). One third of the methods (32) can be used for studying UX of single episodes and test sessions (e.g., group-based expert walkthrough, FaceReader, game experience questionnaire). Remarkably, many questionnaires were reported to be able to deal with all three of these periods (e.g., SAM, USQ, SUMI, presence questionnaire). Only about one-fifth of the methods (13 out of 59) were reported to be able to evaluate the period *before usage* (e.g., Kansei engineering, repertory grid technique, property checklist, fun toolkit), and about one-third (34 out of 95) of the methods can deal with long-term usage (e.g., longitudinal comparison, TUMCAT, www.review.it, evaluation probes).

Development Phases

Most of the methods (about 80%, Table 1, item 7) can be used in the two later development stages, when we can evaluate a functional prototype or product. Of those, 46 methods can *only* be used in those later stages and *not* in the early development phases where only concept ideas or non-functional prototypes are available. Roughly one third (39%) of all methods (37) can be used in at least one of the two early development stages.

Other Requirements

Most methods were reported not to require any special equipment or software (Table 1, Item 12). Remote use of a method (e.g., via a website) is possible in about half of all cases (e.g., multiple sorting methods, ServUX, audio narrative, activity experience sampling, SUMI).

Analysis of Methods for Early Development Phases

Only about one-third of all methods (37) can be used in the early development phases. Since early UX evaluation is important to avoid expensive failures, it is interesting to analyze these methods in more detail.

Early methods: Lab, Field or Online

Of the early development phases, 24 can be used in the conceptual phase and 22 in the non-functional prototype phase. Fifteen early methods can be used in the lab as well as in the field (e.g., Emocards, evaluating UX jointly with usability, Emofaces, Kansei engineering software); 12 are lab-only methods (e.g., multiple sorting method, private camera, conversation, mental mapping), and 8 are field-only methods (e.g., ethnography, longitudinal evaluation, competitive evaluation of prototypes in the wild). Relatively few of the early development phase methods (7) can be applied online. Examples of online methods for early development phases are: prEmo, Emofaces, paired comparison, product personality assignment, sentence completion, www.review.it, QSA-GQM-questionnaire.

Early Methods for Groups of Users

Only 7 early methods use groups as information sources. In our collection, two types of 'group' methods exist: those that study UX in groups of users (as in case of collaborative work), and those that make use of groups as information sources but study individual UX. Only two group-based methods focus on products for use by groups: 'Longitudinal evaluation' and 'Evaluating collaborative user experiences with focus on social interaction and social context'. The rare methods using pairs as information sources do not specifically focus on product use by pairs of users.

Methods and Period of Experience (Content Analysis)

Another important characteristic of a method is, what period of user experience it studies, since methods for evaluating momentary emotions are very different from those evaluating UX over weeks, months, or years. We enumerated the number of UX evaluation methods that uniquely address one type of Period of Experience, and analyzed their strengths and weaknesses to derive requirements, which are categorized according to Table 2. For example, if "no user recruitment required" was reported as a method's Strength, this relates to a requirement of category "Practicability" (see Table 2). With a simulation model based on Chi-square goodness of fit test, 40% was determined as the optimum threshold for a category to be significantly different from the others. Therefore, if a category occupies more than 40% of the total number of requirements, it is deemed as predominant. For instance, for Long-term Usage, there are 15 requirements derived from Strengths, 5 of them are categorized as Utility (33%) and 10 are somewhat evenly distributed over the other four categories. No predominant requirement type emerges. In contrast, out of the 15 requirements derived from Weaknesses, 12 fall into the category of Practicability (80%), one Scientific Quality and two Utility. Obviously, a predominant type can then be recognized (Table 2). For each period of experience such predominant categories were used to further analyze typical Strengths and Weaknesses of the category of methods.

Many of the UX evaluation methods were marked as applicable to Single Episode (e.g., tasks) and Test Session simultaneously, no method focused on Test Sessions only. Hence, we collapsed these UX evaluation methods into one cluster that we named Episode-Test. Besides, another interesting cluster is those UX evaluation methods (mostly questionnaires) that are marked as applicable simultaneously to three types of short-term usage (cf. their long-term counterpart), viz. momentary, single episode and test. Eventually, we came up with five exclusive clusters of UX evaluation methods that uniquely address specific types or combination of types of Periods of Experience. Subsequently we describe each cluster.

Type	Explanation
Scientific quality	Psychometric properties: reliability and validity of the related tool and process
Scoping	Coverage of various facets of real-life UX (e.g., emotion types)
Practicability	Usability (e.g., ease of use), feasibility (e.g., equipment/expertise required) and motivation (e.g., fun)
Utility	Usefulness of evaluative results to stakeholders (e.g., industry/academics)
Specificity	Target at certain domains or user groups

Table 2. Requirement type scheme for analyzing the methods.

Before Usage: Three of the five UX evaluation methods in this category are based on semantic differential technique, one on checklist and one on heuristics. Strengths that are mentioned relate to issues of Practicability mostly: e.g., being fast, cheap (free access, no user recruitment), and easy are mentioned for Playability Heuristics and Property Checklist; not having to rely on statistical analyses is mentioned as one of the positive issue for Product Semantic Analysis. Prominent weaknesses of methods are their scientific quality: low reliability (e.g., repertory grid technique and multiple sorting technique: evaluative criteria vary with users) and questionable validity (e.g., property checklists: experiences as reported or predicted by experts may not represent real user experiences).

Momentary: The 10 UX evaluation methods represent a range of techniques, including questionnaire, self-report, think-aloud, psycho-physiological measures, and heuristics. No strong pattern can be discerned in the Strength and Weaknesses within this cluster. However, many Strengths of UX evaluation methods (36%) in this cluster are seen as having strong scientific quality mainly for reasons relating to validity or validation, e.g., the objective method facial EMG can be used in combination with users' subjective appraisals, the This-or-That method mitigates social desirability with the use of binary scales for young children, measuring physiological responses was reported as being non-disruptive and PrEmo as being well-validated across cultural contexts. While there is no predominant category of weaknesses for this cluster, many weaknesses relate to issues of Practicability (33%) such as specific expertise/equipment/software required (facial EMG, activity experience sampling, PrEmo) and difficult data analysis (sensual evaluation instrument).

Episode-Test: The 21 methods in this cluster also cover a variety of techniques, and their combined uses are more often found (e.g., in situ observations plus retrospective video analysis with users; automatic log with survey). There is a sub-cluster evaluating different types of emotion with simple as well as sophisticated approaches such as providing visual feedback based on integrated physiological data. Unsurprisingly, Practicability issues are of major concern: functional prototype required (e.g., TRUE and the emotion measurement methods 2DES, FaceReader and ESD), domain-specific expertise required (retrospective interview, group-based expert walkthrough) and time consuming video data analysis (OPOS, competitive evaluation of prototypes in the wild). No predominant category of strengths was identified.

Short-term Usage: Interestingly, 10 out of the 18 UX evaluation methods in this cluster are questionnaires of some sort. Scientific Quality is seen as a predominant Strength as well as Weakness. On the one hand, reliability of these measuring scales can be established with statistical manipulation and questionnaires were reported to be validated in many cases (e.g., SUMI, technology acceptance model scale, hedonic/utility scale). On the other hand, validity is seen as a challenge in other cases (e.g., perceived control, PAD, SAM).

Long-term Usage: This category comprises two major schemes: First, measurements take place only after interacting with a product for a relatively long period of time (though the threshold duration remains arbitrary). Second, measurements are undertaken on an ongoing basis for a while. Again, a mix of measuring techniques is employed. It is well anticipated that Practicability is the predominant concern, especially for the second type of long-term usage studies (e.g., longitudinal pilot study, Living Lab study, etc.), such as resourcefulness in terms of time and money. No predominant category of strengths was identified.

DISCUSSION

This section comprises two parts. First, we revisit the five attributes of the requirement scheme, viz. scientific quality, scoping, predictability, utility, and specificity (Table 2), to identify gaps to be bridged for future development of UX evaluation methods. Second, reflecting on the intensive data collection and analysis processes for this study, we address some generic issues pertinent to evaluation methodologies.

Revisiting Requirements Attributes

Specificity: Development needs for group methods

A relatively small category of methods are those with groups of users as their informants. Group-based methods for use in early development phases are even scarcer. Moreover, for most group-based methods, the informants are groups of users, but the evaluation focuses on single users. Only two methods were identified as being capable of explicitly studying experiences of groups of individuals, and for these methods their time-consuming nature was found to be a major concern. Given the ever increasing popularity of virtual communities, social software and collaborative software, there may be a need for practicable group methods.

Scoping: Development need for early-stage methods

Our collection also shows that some types of methods are scarce. Only a few methods are able to evaluate UX in the period before actual use. As ISO states that UX not only applies to actual use, but also to anticipated use [9], it could be worthwhile to develop more methods for that. Those methods that study the period before usage are generally seen as very practicable, but their scientific quality is seen as one of the concerns for using such methods.

In the early phases, there is no functional system that participants could interact with, but they need to use imagination to be able to evaluate the concept or non-functional prototype. Immersion is the only method in the collected method set, which specifically asks the evaluator to imagine how the experience would be like. In this method, the expert evaluator is supposed to keep the concept in mind in her daily life and make notes on the applicability of the concept in different situations. More methods that help imagining and evaluating future experiences would be needed.

Practicability: Streamlining data analysis for online methods
Results of our data analysis show that the use of most methods is not restrained by the type of application being evaluated. Whereas about one-third of the methods can be used in the early phases of a development process, most methods can be used in a later phase. We expected that lab studies would mostly be used in the early phases, but in our method set almost half of all early methods can be used both in the lab as well as in the field. Moreover, there are almost as many early methods that can be used in the field (23 out of 37) as early methods that can be used in the lab (27 out of 37). Furthermore, most methods do not require the availability of special equipment or software. While these findings suggest the flexibility of these methods, practicability is a major concern for many of them. Online UX evaluation methods could have the potential of studying users without having to go into the field. However, especially for the early phases of the development process, online methods are scarce. Whilst some of these online methods are practical because of their being lightweight, cheap and fast, some are problematic because the collected data are unstructured and data analysis would be tediously time-consuming. There is a need to streamline this process.

Utility: Addressing cost-effectiveness of expert reviews
In our method set 13 out of the 96 methods are expert-based methods. However, for 6 of those, users need to be recruited in addition to experts. Seven methods were pure expert-based methods. Expert-based methods were created originally for reasons of practicability, because they are reported to be cheap, fast, and one does not have to recruit users (immersion, property checklist, playability heuristics, expert evaluation). But practicability was also reported as a weakness because of the need to find enough experts with the required expertise, to build a heuristics matrix, and to identify user roles for use in the evaluation (heuristics matrix, perspective-based interaction, expert evaluation). Indeed, for the methods in the period of experience *episode-test*, the need to gather the right domain expertise is mentioned as one of the major concerns (e.g., retrospective interview, group-based expert walkthrough). This issue seems inherited from traditional usability where there are persistent debates about the utility of "discount methods" such as expert reviews. The cost-effectiveness of these types of method needs to be further investigated.

Scientific Quality: Establishing validity
We found that many of the methods that focus on the period of experience *short term usage* are questionnaires. As expected, problems with these are not in their practicability, but a number of them have questionable scientific quality because of a lack of validation studies. Some of these questionnaires have been empirically validated and thus have a high scientific quality. For the same reason, scientific quality is also reported as being high for a number of methods that do snapshot (or momentary) evaluations. However, for some of such methods practicability is a concern in the sense that they require specific equipment, expertise and or software.

Generic Issues for Evaluation Methodologies

We raise the following questions, but may not be able to answer them satisfactorily. Nonetheless, we aim to invite discussions on them with the wider HCI community.

Predefined Measures or Open Evaluation?
Although we did not specifically collect data on UX measures, we assume that one possible way to categorize the UX evaluation methods is to see if they rely on predefined measures for UX or let participants express their experiences in their own words. During our method collection activities, we have noticed that many UX researchers are passionate about having open, qualitative evaluation methods, as predefined metrics may reveal just small parts of the whole UX.

Around half (55) of the methods in our pool do not count on predefined measures but, for instance, let participants describe their experience freely. However, the practicability of methods without predefined measures is lower, since data analysis is harder with qualitative data. Often, specialist know-how is required to draw out findings from qualitative data on some difficult UX aspects such as identification with the product. Companies cannot always afford this during the product development process, but they would benefit from having quick-to-use, validated measures for the different constructs of UX.

A total of 42 of our UX evaluation methods seem to collect UX data via questionnaires, which typically rely on predefined measures. Questionnaires and scales are one of the most versatile but also the most often misused research tools, not only for HCI but also for other domains [4]. It is not quite clear for many of the collected tools to what extent these have been formally tested for validity and reliability. There is a need for researchers to run open evaluations to develop comprehensive and validated UX measures for industry use.

Lab or Field Evaluation?
Since UX is highly dependent on the user's internal state (e.g. motivation) in the current context [5], it is important to

collect UX data in the real contexts of use. The main advantages of field methods are seen in the collection of rich data sets, and in the fact that usage tasks emerge from the users [20]. Evaluation studies conducted in a field setting provide a much more realistic context to obtain reliable UX data, compared to a laboratory environment. However, field methods are often considered to be too time-consuming and resource-demanding, especially for the industrial product development time cycles. Interestingly, the number of lab and field methods in our collection is almost equal (64 vs. 66). The high number of field methods may be due to the high number of methods coming from academia.

Different facets of UX can be evaluated more thoroughly in the field in a real life environment, particularly in the later phases. There is a need to further explore appropriate UX evaluation methods, which are engaging for participants and well-integrated in their daily life. These methods should take people's routines and activities into account (enabling an unobtrusive UX evaluation).

Multi-Method Approaches – When and How?

The benefits in terms of a rich picture of UX and higher scientific quality by collecting data with a combination of UX evaluation methods are well recognized. A common understanding seems to be: the more data is collected, the better. On the other hand, the more data is collected, the more time, resources, and skills are needed in the planning, execution, and analysis phases of the study. Collecting data in various ways often means more work also for the participants, who may become exhausted, potentially compromising the reliability of the data. In the end, there may be too much data from different sources and it would become challenging to consolidate such data and draw solid conclusions. System developers may not, in the end, have the time to utilize more than a fraction of the findings to improve the system. Instead of collecting as much data as possible, we need more guidance on which methods work together well, how to effectively analyze the data from different sources, and what kinds of UX data have been especially useful.

CONCLUSION

In this paper, we report the results of our multi-year effort of collecting user experience evaluation methods both from academia and industry. We now have as many as 96 UX evaluation methods in our collection, with comprehensive information about the type of the method, reported in the Results section. Based on our analysis, we have identified the following needs for methodological developments and further research questions on UX evaluation methods:

1. *Methods for the early phases of development:* How to evaluate concept ideas and non-functional prototypes, when evaluating real use cases in real contexts is not possible?
2. *Validated measures for UX constructs:* Improve the validity of measure-based methods by providing validated measures for different experience focuses and domains, and even for cross-cultural studies.
3. *Methods for social and collaborative UX evaluation*: There is a need for methods to address experiences of groups of individuals. How to evaluate user experience of a group employing online social software in a distributed environment?
4. *Attention for practicability of methods:* For methods to be usefully employed in product development, issues such as resources and skills required, ease of use, ease of data analysis, applicability of results for the development, should be considered.
5. *Effective multi-method approaches:* Which methods work well together? How to effectively collect and analyze the data from different sources?
6. *Deeper understanding of UX:* Development of methods and measures quite often takes place even if the domain itself and theories in the domain are still immature. However, it is important to realize that methods and measure development can substantially be supported by some sound models: as Kurt Lewin already realized: "Nothing is as practical as a good theory"[18].

There is a wide variety of UX evaluation methods deployed in industry and academia. Our mission is to make these methods better known and more accessible to a wider UX community, thereby helping HCI practitioners and researchers to identify the best UX evaluation method for their specific needs (e.g., development phase, kind of experience addressed, and location of UX evaluation). This paper serves to characterize the different types of method available for researchers and practitioners at this point in time. We hope that it will foster the development of UX evaluation methods and prepare the ground for commonly agreed approaches to evaluating users' experiences.

ACKNOWLEDGMENTS

We thank all participants and respondents who have helped us to collect UX evaluation methods over the years. This work was partly funded by a grant from the Academy of Finland.

REFERENCES

1. ENGAGE, *Report on the evaluation of generative tools and methods for 'emotional design'*. Deliverable D15.3. EU project Engage 520998 (2006).
2. Ganglbauer, E., Schrammel, J., Deutsch, S., and Tscheligi, M. Applying Psychophysiological Methods for Measuring User Experience: Possibilities, Challenges and Feasibility. *Workshop on User Experience Evaluation Methods in Product Development.* August 25, 2009. Uppsala, Sweden.

3. Gaver, B., Dunne, T., and Pacenti, E. Design: Cultural probes. *Interactions* 6, 1 (Jan. 1999), 21-29.

4. Green, W., Dunn, G., and Hoonhout, J. Developing the scale adoption framework for evaluation (SAFE). In: *Proc. of the 5th COST294-MAUSE Open Workshop "Meaningful Measures: Valid Useful User Experience Measurement (VUUM)",* Iceland June 2008. Also published in the ACM Library.

5. Hassenzahl, M., and Tractinsky, N., User Experience - a research agenda. In: *Behavior & Information Technology, 25*(2), (2006) pp. 91-97.

6. Hassenzahl, M. 2008. User experience (UX): towards an experiential perspective on product quality. In *Proc. of the 20th international Conference of the Association Francophone D'interaction Homme-Machine*. IHM '08, vol. 339. (2008) ACM, New York, NY, 11-15.

7. Hoonhout, H.C.M. Let the game tester do the talking: think aloud and interviewing to learn about the game experience, In: Isbister, K., Schaffer, N. (eds.), *Game Usability: Advice from the Experts for Advancing the Player Experience,* San Fransisco, CA: Morgan Kaufmann Publishers, (2008) 65-77.

8. HUMAINE D9j: *Final report* on WP9. 30th January, 2008.

9. ISO DIS 9241-210:2010. *Ergonomics of human system interaction - Part 210: Human-centred design for interactive systems* (formerly known as 13407). International Standardization Organization (ISO). Switzerland.

10. Isomursu M, Tähti, M., Väinämö, S., and Kuutti, K. Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies*, Volume 65(4), April 2007, pp. 404-418.

11. Isomursu, M. (2008). User experience evaluation with experimental pilots. In: Väänänen-Vainio-Mattila, K.; Roto, V.; Hassenzahl, M. *Now Let's Do It in Practice: User Experience Evaluation Methods in Product Development*, Workshop at CHI2008 (2008).

12. Jordan, P. *Designing Pleasurable Products.* (2000) Taylor & Francis, London.

13. Karapanos, E., Zimmerman, J., Forlizzi, J., and Martens, J. User experience over time: an initial framework. In *Proc. CHI '09*. (2009) ACM, New York, NY, 729-738.

14. Kaye, J. Evaluating experience-focused HCI. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems CHI '07*. (2007) ACM, New York, NY, 1661-1664.

15. Kujala, S., and Väänänen-Vainio-Mattila, K. Value of Information Systems and Products: Understanding the Users' Perspective and Values, *Journal of Information Technology Theory and Application (JITTA), 9*, 4, 2009.

16. Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., and Kort, J. (2009). Understanding, Scoping and Defining User eXperience: A Survey Approach. *Proc. CHI'09, ACM SIGCHI conference on Human Factors in Computing Systems*.

17. Law, E., Scapin, D., Cockton, G., Stary, M., and Winckler, M. Maturation of Usability Evaluation Methods: Retrospect and Prospect. *COST294-MAUSE Closing Conference Proceedings (2009).* http://141.115.28.2/cost294/upload/533.pdf

18. Lewin, K. *Field theory in social science; selected theoretical papers.* D. Cartwright (ed.). (1951) New York: Harper & Row.

19. Mäkelä, A., Fulton Suri, J. Supporting Users' Creativity: Design to Induce Pleasurable Experiences. *Proc. of the Int. Conf. on Affective Human Factors Design*, (2001) pp. 387-394.

20. Monahan, K., Lahteenmaki, M., McDonald, S., and Cockton, G. 2008. An investigation into the use of field methods in the design and evaluation of interactive systems. In *Proc. of the 22nd British HCI Group Annual Conf. on HCI 2008: People and Computers Xxii: Culture, Creativity, interaction - Volume 1.* (2008) British Computer Society, Swinton, UK, 99-108.

21. Nisbett, R.E. and Wilson, T., Telling more than we can know: verbal reports on mental processes. *Psychological Review*, 1977, 84(3), 231-259.

22. Obrist, M., Roto, V., and Väänänen-Vainio-Mattila, K. User Experience Evaluation – Do You Know Which Method to Use? Special Interest Group in *CHI2009 Conference*, Boston, USA, 5-9 April, 2009.

23. Roto, V., Väänänen-Vainio-Mattila, K., Law, E., and Vermeeren, A. User Experience Evaluation Methods in Product Development. *Workshop in INTERACT'09*, August 25, 2009. Uppsala, Sweden.

24. Shami, N. S., Hancock, J. T., Peter, C., Muller, M., and Mandryk, R. Measuring affect in HCI: going beyond the individual. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*. CHI '08. (2008) ACM, New York, NY, 3901-3904.

25. Stone, D., Jarrett, C., Woodroffe, M., and Minocha, S. *User Interface Design and Evaluation* (The Morgan Kaufmann Series in Interactive Technologies). (2005) Morgan Kaufmann.

26. Tullis, T., and Albert, B. *Measuring the User Experience. Collecting, Analyzing, and Presenting Usability Metrics.* (2008) Morgan Kaufmann.

27. Väänänen-Vainio-Mattila, K., Roto, V., and Hassenzahl, M.. Now Let's Do It in Practice: User Experience Evaluation Methods in Product Development. In *Extended Abstract of CHI 2008,* ACM Press (2008), pp. 3961-3964.