Journal of Usability Studies
Vol. 12, Issue 3, May 2017 pp. 133–154

Lab Testing Beyond Usability: Challenges and Recommendations for Assessing User Experiences

Carine Lallemand
Postdoctoral research associate
University of Luxembourg
ECCS research unit
Esch-sur-Alzette
Luxembourg
carine.lallemand@uni.lu

Vincent Koenig
Senior Lecturer
University of Luxembourg
ECCS research unit
Esch-sur-Alzette
Luxembourg
vincent.koenig@uni.lu

Abstract

In the "third wave" of human-computer interaction (HCI), the advent of the conceptual approach of UX broadens and changes the HCI landscape. Methods approved before, mainly within the conceptual approach of usability, are still widely used, and yet their adequacy for UX evaluation remains uncertain in many applications. Laboratory testing is undoubtedly the most prominent example of such a method. Hence, in this study, we investigated how the more comprehensive and emotional scope of UX can be assessed by laboratory testing.

In this paper, we report on a use case study involving 70 participants. They first took part in user/laboratory tests and then were asked to evaluate their experience with the two systems (perceived UX) by filling out an AttrakDiff scale and a UX needs fulfillment questionnaire. We conducted post-test interviews to better understand participants' experiences. We analyzed how the participants' perceived UX depends on quantitative (e.g., task completion time, task sequence, level of familiarity with the system) and qualitative aspects (think aloud, debriefing interviews) within the laboratory context.

Results indicate that the laboratory setting has a strong impact on the participants' perceived UX, and support a discussion of the quality and limitations of laboratory evaluations regarding UX assessment. In this paper, we have identified concrete challenges and have provided solutions and tips useful for both practitioners and researchers who seek to account for the subjective, situated, and temporal nature of the UX in their assessments.

Keywords

user experience, user testing, evaluation, laboratory evaluation, psychological needs

Introduction

Following the "era" of usability and user-centered design, the human-computer interaction (HCI) field has recently entered the era of user experience (UX) and experience design (Hassenzahl, 2010). This conceptual shift to a more comprehensive and emotional scope of human-computer interactions has been accompanied by the development of new or adapted methods for the design and evaluation of interactive systems (Roto, Obrist, Väänänen-Vainio-Mattila, 2009; Vermeeren et al., 2010). These novel methods mainly aspire to cope with the complexity and subjectivity of UX, as compared to the more objective view of usability. However, a majority of these new methods need more time for consolidation and are slowly transferred into practice (Odom & Lim, 2008).

At the moment, established HCI evaluation methods—such as ex-situ (off-site) user testing or expert evaluation—tend to remain standard practices in both research and practice (Alves, Valente, & Nunes, 2014). In this paper, we use the term "established" to refer to widely used and accepted methods that were proven to conform with accepted standards in our field. A majority of current and established user-centered evaluation methods were developed as usability evaluation methods; yet several studies have shown that these usability methods are now used by extension for the evaluation of UX (Alves et al., 2014). As established HCI evaluation methods are still in use, one can wonder how the shift to UX influences these methods: What are the challenges that experts are facing when evaluating UX using the "usual" methods they were trained to use? What practices remain unchanged yet effective, valid, and reliable? What are the new requirements for UX evaluation?

In this study, we investigated through a UX evaluation use case how UX alters user testing in a laboratory setting (i.e., a controlled environment where an evaluator observes how users interact with a system and collects data about the interaction). In the first section of this paper, we describe how UX raises new challenges and questions about the topic of evaluation. We then report on the experiment we conducted and finally discuss the results from a methodological perspective.

From the Evaluation of Usability to the Evaluation of User Experience

Since its inception, the field of HCI has been primarily concerned with the design of usable systems whose evaluation has been the main focus rather instrumental concerns such as effectiveness, efficiency, or learnability (ISO 9241-11, 1998). The most widely used usability evaluation methods were usability testing and inspection methods (Cockton, 2014). During the last decade, the emergence of UX as a key concept opened up both exciting perspectives and hard challenges. Concurrently with numerous attempts at scoping and defining UX (Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009), a broad discussion on UX evaluations and measures rapidly appeared on the research agenda (Hassenzahl & Tractinsky, 2006; Law, Bevan, Christou, Springett, & Lárusdóttir, 2008). However, the diversity of definitions and interpretations of what constitutes UX (Lallemand, Gronier, & Koenig, 2015) along with the complexity of UX attributes and consequences make it difficult to select appropriate UX evaluation methods (Bevan, 2008). Despite sharing common grounds with the concept of usability, UX spans further by also including emotional, subjective, and temporal aspects involved in the interaction between users and systems (Roto, Law, Vermeeren, & Hoonhout, 2011). UX is more holistic and thus more complex. Researchers generally agree that UX is subjective, holistic, situated, temporal, and has a strong focus on design (Bargas-Avila & Hornbaek, 2011; Roto et al., 2011).

Topics such as interaction meaning, temporal dynamics of an experience, or needs fulfillment through the use of technology challenge the evaluation of UX to an extreme.

To account for the richness and complexity of experiences, UX research attempts to produce viable alternatives to traditional HCI methods. Researchers have thus responded to the challenges underlying UX by developing new methods; nearly 80 of them have been identified and categorized in 2010 (Roto et al., 2009; Vermeeren et al., 2010) and many more methods have been developed during the last four years. Regrettably, novel UX evaluation methods are rarely validated (Bargas-Avila & Hornbæk, 2011) and are slowly transferred into practice (Odom & Lim, 2008). This is partly due to the demands of novel UX methods that still need to be adapted to the requirements of evaluation in an industrial setting (Väänänen-Vainio-Mattila,

Roto, & Hassenzahl, 2008). UX being commonly understood by practitioners as an extension of usability (Lallemand et al., 2015), established usability evaluation methods that remain standard practice for the evaluation of UX (Alves et al., 2014). Bargas-Avila and Hornbæk (2011) reviewed 66 empirical studies on UX and concluded that the most frequent UX evaluation pattern is a combination of "during and after measurements – similar to traditional usability metrics, where users are observed when interacting and satisfaction is measured afterwards" (p. 2694).

UX and Laboratory Evaluation Practices

A laboratory evaluation refers to the evaluation of human-computer interactions in a controlled environment where the evaluator monitors the use of a system, observes users' actions and reactions, and assesses users' feelings about the quality of the interaction. Laboratory evaluations are generally opposed to in-situ (also called "field" or "in-the-wild") evaluations that involve assessing the interaction in its real context of use. Laboratory evaluation sessions generally involve the use of a combination of methods (also known as mixed-methods), the more typical being scenarios of use to observe how users operate (both in a non-interfering way and a posteriori based on video and sound recording of user behavior), think-aloud protocols to capture users' immediate experience, questionnaires to provide a standardized quantitative measure of factors of interest, log file analysis, and finally debriefing interviews. The defining characteristic of user tests is, nevertheless, a concrete system use (Hertzum, 2016).

During the third wave of HCI (Bødker, 2006), new topics such as UX or ubiquitous computing have shaken up established design and evaluation methods. While controlled experiments used to be the gold standard in many disciplines, a recent trend in our field claims for more naturalistic evaluation approaches (Rogers, 2011; Shneiderman, 2008; see also Crabtree et al., 2013). A passionate debate notably animated the Ubicomp community following the publication of Kjeldskov et al.'s intentionally provocative paper "Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field" (2004), where the authors claim that field studies bring not much added value to the usability evaluation process. In the field of UX, the laboratory setting has been described as less effective for evaluating UX than it is for evaluating usability (Benedek & Miner, 2002). With the acknowledgment of the temporal and contextual factors underlying UX, the "turn to the wild" movement has gained influence in research (Rogers, 2011).

Surveys on UX practice show that field studies are considered the most important practice, though they are not widely used (Vredenburg, Mao, Smith, & Carey, 2002). Laboratory evaluations therefore remain common practice, even if more sophisticated tools have now stepped into the lab to support the evaluation of human-computer interactions. The development of psycho-physiological measurements such as eye-tracking, skin conductance activity, or facial expression analysis software and devices allow for an in-depth investigation of human cognitive and emotional processes involved in UX. HCI researchers can of course take advantage of these new methods, though they have to be aware of their limitations and pitfalls (Park, 2009), especially linked to data misinterpretation. Besides these technological tools, new self-reported evaluation scales and questionnaires have been developed (or imported from other fields) to assess several facets involved in UX, such as emotions (Desmet, 2003), hedonism (Hassenzahl, Burmester, & Koller, 2003), aesthetics (Lavie & Tractinsky, 2004), values (Friedman & Hendry, 2012), desirability (Benedek & Miner, 2002), or psychological needs (Hassenzahl, 2010; Sheldon, Elliot, Kim, & Kasser, 2001).

To overcome the limitations of controlled ex-situ experiments, Kjeldskov et al. (2004) proposed to enhance the realism of laboratory setups by arranging the space so as to recreate realistic contexts of use. They successfully recreated a healthcare context to test the usability of a portable working device. While appealing, this idea quickly becomes limited when considering large-scale environments or mobility practices. Technological tools could be used in the lab to cope with the issue, for instance, through the use of simulators or augmented reality devices (Kjeldskov & Skov, 2007).

To summarize, it seems that UX rekindles discussions on field versus laboratory studies. It has also changed laboratory evaluations by promoting the multiplication of measuring devices. The scope of this study however is not to discuss additionally deployed tools but rather the impact of UX on the general principles of laboratory studies.

This study investigates how the shift to UX influences user testing in a controlled setting by thoroughly analyzing the processes and outcomes of laboratory UX testing. Based on the findings derived from our use case study, we aim at identifying the respective strengths and weaknesses of laboratory testing when it comes to the evaluation of UX as compared to the evaluation of usability. Knowing more about the new set of challenges we have to address when assessing UX will allow us, as researchers, to suggest ways of adapting research methods and evaluation practices to the particular characteristics of UX. Thus, we present the results of our UX evaluation use case first to serve as a basis for our subsequent methodological analysis.

Methods

To explore how the conceptual shift to UX could alter laboratory evaluations, we conducted user-testing sessions in a laboratory setting to evaluate UX. Seventy users first took part in user tests and then were asked to evaluate their experience with two systems (the e-commerce website Amazon and a digital camera) by filling out the AttrakDiff scale and a UX needs questionnaire (adapted from Sheldon et al., 2001, see Appendix A). We instructed the participants to think out loud during all of the precited steps, including the completion of the questionnaires. We also conducted post-test interviews in order to better understand participants' experiences.

Participants

Seventy participants (36 males, 34 females) were recruited through several channels (e.g., mailing list, social networks, advertisement in public places) and received €30 in compensation for their time spent. The sample's mean age was 29 (*Min* = 18, *Max* = 48). Regarding their employment status, 50% were employed, 48.6% were students, and 1.4% unemployed. Almost all participants declared feeling at ease with technology (*M* = 5.84 on a 7-point Likert scale, *SD* = 1.22). Regarding the use cases, 83% of the participants were registered on Amazon for more than a year. The average level of familiarity with Amazon's website on a 5-point scale was relatively high (*M* = 3.74, *SD* = 1.09). Participants' average level of familiarity with digital cameras (in general) was also assessed as relatively high (*M* = 3.41, *SD* = 1). Amongst camera owners, 30% of the participants use their camera less than once a month (*n* = 21) and only 7.1% use it several times a week.

Materials

First, we welcomed the participants and explained the functioning within the laboratory. The participants were then made familiar with our strict ethical requirements and signed an informed consent form. After having presented the experiment's general instructions, we asked them to complete a preliminary survey including variables such as age, gender, employment, and familiarity with technology. All materials were in French.

Use cases and testing scenarios

During the user test, the participants had to assess two interactive systems: the e-commerce website Amazon.fr and an Olympus digital compact camera. The choice of these specific use cases was based on a previous study where UX experts were asked to conduct an expert evaluation on four interactive systems, including Amazon and the camera (Lallemand, Koenig & Gronier, 2014). The two systems were presented in a counterbalanced order to avoid sequence biases by distributing practice effects equally across conditions.

In order to stimulate the exploration of the systems, we defined scenarios and tasks, chosen to represent the main actions performed by users on such systems. Five scenarios were related to Amazon and seven scenarios were related to the camera (Table 1). To enhance the ecological validity of the assessed experience, we asked the participants to log in using their own Amazon account. The suggestions made by Amazon's recommender system were therefore real suggestions based on prior items viewed or bought by each user.

Table 1. Scenarios of Use

Amazon (5 scenarios)	Digital camera (7 scenarios)
• Exploring featured recommendations on the home page and adding one item to a wish list • Searching for a book and looking inside to read some pages • Consulting customers' reviews • Adding the book to the shopping cart • Browsing the shop for a pair of shoes	• Taking a picture in Auto mode • Making a short movie • Entering the gallery view • Deleting the movie • Taking a picture in "Magic" mode • Setting image size • Exploring the Help menu

Participants were aware that the performance was not assessed and were instructed to work through the scenarios without time or failure pressure. To encourage a realistic user experience, we also asked participants to freely explore each system before starting the scenarios. After task completion, we collected the participants' opinions during a debriefing interview.

System evaluation: Assessment of UX using the AttrakDiff scale

We decided to assess participants' experiences by using a standardized UX measurement questionnaire. After having achieved the scenarios of each use case, participants were asked to answer the AttrakDiff scale. Hassenzahl, Burmester, and Koller (2003) developed the AttrakDiff questionnaire to measure both the pragmatic and hedonic quality of an interactive product. The measurement items are presented in the format of 28 semantic differentials. The evaluated system's qualities are Pragmatic Quality, Hedonic Quality (subdivided into Hedonic-Stimulation and Hedonic-Identification), and finally Attractiveness. Please note that, according to Mahlke's theoretical model (2008), we considered usability as being measured through the Pragmatic Quality subscale of the AttrakDiff scale.

As all materials were in French, we used the French version of the AttrakDiff (Lallemand, Koenig, Gronier, & Martin, 2015). After having checked the reliability of each AttrakDiff subscale (Table 2), we computed mean scale values for each subscale by averaging the respective items for each participant, and a global AttrakDiff score (ATD_TOTAL) by averaging the values of the four subscales.

Need fulfillment: Assessment of a specific UX-related factor

Beyond the holistic UX assessment provided by the AttrakDiff scale, we also wanted to include an additional UX measure, specifically focused on the fulfillment of psychological needs. Many studies in positive psychology or UX (Hassenzahl, Diefenbach, & Göritz, 2010; Sheldon et al., 2001) suggest that the fulfillment of human psychological needs could act as one of the main drivers for a positive experience. Researchers following this perspective therefore assume that that a system able to fulfil the need for relatedness, the need for competence, or the need for autonomy (to just name a few) will support an optimal and engaging user experience.

Evaluation of needs fulfillment using the UX needs scale. We assessed need fulfillment using an adapted and translated version of the scale developed by Sheldon et al. (2001; see Appendix A). Thirty items divided into seven subscales were used to assess the fulfillment of seven basic needs: Competence (5 items), Autonomy (4 items), Security (5 items), Pleasure (4 items), Relatedness (4 items), Influence (4 items), Self-Actualizing (4 items). We asked the participants to rate the fulfillment of their psychological needs using a 5-point Likert scale (from 1 *Not at all* to 5 *Extremely*). After having checked the reliability of each UX need subscale, we computed mean scale values for each need by averaging the respective items for each participant, and a global need fulfillment score by averaging all items.

Importance of needs fulfillment. Based on the assumption that some needs could be perceived as more important than others depending on the system and the context, we asked participants to report on a 5-point Likert scale (see Appendix B) how important they assessed each need in the context of the interaction with either Amazon or a digital camera.

Results

We used univariate statistics to examine the means and standard deviations of each item as well as to check for possible outliers or entry errors. No outliers or entry errors were found. We used SPSS v22 software to perform statistical analyses.

User Experience Assessment Using the AttrakDiff Scale and the UX Needs Scale

Overall, Amazon was positively assessed on the AttrakDiff scale (M = 4.88), especially regarding its Pragmatic Quality (M = 5.48) and Attractiveness (M = 5.13; Table 2). The UX of the digital camera was also positively assessed (M = 4.35), with the highest rating on Attractiveness (M = 4.71) and the lowest rating on Hedonic-Stimulation (M = 3.85).

Regarding the fulfillment of UX needs, results show that the need that is best fulfilled by Amazon is the need for Security (M = 3.93), whereas the need that is least fulfilled by Amazon is the need for Relatedness (M = 2.12). Regarding the digital camera, the need that is best fulfilled is also the need for Security (M = 3.44), and the least fulfilled is the need for Influence (M = 2.04). As expected, the fulfillment of UX needs is strongly correlated to the perceived UX assessed through the AttrakDiff scale and this holds true both for Amazon, $r(68) = .50$, $p < .001$, and the camera, $r(68) = .65$, $p < .001$.

Table 2. AttrakDiff and UX Needs Scores: Descriptive Statistics and Reliability Analyses

	Amazon website					**Digital camera**				
AttrakDiff scores	**Min**	**Max**	**M**	**SD**	**Cronbach's alpha**	**Min**	**Max**	**M**	**SD**	**Cronbach's alpha**
AttrakDiff global score	3.50	6.46	4.88	0.64	.88	2.39	5.96	4.35	0.83	.92
Pragmatic Quality	3.14	6.86	5.48	0.79	.75	1.57	6.71	4.47	1.18	.87
Hedonic-Stimulation	2.29	6.43	4.16	0.89	.74	1.86	6.29	3.85	1.11	.87
Hedonic-Identification	2.57	6.71	4.76	0.77	.67	2	6.14	4.36	0.85	.70
Attractiveness	2.71	7	5.13	0.89	.86	2.14	6.71	4.71	1.13	.90
UX needs subscales										
Competence	1	5	3.62	0.9	.85	1	5	3.2	1	.90
Autonomy	1	5	3.62	0.9	.74	1.25	5	3.34	.95	.79
Relatedness	1	3.75	2.12	0.95	.82	1	5	2.67	1	.83
Pleasure	1	5	2.81	1.05	.83	1	5	2.71	1.06	.87
Security	2	5	3.93	0.7	.70	1.2	5	3.44	.92	.83
Influence	1	5	2.38	1.04	.86	1	4.25	2.04	.95	.86
Self-Actualizing	1	5	2.62	0.94	.79	1	4.75	2.51	.87	.70

Note. N = 70

Finally, we asked users to rate how important the fulfillment of each need is in the context of an interaction with Amazon or in the context of an interaction with a digital camera. This rating is essential to interpret the results of the UX needs scale: Whenever one wants to assess a dimension of UX, one should also understand how important or meaningful this specific aspect is to the user. Some needs could score low on the needs scale (and one could be tempted to come up with suggestions to improve this dimension) yet this might not influence the experience if these needs are assessed as less (or not) important in the context of the interaction under study. What matters here is the adequacy between the perceived importance of a need and its actual fulfillment through the interaction (Figure 1).

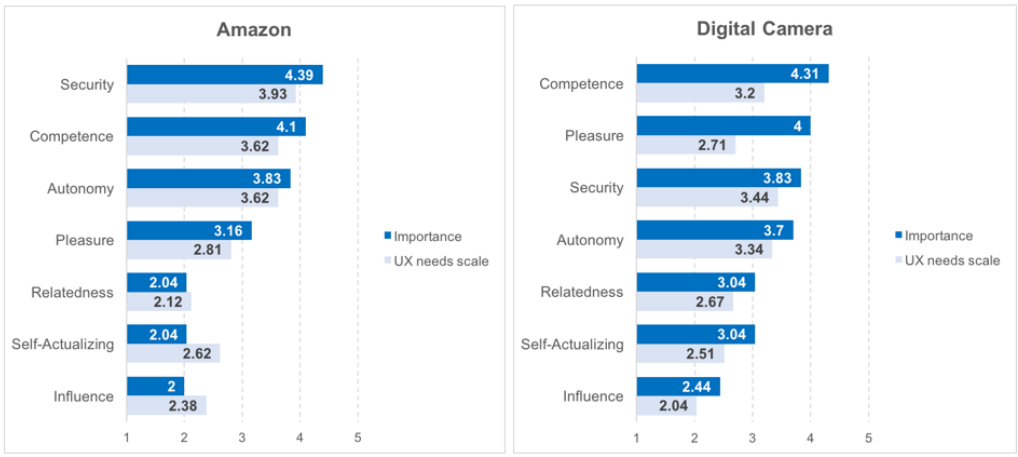

Figure 1. Comparison between perceived importance and perceived fulfillment of UX needs. Needs are presented by order of average perceived importance.

According to the participants, the most important needs to be fulfilled when interacting with a website such as Amazon are the needs for Security (M = 4.39, SD = 0.95), Competence (M = 4.10, SD = 0.98), and Autonomy (M = 3.83, SD = 1.03). Regarding the camera, the most important needs to be fulfilled are Competence (M = 4.31, SD = 0.75), Pleasure (M = 4, SD = 0.99), and Security (M = 3.83, SD = 1.1). In both cases, we notice that the needs for security and competence are in the top three of most important needs to be fulfilled. As Figure 1 shows, in the case of the camera the needs were always rated as more important than actually fulfilled. This could suggest that users' expectations about the UX of the camera were not satisfied. Noteworthy differences between importance and presence are observed for the needs of competence, pleasure, self-actualizing, or relatedness. The results are different in the case of Amazon with a better balance between needs fulfillment and needs importance; some needs being assessed as equal or even more fulfilled than important.

Impact of Task Completion Time and Sequence on Perceived UX

On average, participants worked through the Amazon test scenarios in 11 minutes (Min = 5, Max = 26, SD = 4). The time spent on a task is generally associated with the usability of a system. Table 3 presents the correlations between time spent on a scenario, AttrakDiff ratings, and UX needs fulfillment ratings. As expected, the duration needed by each user to complete the Amazon scenario correlates negatively with the AttrakDiff pragmatic scale, $r(68)$= -.28, p = .018. However, there is no correlation with AttrakDiff's Hedonic or Attractiveness subscales. Similarly, there is no correlation between the time spent on the Amazon scenarios and UX needs scale or subscales: The duration of the tasks apparently does not influence the perceived fulfillment of basic needs. Regarding the digital camera, participants achieved the scenarios in 11 minutes on average (Min = 5, Max = 29, SD = 4). The time spent on the testing scenarios is negatively correlated to the AttrakDiff's pragmatic subscale, $r(68)$ = -.24, p = .044, and positively correlated to the Hedonic-Stimulation subscale, $r(68)$ = .24, p = .047. These results suggest that the more time a user spent on the scenarios, the less the camera is assessed as pragmatic, but the more it is assessed as stimulating. This is an interesting finding that tends to contradict the common interpretation of usability metrics related to the efficiency of a system. While the best efficiency is usually thought as a goal to reach for maximizing the usability of a system, this case shows that additional matters are at stake when it comes to assessing UX rather than usability only. Similar to the Amazon use case, we found no correlation between time spent on the camera scenarios and UX needs scale or subscales.

Table 3. Correlations Between Time Spent on a Scenario and UX Ratings (AttrakDiff and UX Needs Fulfillment)

	Amazon	**Camera**
AttrakDiff - Pragmatic Quality	Negative r(68) = -.28, *p* = .018	Negative r(68) = -.24, *p* = .044
AttrakDiff - Hedonic Quality	NS	Positive with Hedonic-Stimulation r(68) = .24, *p* = .047
AttrakDiff - Attractiveness	NS	NS
UX needs fulfillment	NS	NS

Note. NS = non-significant correlation

We also attempted to understand whether the order in which the systems were presented to the user would influence the perceived UX. Independent-samples t-tests were conducted to compare the effects of testing sequence on the evaluation of UX and fulfillment of needs. Several significant order effects were observed (Table 4). When participants interacted with Amazon as a second use case (Order 2), they assessed it more positively on the AttrakDiff scale (*M* = 5.03) than in the condition where they interacted with Amazon first (*M* = 4.73). At the subscale level, interacting with the camera first and Amazon afterwards led to a better evaluation of Amazon's Pragmatic Quality (*M* = 5.74) than in the case where Amazon was experienced first (*M* = 5.22). The same tendency was observed for the reported fulfillment of UX needs. The order in which systems were presented therefore affected the UX: The difference in perceived UX is higher when the "better" system is presented first. While this of course reminds us to be cautious when designing experiments (e.g., randomizing testing order), it also raises an additional question: How do the users' previous immediate experiences influence an evaluation, and how should we cope with this potential issue?

Table 4. Independent-Samples *T*-Tests Comparing the Effects of Testing Sequence on the Evaluation of UX and Fulfillment of Needs.

		Order 1 (Amazon–Camera)	**Order 2(Camera –Amazon)**	**Diff**	**T-test**
Amazon	AttrakDiff global score	**4.73 (SD=0.57)**	**5.03 (SD=0.67)**	**-.30**	*t*(68)=-2.06 *p* =.043
	AttrakDiff – Pragmatic Quality	**5.22 (SD=0.79)**	**5.74 (SD=0.72)**	**-.53**	*t*(68)=-2.91 *p* =.005
	AttrakDiff - Hedonic Quality	4.36 (SD=0.75)	4.56 (SD=0.9)	-.20	NS
	AttrakDiff - Attractiveness	4.99 (SD=0.79)	5.28 (SD=0.98)	-.29	NS
	UX needs fulfillment	**2.80 (SD=0.63)**	**3.23 (SD=0.63)**	**-.43**	t(68)=-2.87 *p* =.005
Digital Camera	AttrakDiff global score	4.27 (SD=0.88)	4.43 (SD=0.77)	-.15	NS
	AttrakDiff – Pragmatic Quality	4.22 (SD=1.35)	4.72 (SD=0.95)	-.50	*t*(68)=-1.79 *p*=.078
	AttrakDiff - Hedonic Quality	4.19 (SD=1.02)	4.03 (SD=0.95)	-.16	NS
	AttrakDiff - Attractiveness	4.5 (SD=1.11)	4.92 (SD=1.13)	-.42	NS
	UX needs fulfillment	2.79 (SD=0.68)	2.9 (SD=0.59)	-.11	NS

Note. Significant differences are presented in bold. NS = non-significant differences

Our results therefore suggest that both the duration needed for a user to complete a scenario and the sequence in which systems were assessed influenced perceived UX.

Impact of the Level of Familiarity with the Systems on Perceived UX

The self-reported level of familiarity with Amazon is positively correlated to the AttrakDiff global rating, *r*(68) = .44, *p* < .001, and its subscales, especially the Hedonic-Identity subscale, *r*(68) = .47, *p* < .001. The more familiar users were with Amazon, the more they reported the experience as positive. In the case of the digital camera, familiarity with technology is negatively correlated to the evaluation of the Hedonic-Stimulation quality of the device, *r*(68)= -.24, *p* = .041. The more familiar users were with technology, the less they were stimulated by this camera.

Familiarity with technology is not correlated with any need fulfillment in the case of Amazon, and only correlated with the fulfillment of the need for competence in the case of the camera, $r(68) = .27$, $p = .021$. The more users felt at ease with technology, the more competent they felt using the camera. Neither level of familiarity with Amazon nor opinions about Amazon are correlated to the fulfillment of UX needs. Level of familiarity with digital cameras is correlated to the fulfillment of the need for security while using the camera, $r(68)= .25$, $p = .040$.

Understanding Participants' Experiences through Debriefing Interviews

In addition to the quantitative metrics presented above, we interviewed all the participants in order to further understand the adequacy of the laboratory situation for the assessment of the participants' experience.

Single-word experience description

We first collected participants' feelings by asking them to describe their experience with each of the two assessed systems using a single word (Table 5). Amongst 70 words collected (one per participant) to describe the UX of Amazon, 83% had a positive meaning, 13% were neutral, and 4% had a negative meaning. Most cited words were "practical" (11 occurrences), "effective" (8 occurrences), or "good" (8 occurrences). Opinions regarding the digital camera were more heterogeneous with 47% of positive words, 20% of neutral words, and 33% of negative words. Most cited words were "satisfying" (6 occurrences), "novel" (4 occurrences), or "banal" (3 occurrences).

Table 5. Single-Word UX Description for Each of the Two Use Cases

Single-word UX description	**Amazon**		**Camera**	
	Frequency	Valid percentage	Frequency	Valid percentage
Negative	3	4.2	23	32.9
Neutral	9	12.9	14	20
Positive	58	82.9	33	47.1
Total	70	100	70	100

Note. $N = 70$

While interviewing participants we specifically investigated the case of contrasting AttrakDiff and UX needs ratings with the single-word UX description given by each participant (positive UX rating associated with a negative single-word experience description or vice versa). Any time we felt that there was an inconsistency between users' ratings and their experience report, we asked the users to explain why in order to understand the rationale behind their UX evaluation.

We also asked users which elements influenced their UX positively or negatively. For Amazon, they mainly pointed to the content (31%, 138 citations), the usability (22%, 97 citations), and the service experience (16%, 71 citations). The elements influencing their UX while interacting with the camera were mainly the usability (37%, 128 citations), the design (26%, 91 citations), and the features (26%, 88 citations). The ratio between positive and negative elements is much more positive in the case of Amazon (64% positive vs. 36% negative) than in the case of the camera (52% positive vs. 48% negative), which follows the UX scores reported through the questionnaires.

Impact of the testing situation on participants' experiences

Reduced perceived autonomy. Participants reported that the testing situation influenced the need for autonomy, which was perceived as ambiguous. Even if one might feel autonomous when surfing on Amazon or when using a camera, the controlled testing situation places individuals in a context where freedom is inherently limited. During questionnaire completion, several participants reported (by thinking aloud) that their feeling of autonomy was reduced by the situation, even if they could imagine that they would feel autonomous with the systems. Participant 6 for instance stated, "I followed the scenario that you have designed so I did what you wanted me to do. I didn't feel autonomous in that context."

Impact of testing scenarios. Furthermore, several participants reported that, beyond the feeling of autonomy, their experience was influenced by the testing situation in other ways. Many of them mentioned that they performed actions through the testing scenarios that they would not have performed at home because they usually are not using these kinds of systems this way. Sometimes the scenarios would lead to positive experiences; this was, for instance, the case for a participant who discovered nice shoes on Amazon though she generally would only look for books or computer material. But more often in this experiment, this led to frustration and negative experiences; for instance, when users had to modify settings on the camera (e.g., picture size, filter effects) and reported that they would only have used the Auto mode at home and would probably have been very satisfied with it:

> "Without the scenario I would have said that it is easy to use, yet here with all things I would never have done at home, I feel that it is more complicated than expected" (Participant 13).

> "A negative experience is that the Magic mode was not easy to find. But this depends on the scenario, I wouldn't have felt frustrated if I wasn't observed to perform some tasks. I would on the contrary have felt curious and eager to explore and try things out" (Participant 15).

The same applied to Amazon: "This is the first time that I am using the menus, I usually use the search engine only. So here I could see how complex the website was, even though I never realized it before" (Participant 43). So, despite that we tried to keep scenarios easy and we instructed the participants that they could skip any of the scenarios if they wanted to, the laboratory setup indeed modified the felt experience.

Difficulties to assess some needs. Regarding the testing session, a majority of participants reported difficulties to assess some of the UX needs due to the testing situation. Relatedness or Influence needs items were highlighted as problematic because of the absence of people in the lab, especially people who are important to the user. For instance, a participant said, "this camera would probably contribute to the fulfillment of the need for relatedness, if I were at home or on holidays taking pictures of my wife and kids. But here alone in the lab, I truly don't feel that way, so I assessed it as not fulfilled at all" (Participant 10). The need for Self-Actualizing is another typical example of a feeling that is difficult to assess in the lab. Participant 6 stated, "well if I do nice and creative photos then I sometimes feel this way. But here, I know that the picture I took will be deleted and that's it. It will not have any impact."

Participants also highlighted that the assessment of a system depends on a more holistic set of criteria than just the direct interaction with the product. In the case of the camera, Participant 37 for instance stated, "my experience and judgement would depend on the price of the camera. I have no idea here how much it would cost, so I have trouble judging it." As per Amazon, several participants commented on the reputation of the company: "My experience is influenced here by what I already know about Amazon and how they treat their workers. So this is not about how I felt here during the interaction" (Participant 13).

Discrepancies between self-reported and observed. In some cases, we observed major differences in the evaluation made by participants using the questionnaires and the feelings reported during the debriefing interview. For instance, participants could rate their experience with the camera as quite negative because they had experienced several issues while performing the test, but then they could report the same experience as satisfying during the interview because they somehow felt that the device was interesting and could be enjoyable after a short learning period (Participant 2).

Temporality and anticipated experience. Finally, a typical constraint reported by participants is the reduced interaction time in the laboratory setting, which allows for a fragmentary experience only.

> "The camera looks powerful but I would need more time and taking pictures in different contexts (for instance daylight, portrait, sport) to truly assess it. I also couldn't assess whether battery life was OK here because in 15 minutes it is obvious that I would not have run out of battery" (Participant 10).

> "If you could lend me the camera for a week, I would have lived other experiences for sure. Here in 15 minutes I had no time to truly explore it" (Participant 5).

Discussion and Recommendations

The results of our study shed light on issues and challenges to be addressed when evaluating UX using laboratory user testing. In this discussion section, we show that some issues are not novel and were already recognized as problematic for the evaluation of usability (e.g., order effects or the impact of familiarity level on the assessment of a system). However, the extended scope of UX along with its subjective, situated, and temporal nature has brought additional challenges to tackle. Through our laboratory experiment, we were able to identify which aspects of the user testing situation were still suitable for the evaluation of UX and which aspects seemed to be challenged. In the following sections, we describe issues and recommendations related to each of these three UX characteristics.

Challenges with the Subjective Nature of UX

We introduced a psychological needs-driven approach in our study to cater for the subjectivity of UX. This approach is a well-explored area in UX research and appears to be a powerful framework for the design of more experiential interactive systems. UX designers should consider interactive systems as a means to fulfil needs ("be-goals") and not only a means to achieve task oriented "do-goals" (Hassenzahl, 2010). Do-goals have been much more prominent in a usability-driven approach while be-goals reflect the extended scope of UX. We therefore recommend the use of needs-driven approaches in support of testing of subjective aspects of the experience.

As stressed by Rogers et al. (2007), we saw that traditional usability metrics such as task completion time did not inform about the felt experience. While time spent on testing scenarios in both use cases was negatively correlated with the AttrakDiff pragmatic scale (which was expected because previous usability studies have shown links between efficiency and perceived usability), it had no influence on the perceived attractiveness, nor on the fulfillment of UX needs. Interestingly, it however affected the perception of hedonic qualities in the case of the camera: The more time a user spent on the scenarios, the more the camera was assessed as stimulating. This finding suggests that traditional usability or performance metrics such as efficiency do not necessarily reflect a bad experience and could even be clues of a positive experience. Designing for the experiences of curiosity, exploration, or interest (Yoon, Desmet, & van der Helm, 2012), are a few examples of cases where efficiency is not an ultimate goal to reach. In UX design, it is therefore essential not to misinterpret performance metrics and to adopt a holistic perspective on human-computer interactions. Similarly, participants' ratings on standardized UX questionnaires should be interpreted according to the importance that each user gives to the UX dimension under inquiry. In our experiment, adding the needs importance scale allowed us to understand the adequacy between users' expectations and the actual interaction.

Whenever one wants to assess a dimension of UX, one should understand how important or meaningful this specific aspect is to the user and how it will contribute to the overall subjective experience. This could be done by adding another self-reported metric as we did here or alternatively by using methods able to identify the elements of the interaction that are meaningful to the users. They will be able to subsequently report on their experience by using their own vocabulary and personal constructs. In our experiments, participants, for instance, commented on the "awkwardness" of some standardized items that would not be suitable for the context or not able to fully account for their experiences. The repertory grid (Möttus, Karapanos, Lamas, & Cockton, 2016; van Gennip, van der Hoven, & Markopoulos, 2016) or sentence completion methods (Kujala, Walsh, Nurkka, & Crisan, 2013), both arising from the field of psychology, could be alternative ways of assessing UX without constraining the user by a predefined vocabulary as the one typically used in standardized scales.

An additional concern stressed by our participants relates to the artificiality of testing scenarios that influenced the felt experience by directing users towards actions that they would probably not have done in a real-life context. First, users reported a direct negative impact of the testing situation on their assessment of the need for autonomy: As they were guided through the process by achieving standardized scenarios, they felt globally less autonomous and this

influenced their evaluation of the system's ability to make them feel autonomous. Moreover, the artificial actions triggered either positive or negative feelings and distorted users' experiences, thereby biasing the outcomes of UX evaluation. This unfortunately holds true even for the needs that seem easier to assess in a controlled experiment, such as Security or Competence. The scenario where users had to modify settings on the camera, for instance, negatively influenced some users' feeling of competence: At home, they would only have used the Auto mode and would probably have been very satisfied with it, not feeling frustrated or incompetent. These artificial behaviors and task selection biases (Cordes, 2001) were already identified as problematic within usability studies (Kjeldskov et al., 2004, NNGroup, 2014).

In user testing, the tasks defined by the evaluator stipulate, strongly or loosely, what objectives the users should reach (Hertzum, 2016). Introducing user-defined tasks could be a good way to cope with the perceived artificiality of testing scenarios. User-defined tasks are tasks that participants bring into the evaluations as opposed to product-supported tasks (Cordes, 2001). By asking users to define themselves and the tasks that they would need or like to perform with the system or the product, one comes much closer to a meaningful, motivating, and realistic experience. It also allows for identification of tasks that are not supported by the system and for an understanding of how a task meets users' expectations so a more authentic and accurate picture of the experience can emerge.

In our experiment, we could have asked users before the interaction to tell us what tasks they usually would have done with a digital camera or on Amazon, thereby turning users' goals into scenario tasks. We could have then used these user-defined tasks to conduct the user test. Note that this is different from the free exploration time our participants had at the beginning of the session, yet both could have been combined. Obviously, most tasks chosen by participants will be unique, which makes it harder to compare performance metrics and observations between participants. The choice of predefined scenarios versus user-defined tasks depends on the objectives of a study. The first option is mainly suitable for testing pragmatic aspects of an interaction while the second option could better account for the complexity of UX as a holistic concept involving hedonic, emotional, and contextual aspects.

Challenges with the Holistic and Situated Nature of UX

The testing situation and the laboratory setting affected the felt experience in many ways. First, being in a laboratory hindered the fulfillment of specific needs, such as Relatedness or Influence, which are so closely embedded into the social, physical, and daily context that they are not easily reproducible in a lab. This issue was frequently reported during the debriefing interviews and therefore it is hard to claim for the validity of our results regarding the aforementioned needs. The same could apply to the pleasure while interacting with Amazon; the pleasure in that case is mainly derived from buying something that one desires. In a laboratory setting, the tasks are somewhat standardized and even if the users were allowed to freely explore each system, the usage situation was not oriented towards the pleasure of the discovery or buying of appealing products. Similarly, the feeling of self-actualizing could arise from a wonderful photo shoot where one feels particularly creative and spontaneous; however, this is the kind of situation that we cannot capture in a laboratory because it is too much embedded in a real-life context. Our results are in adequacy with Sun and May's study (2013) that compared field-based and lab-based experiments to evaluate UX of mobile devices. The authors recommended using lab experiments when the "testing focus is on the user interface and application-oriented usability related issues" and field experiments "for investigating a wider range of factors affecting the overall acceptability of the designed mobile service" (p. 1).

To adapt lab experiments to the situated nature of UX, some authors have proposed adding contextual features to laboratory setups in order to improve the realism of the laboratory setting (Kjeldskov & Skov, 2003; Kjeldskov et al., 2004; Kjeldskov & Skov, 2007). Kjeldskov et al. (2004) recreated, for instance, a healthcare context in a laboratory to study the usability of a mobile application. In the case of UX however, recreating a meaningful setting in-situ seems challenging as UX is very often embedded in—and influenced by—daily routines and usual social interactions. Nevertheless, if a controlled experimental setting is required, trying to recreate the context of use seems relevant. One could think about adding specific furniture, triggering specific situations through role-playing (Simsarian, 2003) or involving families or friends to co-discover the system (Jordan, 2000; Pawson & Greenberg, 2009). Situating the system or product in a wider context could also be a good idea; in the case of the camera, we could have

presented the product along with a catalog description indicating the main features and the price. As there are several touchpoints influencing the experience during a user's journey, one can think about means to mimic these contextual elements in the lab. Augmented reality devices or simulators could also support a more contextual approach. However, these technological approaches are costly and not yet widely used in industry (Alves et al., 2014).

Last but not least, recent years have seen the emergence of remote user tests as an alternative to laboratory evaluation. They offer several advantages such as reduced cost and administration time (in the case of unmoderated tests) as well as the possibility to involve geographically distributed participants. During synchronous remote tests, the evaluator conducts the evaluation in real time with participants at a distance. Interestingly, studies have shown that this approach could be as effective as laboratory evaluations for the identification of usability issues (Lizano & Stage, 2014). Mainly advertised for their practicality, remote synchronous user tests could be an alternative way to evaluate UX in a more situated way, closer to field testing in some respects (e.g., tests conducted in the user's environment). One should nevertheless be aware that remote testing challenges the moderator role (Wozney et al., 2016).

A final challenge brought by the holistic and situated nature of UX is that of user (data) privacy. Assessing UX in a laboratory requires a thorough reflection on data collection and ethical issues. In the case of our study, we first considered a social network such as Facebook as a potential good candidate to be used in our users' evaluation sessions because of the diversity and intensity of experiences it triggers. However, we were challenged by privacy issues. While privacy issues were already relevant in the context of usability (and whenever we ask users to perform actions that are observed and recorded), additional challenges arise when dealing with UX. Assessing a realistic Facebook experience would have implied users logging on their own personal account (with their own friends and timeline), whereas at the usability level we probably could have tested the system using a fake account. Systems and products are increasingly providing a more personalized experience for their users; hence, privacy issues will become more frequent when assessing the UX of a product on the market in the presence of an observer (or a recording device). Of course, this doesn't apply to early prototypes or new products, but in those cases the challenge will be to simulate a personalized experience in order to assess their potential UX. Amongst our use cases, Amazon allowed us to assess a personalized service with less privacy issues, although participants still had to agree to log in using their passwords on our testing computer and to show the recommended products based on their previous purchases. The digital camera was less problematic from a privacy perspective because it did not belong to the users themselves and therefore did not contain private pictures. While privacy issues can be dealt with up to a certain degree, researchers and practitioners should carefully weigh the pros and cons of studies involving users' very private data, both for ethical reasons and because unveiling such data could trigger a feeling of awkwardness in some users. Several papers on ethical issues raised by UX have been published to raise awareness on the topic (Barcenilla & Tijus, 2012; Brown, Weilenmann, McMillan, & Lampinen, 2016; Munteanu et al., 2015).

Challenges with the Temporal Nature of UX

Another limitation of laboratory UX evaluation relates to the dynamics of UX, which is difficult to assess in a single session. We were already able to observe the impact of time on UX, especially by noticing a difference between the momentary evaluation made by users through the questionnaires and the more reflective evaluation they reported during the debriefing interview. Results also show that already known potential issues related to laboratory evaluations remain problematic in the conceptual approach of UX. Sequence biases were observed and influenced both the perceived experience assessed through the AttrakDiff scale and the reported fulfillment of UX needs.

Without adopting a novel method, could we adapt laboratory evaluations to improve the assessment of the temporal dimension of UX? As laboratory evaluations often entail a combination of evaluation methods, we could add specific tools to better understand UX over time at a micro-level during a testing session. First, it seems essential to investigate users' history by inquiring about their expectations, previous experiences and level of familiarity with the system (or similar ones), opinions about the system, or even anticipated UX. Then, one could use tools to assess the changes in UX during the session. Mood maps aim at documenting the emotional states of users over time by asking users to frequently report their emotional

state during the test. These maps could be used to better catch momentary frustrations and to match mood with specific parts of the interaction, thereby informing designers about what specifically should be improved. It is also possible to ask users to answer several questionnaires before, during, and after the interaction. Other tools, such as retrospective assessment curves (Karapanos, Martens, & Hassenzahl, 2012; Kujala, Roto, Väänänen-Vainio-Mattila, Karapanos, & Sinnelä, 2011), could be used to represent the evolutions of UX over time during the session. While UX curves were primarily designed to assess UX over long periods of time, UX curves, such as iScale (Karapanos et al., 2012), could also be used on a shorter timeframe. Finally, thinking aloud protocols along with observation and debriefing interviews have been shown in our study to be effective at detecting changes in the UX. If one wants to be more accurate in detecting changes in emotions or behaviors, one could use novel devices in the lab to provide psychophysiological measurements (Park, 2009), eye-tracking data, or facial expressions assessment (Zaman & Shrimpton-Smith, 2006). However, there is much more than that to account for UX temporality and this reflects the growing interest for long-term UX evaluation methods, such as longitudinal methods or retrospective UX assessments (Karapanos et al., 2012; Kujala et al., 2011).

With regards to another issue related to time, one should be aware that the duration of the session itself constrains the experience and resulting evaluation. For example, several participants mentioned that they did not have enough time to truly explore and appreciate the features of the camera, or to truly enjoy exploring products they like on Amazon. As shown in Karapanos et al.'s model of temporal aspects in UX (Karapanos, Zimmerman, Forlizzi, & Martens, 2010), the first period of use is characterized by an orientation phase where the user discovers the system. At this stage, strong UX-related factors such as functional dependency or emotional attachment are absent from the interaction. The evaluation of UX in a single, short user testing session therefore remains incomplete. For long-term UX to be assessed in a laboratory, one could think about multiple sessions involving the same participants; however, this approach is costly. The living lab method (Ley et al., 2015) also generates increasing interest and has the potential to address the situatedness and the temporal challenges brought by UX. In this approach, the environment can be completely appropriated by the users, while fulfilling both the requirements of, for example, a systematic observation (through appropriate observation and recording equipment) and those of a natural environment (through the location and specific equipment available). Users can have continuous and ongoing activities in this space (over weeks, months, or even years), thus leveraging their experiences beyond punctual snapshots.

In practice, longitudinal or retrospective methods are more suitable to address the challenges related to the dynamics of UX. A thorough assembly of methods is therefore required if the cumulative UX is to be assessed, that is, the combination of anticipated, momentary, and episodic experiences.

Beyond the Lab: Alternative Evaluation Methodologies

In the preceding section we have elaborated on the challenges for using laboratory evaluations in a conceptual approach of UX. In this section we discuss some alternatives that can be used to evaluate UX in a more naturalistic context by taking into account all UX related factors such as temporality.

Several researchers have argued in favor of more ecological evaluation methods of UX (Crabtree et al., 2013; Rogers, 2011), highlighting that only "in the wild" studies allow for understanding the complexity and richness of experiences (Shneiderman, 2008). Moreover, some authors also claim that field studies provide more valuable insights, thereby better serving design purposes. The main drawback of field studies, however, is the time and cost required to conduct them, typically more than twice the time of laboratory evaluations (Kjeldskov et al., 2004; Rogers et al., 2007). This issue is even more critical if one wants to use a field study as a longitudinal method. Real settings also challenge observation and data collection as one should try to observe and record interactions without interfering too much in the situation. Finally, field studies require working prototypes and are therefore not suitable for early UX evaluation.

A diary study seems to be a good candidate for a research methodology, apparently meeting all requirements to capture the experience from the user point of view by considering all

aforementioned factors. Following Allport (1942), who was encouraging the use of personal documents in psychological science, Bolger, Davis, and Rafaeli (2003) claimed that diary methods were able to "capture life as it is" by reporting events and experiences in their natural, spontaneous context. The advantages of diary methods for the study of UX are indeed numerous. First, diary methods allow studying and characterizing temporal and contextual dynamics of UX; this constitutes a real added value in comparison to more widespread methods like interviewing or think-aloud protocols. It also provides more accurate data on the observed phenomenon because the likelihood of retrospection is reduced (Bolger et al., 2003). Validity and reliability of the collected data is therefore expected to be higher than those related to a methodology implying retrospection of an event, such as UX curves for instance (Kujala et al., 2011). Diaries can help determine the antecedents, correlates, and consequences of daily experiences and therefore can help researchers to better understand the experiences in context. However, diary studies also have main disadvantages related to the cost and time associated with the recruitment of users, training, or briefing sessions and data analysis. They are bound to the expressive abilities of participants and could therefore not be used with any population type (Allport, 1942). As diaries involve self-reporting data only, they also have the drawback to be an indirect approach to data collection; they do not provide first-hand insight into the user experiences.

All in all, it seems hard to conciliate both the capture of the experiential and emotional flow during interaction, and the cumulative and reflective experience. This is a choice to be made according to the objectives of the study and expected outcomes. To address the limitations of single evaluation methods, it is of course possible to adopt a mixed-method approach by combining several methods (Ardito, Buono, Costabile, De Angeli, & Lanzilotti, 2008; Schmettow, Bach, & Scapin, 2014). No UX evaluation method is perfect in the sense of a one-size-fits-all solution and one needs to look at the pros and cons of each method before deciding how to evaluate UX. The trade-off between costs and benefits plays a major role in the choice and adoption of an evaluation method (Vredenburg et al., 2002). Consequently, if UX research wants to foster the adoption of more ecological or longitudinal approaches to UX evaluation, we should put more emphasis on their benefits in comparison to established methods, which are less demanding and costly.

Conclusion

By gaining better insights into how the conceptual approach of UX alters laboratory user testing, we showed that established user laboratory evaluation needs to be adapted in order to fit the nature and characteristics of UX. Furthermore, we should also be aware that practitioners adapt the methods to fit their needs and match specific project circumstances (Cockton, 2014; Woolrych, Hornbæk, Frøkjær, & Cockton, 2011). This is why it is important to investigate and communicate on the strengths and limitations of UX evaluation methods (also by considering methods as collections of resources, as suggested by Woolrych et al., 2011), thereby supporting UX experts in selecting the most suitable method and combination of resources according to the application domain, project constraints, or organizational factors.

While we should pursue to investigate further into the methods and metrics for UX evaluation, a better transfer from research to practice would also support the dissemination of novel evaluation methods, which were specifically designed for the assessment of UX. By raising awareness on the relevance of field studies (for evaluating UX in context) or in longitudinal studies (to evaluate the dynamics of experiences), we could provide UX practitioners with a larger palette of methods. In return, researchers could benefit from practitioners' feedback and data, leading to a win-win situation.

In conclusion, while the conceptual approach of UX definitely alters established HCI methods, we should see this as an opportunity to adapt and improve our research and evaluation practices. The question at hand is not whether we could still use established HCI methods for the evaluation of UX or not, but rather how to adapt existing UX evaluation methods, develop new ones and, over all, be able to wisely select the most suitable method depending on the objective of our study. This will ultimately support the design of better and more experiential systems and services.

Tips for Usability Practitioners

We present the following tips for practitioners conducting similar research:

- Be aware of the subjective, temporal, and situated nature of UX and evaluate each aspect individually with regards to the objectives and available methods in your project. This does not automatically rule out laboratory-based evaluations because they have the advantage of limiting the confounding variables, while varied contexts in the wild may introduce more confounds to the research findings.
- To meet the challenge of UX subjectivity in the lab, first make sure to explore how important or meaningful each UX dimension is to a user in the context of the interaction with your system or product. This will help you to correctly interpret data. A low score on a UX dimension considered of minor importance isn't necessarily an issue. You can also replace your predefined scenarios by user-defined tasks in order to evaluate the system from the user's perspective.
- To meet the challenge of UX situatedness in the lab, add contextual features to laboratory setups in order to improve the realism of the laboratory setting. This mix between in-vitro experiments and in-situ observations is called in-sitro. Alternatively, use evaluation methods showing a higher ecological validity such as remote moderated (synchronous) testing or field observation.
- To meet the challenge of UX temporality, always remain aware that UX is a cumulation of anticipated, momentary, and episodic experiences. While singular episodes and moments can be assessed specifically, the resulting cumulated UX requires combined approaches. The living lab approach, as well as the use of longitudinal (e.g., diaries) or retrospective assessment methods (e.g., UX curves) inform on the dynamics of UX.
- Understand the core of users' experiences beyond the assessment of perceived system qualities supported by standardized UX evaluation scales (e.g. Attrak scale, UEQ, meCUE). The psychological needs-driven approach used in our study is an example of this.
- Regard what used to be considered the usability "gold metrics" in a more nuanced way in UX design. Maximizing efficiency is for instance not always desired if one intentionally wants to design for curiosity, interest, or exploration. Just as for other dimensions, one should explore what is meaningful to the user in the context of this interaction.

Acknowledgments

We thank Mrs. Sophie Doublet who provided helpful comments on previous versions of this document and Dr. Salvador Rivas for his valuable contribution to the adapted UX needs questionnaire. This project was supported by the Fonds National de la Recherche, Luxembourg (n° 1205972).

References

Allport, G. W. (1942). *The use of personal documents in psychological science*. New York: Social Science Research Council.

Alves, R., Valente, P., & Nunes, N. J. (2014). The state of user experience evaluation practice. *Proceedings of the 8th Nordic Conference on Human-Computer Interaction* (pp. 93–102). New York, NY: ACM Press. doi:10.1145/2639189.2641208

Ardito, C., Buono, P., Costabile, M. F., De Angeli, A., & Lanzilotti, R. (2008). Combining quantitative and qualitative data for measuring user experience of an educational game. *Proceedings of Meaningful Measures: Valid Useful User Experience Measurement (VUUM), 5th COST294-MAUSE Open Workshop, 18th June 2008.* Reykjavik, Iceland.

Barcenilla, J., & Tijus, C. (2012). Ethical issues raised by the new orientations in ergonomics and living labs. Work, 41, pp. 5259–5265. doi: 10.3233/WOR-2012-0015-5259

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience. *Proceedings of the ACM Conference*

on Human Factors in Computing Systems, CHI 2011 (pp. 2689–2698). doi:10.1145/1978942.1979336

Benedek, J., & Miner, T. (2002). Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals' Association Conference UPA'02*, Orlando, FL.

Bevan, N. (2008). Classifying and selecting UX and usability measures. *Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM) June 18th 2008*, Reykjavik, Iceland.

Bødker, S. (2006). When second wave HCI meets third wave challenges. *Proceedings of the 4th Nordic Conference on Human-Computer Interaction* (pp. 1–8). New York, NY: ACM Press. doi:10.1145/1182475.1182476

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54*, 579–616.

Brown, B., Weilenmann, A., McMillan, D., & Lampinen, A. (2016). Five provocations for ethical HCI research. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press. doi: 10.1145/2858036.2858313

Cockton, G. (2014). Usability Evaluation. In M. Soegaard & D. Rikke Friis (Eds.). *The Encyclopedia of Human-Computer Interaction* (2nd ed.) Aarhus, Denmark: The Interaction Design Foundation. Available online at https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/usability-evaluation

Cordes, E. R. (2001). Task-selection bias: A case for user-defined tasks. *International Journal of Human-Computer Interaction, 13*(4), 411–419. doi:10.1207/S15327590IJHC1304_04

Crabtree, A., Chamberlain, A., Grinter, R., Jones, M., Rodden, T., & Rogers, Y. (2013). Article 13: Introduction. Special issue of "the turn to the wild." *ACM Transactions on Computer-Human Interaction (TOCHI), 20*(3). doi:10.1145/2491500.2491501

Desmet, P. M. A. (2003). Measuring emotion: Development and application of an instrument to measure emotional responses to products. In M. A. Blythe, A .F. Monk, K. Overbeeke, & P. C. Wright (Eds.), *Funology: From usability to enjoyment* (pp. 111–123). Dordrecht, Netherlands: Kluwer Academic Publishers.

Friedman, B., & Hendry, D. (2012). The envisioning cards: A toolkit for catalyzing humanistic and technical imaginations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI'12 ; pp. 1145–1148). New York, NY: ACM. doi:10.1145/2207676.2208562

Hassenzahl, M. (2010). Experience design: Technology for all the right reasons. *Synthesis Lectures on Human-Centered Informatics, 3*(1), 1–95. doi: 10.2200/S00261ED1V01Y201003HCI008

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttrakDiff: A questionnaire for measuring perceived hedonic and pragmatic quality]. In J. Ziegler, & G. Szwillus (Eds.), *Mensch & Computer 2003*: *Interaktion in Bewegung* (pp. 187–196) [Human & Computer 2003: Interaction in Motion]. Stuttgart, Germany: B. G. Teubner.

Hassenzahl, M., Diefenbach, S., & Göritz, A. (2010). Needs, affect, and interactive products: Facets of user experience. *Interacting with Computers*, 22(5), 353–362.

Hassenzahl, M., & Tractinsky, N. (2006). User experience: A research agenda. *Behavior & Information Technology*, 25, 91–97.

Hertzum, M. (2016). A usability test is not an interview. *Interactions,* 23(2), pp. 82–84. doi: 10.1145/2875462

International Organization for Standardization (ISO). (1998). ISO 9241-11:1998: Ergonomic requirements for office work with visual display terminals (VDTs)—Part 11: Guidance on usability. Geneva, Switzerland: International Organization for Standardization.

Jordan, P. W. (2000). *Designing pleasurable products: An introduction to the new human factors*. London, UK: Taylor & Francis.

Karapanos, E., Martens, J.-B., Hassenzahl, M. (2012). Reconstructing Experiences with iScale. *International Journal of Human-Computer Studies, 70*(11), 849–865. doi:10.1016/j.ijhcs.2012.06.004

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J.-B. (2010). Measuring the dynamics of remembered experience over time. *Interacting with Computers, 22*(5), 238–335. doi:10.1016/j.intcom.2010.04.003

Kjeldskov, J., & Skov, M. B. (2003). Creating a realistic laboratory setting: A comparative study of three think-aloud usability evaluations of a mobile system. *Proceedings of the 9th IFIP TC13 International Conference on Human Computer Interaction (Interact 2003*; pp. 663–670). IOS Press.

Kjeldskov, J., & Skov, M. B. (2007). Studying usability in sitro: Simulating real world phenomena in controlled environments. *International Journal of Human-Computer Interaction, 22,* 1–2.

Kjeldskov, J., Skov, M. B., Als, B. S., & Høegh, R. T. (2004). Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. *Proceedings MobileHCI 2004* (pp. 529–535), Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-540-28637-0_6

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

Kujala, S., Walsh, T., Nurkka, P., & Crisan, M. (2013). Sentence completion for understanding users and evaluating user experience. *Interacting with Computers*, *26*(3), 238–255. doi:10.1093/iwc/iwt036

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. doi:10.1016/j.chb.2014.10.048

Lallemand, C., Koenig, V., & Gronier, G. (2014). How relevant is an expert evaluation of user experience based on a psychological needs-driven approach? *Proceedings of the 8th Nordic Conference on Human-Computer Interaction* (pp. 11–20). New York, NY: ACM Press. doi:10.1145/2639189.2639214

Lallemand, C., Koenig, V., Gronier, G., & Martin, R. (2015). Création et validation d'une version française du questionnaire AttrakDiff pour l'évaluation de l'expérience utilisateur des systèmes interactifs [A French version of the AttrakDiff scale: Translation and validation study of a user experience assessment tool]. *Revue européenne de psychologie appliquée [European Review of Applied Psychology], 65*(5), 239–252.

Lavie, T., & Tractinsky, N. (2004). Assessing Dimensions of Perceived Visual Aesthetics of Web Sites. *International Journal of Human-Computer Studies, 60*(3), 269–298.

Law, E., Abrahão, S., Vermeeren, A., & Hvannberg, E. (2012). Interplay between user experience evaluation and system development: State of the art. Conference paper: I-UxSED 2012, Interplay between User Experience and Software Development*. Proceedings of the 2nd International Workshop on the Interplay between User Experience Evaluation and Software Development*, Copenhagen, Denmark.

Law, E., Bevan, N., Christou, G., Springett, M., & Lárusdóttir, M. (2008). *Proceedings of Meaningful Measures: Valid Useful User Experience Measurement (VUUM)*, Reykjavik, Iceland.

Law, E., Roto, V., Hassenzahl, M., Vermeeren, A. & Kort, J. (2009). Understanding, scoping and defining UX: A survey approach. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press.

Ley, B., Ogonowski, C., Mu, M., Hess, J., Race, N., Randall, D., Rouncefield, M., & Wulf, V. (2015). At home with users: A comparative view of living labs. *Interacting with Computers*, *27*(1), 21–35.

Lizano, F., & Stage, J. (2014). Remote synchronous usability testing as a strategy to integrate usability evaluations in the software development process: A field study. *International Journal on Advances in Life Sciences, 6(*3 & 4), 184–194.

Mahlke, S. (2008). *User experience of interaction with technical systems. theories, methods, empirical results, and their application to the design of interactive systems*. Saarbrücken, Germany: VDM Verlag.

Möttus, M., Karapanos, E., Lamas, D., & Cockton, G. (2016). Understanding aesthetics of interaction: A repertory grid study. *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. New York, NY: ACM Press

Munteanu, C., Molyneaux, H., Moncur, W., Romero, M., O'Donnell, S., & Vines, J. (2015). Situational ethics: Re-thinking approaches to formal ethics requirements for human-computer interaction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press. doi: 10.1145/2702123.2702481

NNGroup (2014). Turn user goals into task scenarios for usability testing (published Jan 12, 2004). Retrieved from http://www.nngroup.com/articles/task-scenarios-usability-testing/

Odom, W., & Lim, K.-Y. (2008). A practical framework for supporting "third-wave" interaction design. *Proceedings of alt.chi 2008, Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press.

Park, B. (2009). Psychophysiology as a tool for HCI research: Promises and pitfalls. In J. A. Jacko (Ed.), *Lecture Notes in Computer Science: Human-Computer Interaction* (Vol. 5610*). New Trends* (pp. 141–148). doi: 10.1007/978-3-642-02574-7_16

Pawson, M., & Greenberg, S. (2009). Extremely rapid usability testing. *Journal of Usability Studies*, *4*(3), 124–135.

Rogers, Y. (2011). Interaction design gone wild: Striving for wild theory. *Interactions, 18*(4), 58–62.

Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W., Kurtz, A., Hall, R. E., Hursey, J., & Toscos, T. (2007). Why it's worth the hassle: The value of in-situ studies when designing Ubicomp. *Proceedings of the 9th international conference on Ubiquitous computing (UbiComp '07*; pp. 336–353*)*. Berlin, Heidelberg: Springer-Verlag.

Roto, V., Law, E., Vermeeren, A., & Hoonhout, J. (2011). User experience white paper: Bringing clarity to the concept of user experience. *Result from Dagstuhl Seminar on Demarcating User Experience, Sept. 15- 18, 2010.* Finland.

Roto, V., Obrist, M., Väänänen-Vainio-Mattila, K. (2009). User experience evaluation methods in academic and industrial contexts. *Proceedings User Experience Evaluation Methods in Product Development (UXEM'09).* Workshop in Interact'09, Sweden.

Schmettow, M., Bach, C., & Scapin, D. (2014). Optimizing usability studies by complementary evaluation methods. *Proceedings of the 28th British HCI Conference (BSC-HCI 2014)*. Southport, UK.

Shneiderman, B. (2008). Science 2.0. *Science, 319*, 1349–1350.

Sheldon, K. M., Elliot, A. J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology, 89*, 325–339. doi: 10.1037//O022-3514.80.2.325

Simsarian, K. T. (2003). Take it to the next stage: The roles of role playing in the design process. *Extended Abstracts on Human Factors in Computing Systems (CHI EA'03)*. New York, NY, USA: ACM Press. doi:10.1145/765891.766123

Sun, X., & May, A. (2013). Comparison of field-based and lab-based experiments to evaluate user experience of personalised mobile devices. *Advances in Human-Computer Interaction*, Article ID : 619767. doi:10.1155/2013/619767

Väänänen-Vainio-Mattila, K., Roto, V., & Hassenzahl, M. (2008). Towards practical user experience evaluation methods. *Proceedings of Meaningful Measures: Valid Useful User Experience Measurement (VUUM), 5th COST294-MAUSE Open Workshop, 18th June 2008* Reykjavik, Iceland.

van Gennip, D., van der Hoven, E., & Markopoulos, P. (2016). The phenomenology of remembered experience: A repertoire for design. *Proceedings of ECCE 2016*. New York, NY, USA: ACM Press.

Vermeeren, A, Law, E., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: Current state and development needs. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. New York, NY: ACM Press.

Vredenburg, K., Mao, J.-Y., Smith, P. W., & Carey, T. (2002). A survey of user-centered design practice. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'02*; pp. 471–478). New York, NY, USA: ACM Press.

Woolrych, A., Hornbæk, K., Frøkjær, E., & Cockton, G. (2011). Ingredients and meals rather than recipes: A proposal for research that does not treat usability evaluation methods as indivisible wholes. *International Journal of Human-Computer Interaction, 27*(10), 940–970.

Wozney L. M., Baxter P., Fast H., Cleghorn L., Hundert A. S., & Newton, A. S. (2016). Sociotechnical human factors involved in remote online usability testing of two ehealth interventions. *JMIR Hum Factors, 3*(1). doi: 10.2196/humanfactors.4602

Yoon, J., Desmet, P. M. A., & van der Helm, A. (2012). Design for interest: Exploratory study on a distinct positive emotion in human-product interaction. *International Journal of Design*, *6*(2), 67–80.

Zaman, B., & Shrimpton-Smith, T. (2006). The FaceReader: Measuring instant fun of use. *Proceedings of the 4th Nordic Conference on Human-Computer Interaction*. New York, NY: ACM Press.

About the Authors

Carine Lallemand, PhD
Dr. Lallemand is a postdoctoral researcher at the University of Luxembourg and Vice-President of the French UXPA chapter FLUPA. Her research work is focused on the development, adaptation, and validation of UX design and evaluation methods. She is the first author of the handbook "Méthodes de design UX" (Eyrolles, 2015, 2nd edition 2017).

Vincent Koenig, PhD
Dr. Koenig is a senior research scientist at the University of Luxembourg, heading the HCI research group and usability lab, part of the Institute of Cognitive Science and Assessment. His work covers: User-centered design, usability, user experience, computer-based assessment, automotive HCI, gamification, and usable and socio-technical security.

Appendix A: UX Needs Fulfillment Questionnaire

Instructions: The following questionnaire contains descriptions of complex feelings that we ask you to rate relative to your interaction with (insert name of the system / product / service). All the sentences start with "During this interaction, I felt…" Please rate each sentence on a scale ranging from 1 *Not at all* to 5 *Extremely*.

During this interaction, I felt…	
AUT1	… that my actions were based on my interests.
AUT2	… free to do things my own way.
AUT3	… free from any pressure or influence.
AUT4	… free of having to make meaningful choices.
CP1	… that I was successfully completing tasks.
CP2	… I mastered complex situations.
CP3	… very capable in what I did.
CP4	… I could achieve my goals.
CP5	… I performed well.
REL1	… a sense of contact with other people in general.
REL2	… close and connected with people who are important to me.
REL3	… cared for.
REL4	… aware of others' emotions, activities, or mood.
PL1	… that I was experiencing new activities.
PL2	… that I experienced enjoyable sensations.
PL3	… physical or emotional pleasure.
PL4	… that I discovered new sources and types of stimulation.
SEC1	… that things were structured and predictable.
SEC2	… that I could frequently apply my routines and habits.
SEC3	… that I could act in a safe and secure way.
SEC4	… I understood how things worked.
SEC5	… in control.
INF1	… that I was a person whose opinion counts for others.
INF2	… that I influenced others.
INF3	… as someone that others take as a person who can give guidance.
INF4	… I am a likeable person.
SELF1	… my actions were with purpose.
SELF2	… my actions conformed to my values.
SELF3	… a sense of fulfillment. .
SELF4	… being a person of value

N.B. The seven subscales should be presented in a random order. Participants respond on a 5-point Likert scale ranging from *Not at all* to *Extremely*.

Appendix B: Individual UX Needs Fulfillment Importance Rating

Instructions: Please rate the importance of the following feelings relative to your use of (insert name of the system / product / service). Do not consider how important these feelings are in your daily life, but please focus specifically on how important they are when you interact with (insert name of the system / product / service).

Each of the following sentences includes two ideas that may seem distinct and cause you to hesitate. If this is the case, base your judgment on the feeling that suits you the most.

When using _____ , how important is it for me …	
IMP_AUTONOMY	to be the cause of my own actions or not to be influenced.
IMP_RELATEDNESS	to have contacts with people who are important to me or to feel part of a community.
IMP_COMPETENCE	to be capable or effective in my actions.
IMP_SECURITY	to be safe or in control of the situation.
IMP_PLEASURE	to enjoy myself or to feel stimulated by new things.
IMP_INFLUENCE	to be liked, respected, or have an influence over others.
IMP_SELFACTU	to develop my best potential or to make my life meaningful.

N.B. The seven items should be presented in a random order. Participants respond on a 5-point Likert scale ranging from *Not important at all* to *Very important.*