A Systematic Mapping Study on Research Contributions on UX Evaluation Technologies

Luis Rivero and Tayana Conte
Usability and Software Engineering Group - Universidade Federal do Amazonas
Manaus, Brazil
{luisrivero, tayana}@icomp.ufam.edu.br

ABSTRACT
User eXperience (UX) refers to quality attributes related to the feelings and emotions of the users. In order to evaluate UX, several technologies (tools, methods, techniques, others) have been proposed that range from using questionnaires to employing biometrics. However, it is essential to characterize UX evaluation technologies to obtain evidence on the contexts in which using a specific technology can provide better evaluation results. In this paper, we present the state of the art on how to evaluate the UX of software applications. We performed a systematic mapping study with an initial sample of 2101 papers from which 227 relevant papers describing UX evaluation technologies have been identified. The results suggest that there is a need for specific UX evaluation technologies that are easy and comfortable to use from the point of view of users, while supporting practitioners in the correction of the aspects that cause poor experiences.

Author Keywords
User eXperience; Software Quality; Evaluation Technologies; Systematic Mapping Study

ACM Classification Keywords
H A General Literature, D.2.5 Design, D.2.10 Testing and Debugging, H.5 Information Interfaces and Presentation, H.5.2 User Interfaces.

INTRODUCTION

Usability is one of the main attributes that represent quality in use considering the aspects of how easy to learn and use a software is [2]. Despite the increasing attention that usability has achieved [21], a new term, "User eXperience" (UX), has emerged as an umbrella phrase for new ways of understanding and studying the quality in use of interactive products [3].

The ISO standard 9241-110 [11] defines UX as the users' perceptions and responses with regards to their interaction with a system or product. According to Hassenzahl [9], user experience goes beyond task execution in an application and focuses on hedonic aspects of use such as fun and pleasure. Consequently, UX research focuses on new approaches to design interactive products accommodating experiential qualities [10].

New UX evaluation technologies (i.e. methods, techniques, tools, processes, artifacts, others) have been proposed to assist practitioners in assessing if the final application meets quality standards in terms of UX [22]. Considering the importance of evaluating UX to improve the quality of the software under development, a number of literature reviews on UX evaluation technologies have been carried out to verify which proposals exist and in which contexts they are useful [1,3,4,7,17,19,22]. Their findings show the potential of UX evaluation technologies in specific contexts (such as the evaluation of recommender systems, games and others), what aspects are being evaluated and how specific technologies have been adapted for the evaluation of UX attributes in those contexts. Nevertheless, few of the cited reviews focus on providing a broad picture of the state of the art of UX evaluation methods. Although Vermeeren et al. [22] carried out a review where 96 UX evaluation methods were identified; their review considers technologies that were proposed up until the year 2009. Therefore, there is still a need to verify which technologies have been proposed since then.

A Systematic Mapping (SM) study is a method for categorizing and summarizing the existing information about a research question in an unbiased manner [18]. This kind of studies help to identify gaps in current researches in order to suggest areas for further investigation [6]. Consequently, to elicit the state of the art of how to evaluate the UX of software applications in different contexts, we conducted a SM. The main contribution of this paper is the analysis and summary of how the UX evaluation technologies have been proposed and applied from the year 2010 until the year 2015.

BACKGROUND AND RELATED WORK

User eXperience

Several definitions have been given to the term User eXperience. For instance, Isomursu et al. [12] state that UX is the totality of the subjective experience of using an application in a situation. Authors who agree with this

 IHC'17, Proceedings of the 16th Brazilian symposium on human factors in computing systems. October 23-27, 2017, Joinville, SC, Brazil. Copyright 2017 SBC. ISBN 978-85-7669-405-2 (online).

definition state that as UX is formed in a dynamic relationship between the user and the device, the application and the usage environment, it cannot be evaluated in a vacuum. Additionally, UX also brings out aspects such as beauty, fun, pleasure, and personal growth that satisfy general human needs but have little instrumental value [16]. Similar definitions are provided by other authors [3,9]. In this sense, the key focus of UX is considering attributes beyond the instrumental such as users' emotions before, during and after their interaction with a software application.

Among the researchers who studied how UX affects product usage, Gaver and Martin [8] stressed the importance of a range of non-instrumental user needs, such as surprise, diversion, or intimacy, to be addressed by technology. Additionally, Jordan [13] suggests that different aspects of pleasure are important to enhance the user's interaction with it.

The above studies suggest that in order to evaluate UX, new methods that consider attributes such as emotions should be proposed for the evaluation of software applications. In the following subsection, we present how UX can be evaluated and the results from some reviews that have been performed to identify UX evaluation technologies.

UX Evaluation

Evaluating the UX of an application is a challenging task, as users may find it difficult to express their experiences if directly asked to [16]. Hence, besides objective measures such as task execution time or the number of errors, it is necessary to develop evaluation approaches that determine what is relevant from the user point of view [22].

To assist the observation of UX, Desmet [5] suggests the use of two types of instruments: non-verbal (objective) and verbal (subjective). Non-verbal instruments comprise mechanisms that allow observing the expressive or the physiological component of an emotional response. On the other hand, verbal self-report instruments typically assess the subjective feeling component of emotions. Subjective feelings can only be observed through self-report.

To investigate the different available UX evaluation approaches, a number of reviews has been performed. For instance, Vermeeren et al. [22] carried out a review where 96 UX evaluation methods were identified. They characterized the methods according to their origin, type of collected data, evaluated type of applications and others, and found out that there is a need for methods for the early phases of development and that special attention should be given to proposing UX methods that are practical to use. Bargas-Avila and Hornbæk [3], on the other hand, focused on identifying papers where empirical studies where performed in the field of UX. The authors identified a total of 51 studies, indicating the type of employed methods within the studies and the UX dimensions that were assessed. Although their review was not specifically developed for characterizing UX evaluation methods, the authors found 15 papers describing UX evaluation methodologies. Another generic review was performed by Law [17]. The author explored UX metrics and types of methods that could provide those metrics in a review paper. Such review can provide a basis for understanding the UX and future needs in the field of UX evaluation. Bargas-Avila and Hornbæk [4] also provided a ranking of UX evaluation technologies and the aspects that these methods evaluated through the analysis of the data from their previously published review [3]. Moreover, Rajeshkumar et al. [19] carried out an analysis of UX evaluation methods identified in other reviews (considering the review by Vermeeren et al. [22]) providing an overview of how UX methods have been applied and a categorization to assist practitioners in their application.

Other reviews focused on a specific category of UX evaluation methods. For instance, the reviews by Frey et al. [7] and Balters and Steinert [1] focused on the evaluation of UX by means of physiology measurement. According to Frey et al. [7], physiological sensors and neuroimaging allow exploring concepts such as workload, attention, vigilance, fatigue, error recognition, emotions, engagement, flow and immersion, which could assist in the evaluation of UX. Furthermore, Balters and Steinert [1] provide an overview of current studies using physiology sensors in engineering and human–computer interaction settings.

Although there are several reviews that provide indicators in the field of UX evaluation, there is still a need for further reviews that provide an updated overview of the current state of UX research, since most of the recently published reviews were focused on specific categories of UX evaluation methods [1,7]. Also, there is a need that such reviews are systematically executed so their results are more reliable. Thus, it is necessary to dig deeper, finding further information on these methods and the situations in which they prove useful for assessing the UX and identifying problems to be corrected. In the next subsections, we describe how we analyzed UX evaluation methods through a systematic mapping study.

RESEARCH METHOD

A Systematic Mapping (SM) study uses systematic and explicit methods to identify, evaluate and interpret all relevant studies for clearly defined research questions. We have performed our systematic mapping study by considering the guidelines provided by Kitchenham and Chartes [14], dividing our SM in three stages: Planning, Conducting, and Reporting of the results.

Planning of the Systematic Mapping Study

Research Questions

The goal of our study is to examine the current use of UX evaluation technologies in the development of applications from the point of view of the following research question:

"*What technologies (methods/techniques/tools/others) have been proposed for the evaluation of user experience in the development of applications and how have these methods been used?*"

Our research question is broad in order to cover what researchers are evaluating when referring to the term UX. When considering previous methods that are related to UX but do not use such term (e.g. Kansei Engineering, Workload Scale, others) as they were proposed before the conception of the UX term, readers can refer to other reviews [13,22]. To address our research question, it has been decomposed into more detailed sub-questions, which allow characterizing the retrieved UX evaluation technologies. Table 1 shows these sub-questions along with their motivation.

Search Strategy

As suggested by Kitchenham and Charters [14], we used the PICOC (Population, Intervention, Comparison, Outcome, and Context) criteria to improve and structure our search for papers. We refined the PICOC criteria to frame the research question defined above, as follows:

- Population (P): The population is software applications. However, as describing all types of applications (e.g. Web, Mobile, Desktop, others) in the search string can be very exhausting and there is a risk of forgetting synonyms for any of the categories of software applications, the population will be defined in the search string as a restriction of research area.
- Intervention (I): Technologies that can be applied in the software development process.
- Comparison (C): Not applicable, since the goal is not to make a comparison between technologies, but to characterize them.
- Outcome (O): User eXperience Evaluation.
- Context (C): Any technology evaluating UX can be retrieved as there is no comparison.

In order to develop an appropriate search string, we have considered the reviews by Bargas-Avila and Hornbæk [3] and Vermeeren et al. [22] as basis. We followed the suggestions by Bargas-Avila and Hornbæk [3] regarding the term for representing User eXperience. The authors searched for papers employing the term "user experience" as they pointed out that it better represented the new movement. However, as user experience could be shortened to UX, we also employed that term. Additionally, we reviewed the papers identified in the review by Vermeeren et al. [22] to find out keywords representing technologies for UX evaluation. To carry out such analysis, we considered the papers within that review that refer to user experience evaluation technologies that can be employed in the evaluation of software applications, even though the authors found UX evaluation methods from other areas (e.g. Design and Psychology) that were applicable to products in general. This decision was made, as our review tried to retrieve papers that described technologies applicable for the development of applications.

ID	Research Sub-Question	Goal
SQ1	Type of Technology (What type of technology do the authors propose?)	To identify the type of technology described in the paper. In other words, we aim at identifying how the technology is collecting the UX data.
SQ2	UX Data Source (Who provides the UX information that is collected by using the technology?)	To identify the people who provide insights/opinions towards the evaluated application.
SQ3	Location (Where can the method be applied?)	To identify in which location the technology can be applied.
SQ4	Type of Assessed Application (What type of application is assessed?)	To identify which types of software applications that could be evaluated with the UX technology.
SQ5	Type of Assessed Artifact (What type of artifact is assessed?)	To identify what type of software development artifacts the technology could evaluate.
SQ6	Assessed Period of Experience (What period of experience is studied?)	To identify the time in which the experience is evaluated.
SQ7	Collected Data (What type of data is collected?)	To identify the type of data that the UX technology gathers.
SQ8	Supports Correction of the Identified Problems (Does the technology provide means to facilitate/support the correction of the encountered problems?)	To identify if the technology also included steps for dealing with the identified UX problems
SQ9	Availability (Are all materials necessary for the application of the method available?)	To identify if the technology was available for use

Table 1. Research sub-questions and their motivation.

Considering the specifications described above, our search string was developed following the restrictions below for each of the PICOC criteria [14]:

- Population: LIMIT-TO (SUBJAREA,"COMP") AND
- Intervention (Terms found in the papers identified by Vermeeren et al. [22]): (Method OR Instrument OR Tool OR Questionnaire OR Approach OR Technique OR System OR Scale OR Scheme OR Framework OR Model) AND
- Outcome: (User Experience OR UX) AND (Terms found in the papers identified by Vermeeren et al. [22]) (Assessment OR Measurement OR Evaluation OR Testing OR Recognition OR Measure OR Evaluating OR Tracking, Assess) AND
- Time span: (PUBYEAR > 2009) since Vermereen et al. (2010) already covered until the year 2009.

The final search string was applied in Scopus to identify research papers. Scopus is search engine that provides high flexibility in the execution of search operations and is one of the largest databases indexing abstracts and citations

[14]. We applied our search string in September, 2015. Therefore, some papers that were published in the year 2015 have also been considered in our systematic mapping.

Selection and Data Extraction Strategy

To select which papers would be considered in our review, we created a set of criteria that would objectively define whether a paper contained relevant information on UX evaluation technologies and was available for analysis. These inclusion (see code IX, where X is the number of the inclusion criterion) and exclusion (see code EX, where X is the number of the exclusion criterion) criteria were defined as shown in Table 2. Also, to standardize our extraction process, we extracted information from each of the selected papers according to our research sub-questions in Table 1 by adopting a set of possible answers to categorize the extracted technologies.

Conducting of the Systematic Mapping Study

Our review identified 2101 papers published between 2010 (January) and 2015 (September). Initially, we carried out the first filter of our review, i.e. we read the title and abstract of each paper to verify if they met the inclusion criteria. The papers that were accepted in the first filter were then downloaded and, after reading the entire paper, we decided whether they met the inclusion criteria or were not in the scope of our review (i.e. met the exclusion criteria). When a potential paper was not accessible due to a subscription requirement, the corresponding author was contacted to gain access to the paper.

Table 3 shows the number of returned, accepted, and rejected papers from the year 2010 until the year 2015. Our review selected a total of 227 papers that were in accordance with the inclusion criteria. Also, from the initial set of papers, 9 presented secondary studies (literature reviews or systematic mappings studies or systematic literature reviews). These papers were discussed in our Background and Related Work Section.

ID	Inclusion Criteria
I1	Paper describing technologies for evaluating user experience in software applications.
ID	**Exclusion Criteria**
E1	Papers that do not describe a technology for evaluating user experience or do not evaluate UX in software applications.
E2	Papers in which user experience is only mentioned to advertise a specific technology (e.g. the paper describes that a specific technology has a high quality user experience).
E3	Papers that can be considered Grey literature (e.g. a course summary, thesis and dissertations), assuming that good quality grey literature research will appear as journal or conference papers [15].
E4	Duplicated papers, i.e. papers that were already retrieved by the search engine.
E5	Papers in languages different than English.
E6	Papers not available for download after trying to contact the authors.

Table 2. Inclusion and exclusion criteria for selecting papers on UX technologies.

Year	Total Returned	Accepted (1st Filter)	Accepted (2nd Filter) and Extracted
2010	262	64	32
2011	346	62	29
2012	364	84	45
2013	459	107	47
2014	509	125	53
2015	161	45	21
Overall	**2101**	**487**	**227**

Table 3. Results of the conducting stage.

RESULTS

The overall results, which are based on counting the primary studies that are classified in each of the answers to our research sub-questions, are presented in Table 4. Note that the answers to sub-questions SQ1 to SQ5 are not exclusive; which means that a paper can be classified into one or more of the possible answers (i.e. the sum of the percentages can be over 100%). Readers who are interested in viewing the references of the selected papers within this review and their classification can find them in [20].

Publication Year and Venue

The 227 identified papers were published between 2010 and September 2015. From a time perspective (see Figure 1), the number of publications has been increasing in the past few years. As this SM was conducted in September 2015, not all conferences held in 2015 had their publications in indexed searchable digital libraries. This may be the reason for the low number of papers in that year.

In this SM, we considered peer-reviewed venues (including journals, conferences and workshops). Figure 2 provides an overview of the distribution of papers per scientific journal. The top three journals are the Interacting with Computers Journal, the Journal of Universal Computer Science and the International Journal of Human Computer Studies.

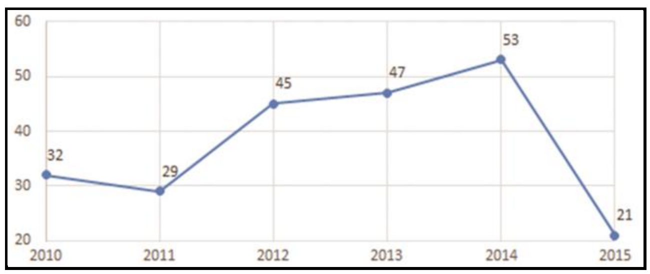

Figure 1. Distribution of papers per year.

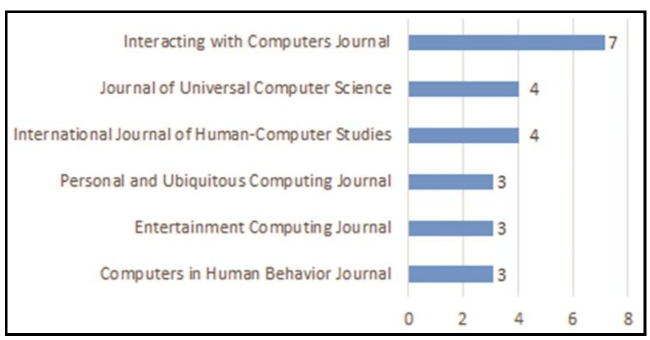

Figure 2. Distribution of papers per journal.

		Number of Papers	Percentage of Papers
SQ1 - Type of Technology	Written Reporting	153	67.4
	Oral Reporting	50	22.0
	Observation/Monitoring	104	45.8
SQ2 - Information Source	Users	219	96.5
	The Development Team	5	2.2
	UX Experts	12	5.3
SQ3 - Location	Controlled environment	188	82.8
	Field	45	19.8
SQ4 - Type of Assessed Application	Generic	77	33.9
	Web Application	31	13.7
	Mobile Application	20	8.8
	Others	100	44.1
SQ5 - Assessed Period of Experience	Before Usage	18	7.9
	During Usage – Single Ep.	100	44.1
	During Usage – Long Term	15	6.6
	After Usage	163	71.8
SQ6 - Collected Data	Qualitative	31	13.7
	Quantitative	133	58.6
	Both	63	27.8
SQ7 - Supports Correction of Identified Problems	Yes	15	6.6
	No	212	93.4
SQ8 - Availability	Available For Free	131	57.7
	Available Under a License	22	9.7
	Not Available	86	37.9

Table 4. Results per research sub-question from the systematic mapping study.

With regards to the distribution of papers per conference and workshop, Figure 3 shows that the top venues were: the Conference on Human Factors in Computing Systems (CHI), the Nordic Conference on Human-Computer Interaction (NordiCHI), the International Conference on Interaction Design and Children (IDC) and the International Conference of Design, User Experience and Usability (DUXU). There were other conferences, journals and workshops with only two or one published paper, but we did not represent them in the figures. The complete reference of each paper with their respective conferences, journal or workshops can be found in [20].

Findings per Research Sub-Question

The following subsections present the analysis of the results for each of the research sub-question according to Table 4. This analysis may be an interesting topic for researchers and practitioners willing to gain knowledge on the current available UX evaluation technologies and their advantages and disadvantages according to their categorization. Some papers from our review have been cited as [SXX], where XX is the number of the citation. Due to lack of space, to view the complete list methods and their categorization, interested readers can refer to [20].

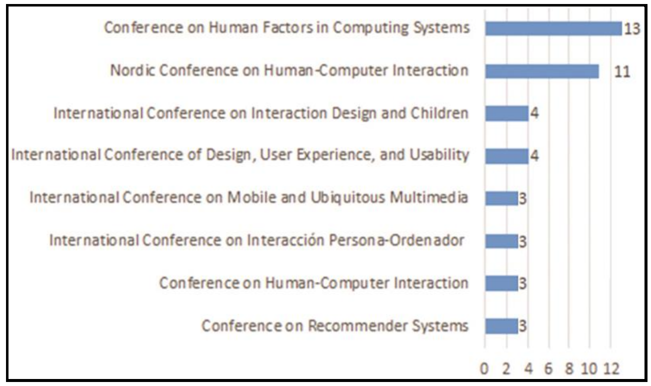

Figure 3. Distribution of papers per conference and workshops.

Results for SQ1 - Type of Technology

We categorized the type of the identified methods within the papers according to the way in which UX data was collected. As a result, 78.2% of the papers reported technologies in which the UX data was provided by writing or speaking or both. Additionally, 45.8% of the papers presented UX evaluation technologies that monitored the user's responses or behavior to gather UX data. Below, we present the technologies we identified in these categories.

Technologies such as scales, forms and checklists allow users to report their experience without interacting with an evaluator. Scales and forms are quick and easy to use. They allow gathering data from distributed users covering different user profiles, and can enable gathering data both on positive and negative experiences [S3]. However, most of these scales still try to quantitatively evaluate the emotions felt by users and may fail to provide information on the causes for poor experiences. On the other hand, UX researchers are developing specific questionnaires in which users can report aspects of their day or interaction that affected their experience [S2]. Other written/online forms employ specific questions to make users recall information. Although it may be difficult for some users to remember specific events, a retrospection method can allow users to report on the main events that affected their experience. Finally, some checklists have also been developed to guide evaluators during the reporting of their experiences [S7].

In Interviews, evaluators ask predefined questions that aim at extracting UX data and may provide means to understand the reasons why a specific feature of an application is impacting the UX [S11]. However, the presence of the moderator may have a negative impact on the evaluation session, making users feel uncomfortable by asking them about their experience in person. To mitigate this problem, Exploration with Acquaintances has emerged as an alternative for letting users discuss the positive and negative aspects of their experience, however without the intervention of an evaluator. When using this type of method, users can talk to people they know, making them feel more comfortable [S5]. The main advantage of this approach is that users tend to comment based on the opinions of others and this method could encourage shy users to open to their friends and report on their experience. This type of method can also be useful for monitoring users' reactions towards an application when socializing. Nevertheless, such an informal evaluation environment can cause users to focus on aspects that are not relevant for the improvement of the application. Finally, Probes are another alternative for motivating and inspiring users in reporting their experience orally [S9]. One can use materials such as multimedia and objects to engage users in the evaluation process of an application. The main issue with this type of method is the cost of producing the materials necessary for performing the evaluation.

Another method that monitors the physiological response or behavior when interacting with an application is experience sampling. This method allows users to report their experience at specific moments of their day [S8]. While a device periodically senses their physiological responses to stimulus, users are asked how they are feeling, and the context they are in. The main advantage of this method is having users experiencing the applications in real usage scenarios, gathering valuable information to improve UX.

Controlled User Monitoring is another type of method which allows gathering information on the users' psychophysiological responses or their reaction towards a stimulus [S1]. This type of technology, however, is applied in controlled environments (due to the attached sensors), which can have an effect over how the user behaves. In this sense, we identified that facial expressions, heart rate, respiration rate (and others) still remain the main measures for identifying users' emotions when interacting with an application. Nevertheless, the use of eye tracking, hand movement and body language is now being employed to measure users' mental workload, engagement and others.

Results for SQ2 - Information Source

This sub-question sheds light into who provides the data that will be analyzed to identify UX problems and improvement opportunities in the application. Regarding the use source of UX data, users are the main source of information (96.5%), followed by UX experts (5.3%) and the development team (2.2%).

The main advantage of having users express their emotions is that they can reveal information that only someone for whom the application is being developed can provide [S2]. As UX is subjective, users can describe specific attributes of the application that other stakeholders cannot. On the other hand, evaluations with users can be affected by how the data is collected. For instance, users who are shy may have problems expressing their emotions. Therefore, objective evaluation technologies that gather information on users' emotional response, or draw conclusions based on users' behavior can mitigate this problem. Other methods that make users feel more comfortable during their application are questionnaires or exploration with acquaintances. By applying questionnaires containing scales to several participants, the evaluators can identify patterns in the response to the use of the system/stimulus while allowing the subjects to answer on their own, without the intervention of an evaluator [S3]. Also, methods in which users participate together in the evaluation can allow them to build their comments on the comments of others (i.e. the opinion of one user may trigger another comment from another user, facilitating the reporting of their UX).

Some methods try to make software developers and UX experts create empathy with the target users. However, the main disadvantage of allowing UX experts or the development team to provide UX feedback is that they may be biased as they may not provide accurate results [S7].

Results for SQ3 – Location

Regarding the location where the technologies can be applied, most of the selected papers (82.8%) described technologies that need to be applied under controlled environments (i.e. labs or simulated usage scenarios). Methods in which users are being observed while interacting with the application in controlled environments can allow evaluators to gather specific data they are interested in. For instance, users can report their experience during and after an interaction in a specific location, or they can be monitored to gather objective data [S10]. This type of evaluations allows verifying what the users are feeling while being observed. Also, while asking questions during or after experiencing the application according to their observations, the evaluators can make sense of their observations and better understand the reasons for a poor or positive UX. Nevertheless, as users are being observed (even in specific usage scenarios) they might feel uncomfortable, affecting the results of the evaluation.

Only 19.8% of the papers describe technologies that can be applied in real usage situation of the software under evaluation. The main advantage of allowing users to evaluate UX in real uncontrolled usage situation, is that the data is more representative of a real use scenario. However, methods that allow gathering such feedback are scarce, and those that are available might not allow evaluators to intervene [S8]. Some experience sampling methods may allow evaluators to prompt specific questions during real usage interaction while monitoring users' physiological responses. Nevertheless, the prompted questions may disrupt the interaction, or be inapplicable in long trials.

Other methods that allow evaluators to gather data in real usage situations, allow users to film themselves or report their experience with the application. Diaries can allow users to express themselves without being interrupted or being asked personal questions [S4]. The main problem with these methods is that the data that the users choose to report might not be useful to the evaluators, and training may be necessary so users know what type of data to report.

Results for SQ4 - Type of Assessed Application

Regarding the evaluated type of application, we identified that new types of applications are arising and more methods for applications of ubiquitous computing are gaining attention in the field of UX. We identified that there are methods that are broad enough to be applied in the evaluation of any type of interface (33.9%), while others focus on Web applications (13.7%) and Mobile applications (8.8%), yet these numbers are still low.

The other types of applications category had the highest number of papers (44.1%). The methods within these papers evaluated games, recommended systems, management tools, e-learning systems, virtual reality applications, museum guide applications and others. In this sense, games are receiving more attention due to the importance of user experience in the engagement of users.

Results for SQ5 - Assessed Period of Experience

The observed period of experience details when, in the application usage, the UX observation can be carried out. If a UX instrument can be applied before using the software application, we are measuring the expectations of the user. On the other hand, UX observations performed during application usage (in momentary episodes) are more complex, as the observer must gather data while the informant is actually carrying tasks or exploring the hedonic/pragmatic aspects of the software application.

It is possible to gather data after an episodic experience. By asking questions or making a retrospection analysis, the observers can gather information on aspects that the users reflected on regarding their UX. Finally, assessing UX over time allows gathering UX data before, during and after the different cumulative episodic experiences of the user. Although this type of assessment is richer in information on the aspects that affect product attachment, its main disadvantage is that the users may not be available to carry out an evaluation for such a long period, or the development team may not have enough available time.

Our results suggest that methods evaluating UX before usage (7.9%) generally focus on user requirements (both pragmatic and hedonic) and what features the software application should have. Also, there is a shortage of methods for evaluating UX in long usage cycles (6.6%). However, these methods may not be cost effective, as they require a long observation time, or a lot of effort from users. Thus, the most applied types of evaluation regarding UX period is after using the application (71.8%) and during single episodic experiences (44.1%). The former allows evaluators to carry out retrospective analyses in order to verify the aspects that affected the UX. These methods collect data through written questionnaires or tools, interviews or think aloud protocols. Additionally, in the evaluation of single episodic experiences, sensors can be applied to gather UX data or relate the users' reports to what they actually experienced according to their physiological responses.

Results for SQ6 - Collected Data

Around 58.6% of the identified methods allow collecting quantitative data. Most of these methods gather data through questionnaires and monitoring tools, which apply scales or report on the users' physiological responses respectively. For instance, in questionnaires applying scales, evaluators can count and compare the score of a specific attribute, providing an idea of how users perceive an application and their UX with it. The main advantage of gathering quantitative data is that it is easier to analyze. However, only gathering quantitative data does not provide information to the evaluators on what are the causes for the specific scores or measures of the evaluated stimulus.

Qualitative UX evaluation methods (13.7%) have focused on gathering data through observations, open questions and interviews or think aloud protocols. Researchers in this area are trying to find new ways of facilitating the reporting by users, mainly in oral reporting. Furthermore, the use of probes is another alternative for gathering qualitative data. For instance, some papers describe technologies where cards and specific objects have been designed to facilitate users report their experience [S9]. In this sense, qualitative data such as reasons for an emotion, dislikes, improvement opportunities, and others can allow evaluators to have a holistic view of the users' experience. Nevertheless, it may be more difficult to interpret the results or understand the reported experiences.

To combine the advantages of gathering quantitative and qualitative data, researchers are developing or combining UX evaluation methods. As a result, 27.8% of the papers describe technologies for quantifying UX in terms of positive and negative emotions instances (e.g. number of smiles, degree of arousal, agreement with items, others) while trying to understand the causes for these reactions (by observing the context of use or directly asking the users) [S11]. When using these methods, evaluators can indicate the problem and explain why it happened.

Results for SQ7 - Supports Correction of Identified Problems

When considering the reviews listed in the Related Work Section, there is a lack of information on whether or not the proposed technologies support the correction of the identified problems. Therefore, SQ8 was essential to understand how many of the current proposals provided any type of support and what kind of support this was. In this sense, only 6.6% of the papers described methods in which correction support was provided. From these methods, some recommend [S6]: (a) employing the identified violated UX principals and make changes in the application to reflect those rules; (b) asking the user what could be improved; and (c) extracting improvement opportunities from the obtained data. Nevertheless, few methods suggest the specific steps for modifying the interface or the process for extracting the improvements opportunities. Additionally, as UX is personal, it may be difficult for users to suggest solutions that may apply to the majority of users without being previously tested. This result suggests that new proposals in UX evaluation could also consider indicating how to correct the identified UX problems in a way that practitioners can make improvements in the application, enhancing its UX.

Results for SQ8 – Availability

Finally, regarding the availability of the methods, 57.7% of the papers describe methods that are available for free, providing the necessary materials for their application. However, 9.7% of the papers describe methods that are only available after paying for accessing for specific equipment or artifacts, while 37.9% of the papers describe technologies that were described but do not provide information on their artifacts nor their necessary materials.

We noticed that there are methods that propose scales or questionnaires for UX evaluation do not provide the entire set of evaluated items/questions [S4]. Also, other methods that propose processes or tools do not provide the necessary artifacts to be applied neither are available for download [S10]. This feature affects how these methods can be applied in the market/research. Furthermore, there are some methods that require specific equipment to be applied, which enhances their cost [S9]. Thus, there is a need for more technologies that can be applied without costs or that are available to be employed in real development scenarios.

DISCUSSION

Implications for Research and Practice

For researchers, we discovered that most technologies focused on different aspects not always present on every technique. On one hand, quantitative data is mostly analyzed when monitoring users' physiology or categorizing the overall opinion of users through scales. On the other hand, qualitative data is analyzed when digging deeper into the users' opinion, or when trying to understand the causes for a poor experience. Therefore, new UX evaluation technologies should take into account all existing aspects in order to include them, not duplicate them, and not confuse them; therefore, allowing the UX evaluation technologies to provide more consistent and complete evaluation reports.

Our results show that there is a shortage of UX evaluation technologies able to provide improvement suggestions once a problem is found. Methods providing such information could be useful for software development teams with low experience in the use of UX evaluation approaches and the correction of the identified problems. Therefore, new research should guide the development of approaches that provide assistance to practitioners willing to improve the quality of the evaluated application in terms of UX.

Our findings regarding the evaluated period of experience suggests that further attention should be given to developing methods for evaluating UX before usage and during long term usage situations. Additionally, the methods that are already published should also be made available in order to allow their use in the software development market. Although some methods still require specific equipment to be applied, the others that do not should describe the process for their application and the necessary items/questions/artifacts that should be used.

We found out the current UX evaluation technologies mostly focus on the evaluation of prototypes or finished versions of an application. Since correcting software problems during the last stages of the development process can be a costly activity, new research could be oriented towards evaluating early artifacts such as the requirement specifications or ideas of an application.

Practitioners must consider that none of the describe UX evaluation technologies is able to identify all problems, specifically when considering different user profiles. However, they can be combined to improve evaluation results and find more problems according different types of users. Therefore, practitioners can apply these technologies in different development phases to gather both quantitative and qualitative data and increase their effectiveness.

Limitations from the Systematic Mapping Study

To mitigate the threat of not including a relevant paper, we prepared our search string considering previous reviews [3,22] in order to consider all possible terms used up to 2010 for denoting UX evaluation technologies. After verifying the returned papers from applying our search string, the papers used as basis for developing our search string were returned. Thus, by applying that measure, the probability of missing relevant papers has been reduced.

The selection process largely depends on the personal knowledge and experience of the researchers who conducted this systematic mapping study, which might have introduced bias to the selection results. Thus, we applied inclusion and exclusion criteria for paper selection and a second researcher reviewed the paper selection process providing feedback on its execution.

The data extraction results might have been negatively affected due to the bias of the researcher who extracted the data. Bias on data extraction may result in an inaccuracy of the extracted data items. This bias was mitigated with two measures: (a) a list of extracted data items was specified in detail to reduce possible misunderstandings on the data items to be extracted; and (b) we carried out discussions on extracted data items from the researchers throughout the whole data extraction process to improve the consistency and correctness of the data extraction results.

CONCLUSIONS AND FUTURE WORK

This paper presented an analysis of UX evaluation technologies continuing the work by Vermeeren et al. [22] through a systematic mapping study. From the initial set of

2101 papers, we extracted 227 papers reporting technologies from the years 2010 to 2015.

We noticed that there is a need for methods that allow identifying more qualitative data besides rating the UX, as this can help evaluators understand the cause of poor experiences or ways to improve them. In this sense, we also noticed that few methods provide means to facilitate the correction of the identified problems. This is essential, as novice software practitioners require guidance in order to make changes in the evaluated application to improve its quality. Additionally, there is a need for further methods gathering UX data at different time stamps of the use of an application, while considering specific features of types of applications (e.g. mobile, web, others). Finally, there is a need for methods that are available for practitioners without the need of costly equipment, enabling small teams to improve and increase the acceptance of their applications.

Each type of technology has been discussed for each of our sub-questions to provide in which context, they could be more appropriate. Also, each technology has been categorized, providing a list in [20]. For practitioners, these results could be useful for integrating UX evaluation technologies into the software development process; whereas for researchers, it would be interesting to propose (or adapt) technologies according to the identified gaps.

When comparing our review with the ones cited in the related work section, we identified that several of the factors cited by them were similar to the ones we identified in our review. For instance, some methods focus on the evaluation of emotions in terms of arousal, dominance and control, while others focus on how specific emotions are evoked by the software application. In this case, when developing games or applications for children, the development team focuses on fun and engagement. Other methods, on the other hand, focus on specific attributes of product attachment over time, and a combination of hedonic and pragmatic attributes through scales. Still, we highlight that our review is broader, which can provide an overview of how UX evaluation technologies have been applied. As future work, we intend to update this SM in order to increase the body of knowledge with new technologies, allowing identifying new research gaps.

ACKNOWLEDGEMENTS

We thank the financial support granted by the CAPES AUXPE-PROCAD/NF 175956/2013; FAPEAM project: 062.00578/2014; and CNPq for the scholarship granted to the first author of this paper.

REFERENCES

1. Stephanie Balters, and Steinert Martin. 2015. Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices. *Journal of Intelligent Manufacturing* 28, 1: 1-23.
2. Simone Barbosa, and Bruno Silva. 2010. *Human Computer Interaction*. Elsevier Brasil. (*In Portuguese*)
3. Javier A. Bargas-Avila, and Kasper Hornbæk. 2011. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2689-2698.
4. Javier A. Bargas-Avila, and Kasper Hornbæk. 2012. Foci and blind spots in user experience research. *Interactions* 19, 6: 24-27.
5. Pieter Desmet. 2003. Measuring emotion: Development and application of an instrument to measure emotional responses to products. In *Funology*, 111-123.
6. Adrian Fernandez, Emilio Insfran, and Silvia Abrahão. 2011. Usability evaluation methods for the web: A systematic mapping study. *Information and Software Technology* 53, 8: 789-817.
7. Jérémy Frey, Christian Mühl, Fabien Lotte, and Martin Hachet. 2013. Review of the use of electroencephalography as an evaluation method for human-computer interaction. In *arXiv preprint arXiv*, 1311.2222.
8. Bill Gaver, and Heather Martin. 2000. Alternatives: exploring information appliances through conceptual design proposals. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 209-216.
9. Marc Hassenzahl. 2003. The thing and I: understanding the relationship between user and product. In *Funology*. 31-42.
10. Marc Hassenzahl, Sarah Diefenbach, and Anja Göritz. 2010. Needs, affect, and interactive products–Facets of user experience. *Interacting with computers* 22, 5: 353-362.
11. International Organization for Standardization, ISO DIS 9241-210. 2010. Ergonomics of human system interaction - Part 210. *International Standardization Organization (ISO)*.
12. Minna Isomursu, Marika Tähti, Soili Väinämö, and Kari Kuutti. 2007. Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies* 65, 4: 404-418.
13. Patrick W. Jordan. 2002. *Designing pleasurable products: An introduction to the new human factors*. CRC press.
14. Barbara Kitchenham, and Stuart Chartes. 2007. Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3 EBSE Technical Report*. EBSE.

15. Barbara Kitchenham, Rialette Pretorius, David Budgen, O. Pearl Brereton, Mark Turner, Mahmood Niazi, and Stephen Linkman. 2010. Systematic literature reviews in software engineering–a tertiary study. *Information and Software Technology* 52, 8: 792-805.
16. Sari Kujala, Virpi Roto, Kaisa Väänänen-Vainio-Mattila, Evangelos Karapanos, and Arto Sinnelä. 2011. UX Curve: A method for evaluating long-term user experience. *Interacting with Computers* 23, 5: 473-483.
17. Effie Lai-Chong Law. 2011.The measurability and predictability of user experience. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems*, 1-10.
18. Elizabete Munzlinger, Fabricio Batista Narcizo, and José Eustáquio Rangel De Queiroz. 2012. Systematization of literature reviews in the HCI area. In *Companion Proceedings of the 11th Brazilian Symposium on Human Factors in Computing Systems*, 51-54. (*In Portuguese*)
19. S. Rajeshkumar, Omar Ridha, and Murni Mahmud. 2013. Taxonomies of User Experience (UX) evaluation methods. In *International Conference on Research and Innovation in Information Systems* (ICRIIS), 533-538.
20. Luis Rivero, and Tayana Conte. 2016. TR-USES-2016-002 Systematic Mapping on UX Technologies: Extraction of User eXperience Evaluation Methods. *USES Research Group Technical Report.* Available at: http://uses.icomp.ufam.edu.br/relatorios-tecnicos/
21. Patricia Tavares, Luiz Agner, and Simone Bacellar Leal Ferreira. 2010. Usability Evaluation of mobile devices for the collection of neighborhood data through interviews based on scenarios and tasks. In *Proceedings of the IX Symposium on Human Factors in Computing Systems*, 11-20. (*In Portuguese*)
22. Arnold P. O. S. Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 521-530.

REFERENCES FROM THE SYSTEMATIC MAPPING

[S1] Le An, Songfan Yang, and Bir Bhanu. 2015. Efficient smile detection by extreme learning machine. *Neurocomputing* 149, 1: 354-363.

[S2] Carol M. Barnum, and Laura A. Palmer. 2010. More than a feeling: understanding the desirability factor in user experience. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 4703-4716.

[S3] Marios Belk, Panagiotis Germanakos, Panagiotis Zaharias, and George Samaras. 2012. Adaptivity Considerations for Enhancing User-Centric Web Experience. In *The Fifth International Conference on Advances in Computer-Human Interactions* (ACHI 2012), 348-353.

[S4] Luca Colombo, and Monica Landoni. 2014.A diary study of children's user experience with EBooks using flow theory as framework. In *Proceedings of the 2014 conference on Interaction design and children*, 135-144.

[S5] Reeva Lederman, Greg Wadley, John Gleeson, Sarah Bendall, and Mario Álvarez-Jiménez. 2014. Moderated online social therapy: Designing and evaluating technology for mental health. *ACM Transactions on Computer-Human Interaction* 21, 1: 5.

[S6] Philip Lew, Luis Olsina, and Li Zhang. 2010. Integrating quality, quality in use, actual usability and user experience. In *6th Central and Eastern European Software Engineering Conference (CEE-SECR)*, 117-123.

[S7] Llúcia Masip, Marta Oliva, and Toni Granollers. 2012. The open repository of heuristics. In *Proceedings of the 13th International Conference on Interacción Persona-Ordenador*, 4.

[S8] Pejman Mirza-Babaei, and Lennart Nacke. 2014. Introducing the biometric storyboards tool for games user research. In *Games Media Entertainment* (GEM), 1-7.

[S9] Georg Regal, Marc Busch, Christina Hochleitner, Peter Wokerstorfer, and Manfred Tscheligi. 2014. TACTUX: a tactile user experience assessment board. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 983-986.

[S10] Jacopo Staiano, María Menéndez, Alberto Battocchi, Antonella De Angeli, and Nicu Sebe. 2012. UX_Mate: from facial expressions to UX evaluation. In *Proceedings of the Designing Interactive Systems Conference*, 741-750.

[S11] Jeroen Vanattenhoven, and Dries De Roeck. 2009. Investigating the Use of the Experience Clip Method. In *International Conference on User Centric Media*, 79-86.