Which Instrument Should I Use? Supporting Decision-Making About the Evaluation of User Experience

Ticianne Darin(✉), Bianca Coelho(✉), and Bosco Borges(✉)

Virtual University Institute, Federal University of Ceará, Humberto Monte, S/N, Fortaleza, Brazil
ticianne@virtual.ufc.br, biancasmd@alu.ufc.br, boscofilho4@gmail.com

Abstract. User Experience (UX) has been intensively investigated lately, resulting in the proposal of several evaluation instruments, methods, and techniques. However, the definition of UX and its constructs is still a work in progress, making User Experience a concept open to various interpretations. Consequently, the development of UX evaluation methods and instruments rely on very different assumptions, often making professionals and beginning researchers uncertain about choosing the right methods to evaluate user evaluation aspects. Aiming to help fill in this gap, in this work we present the results of a systematic snowballing procedure conducted to investigate the characteristics of the UX evaluation instruments that have been proposed and used by HCI community in the last years. We compiled information about 116 instruments aiming to assist researchers and practitioners in making informed choices about which instruments can support UX data collection, according to their research goals. In addition to that, the data analysis provided a glance on the directions the research on UX evaluation instruments is taking.

Keywords: User Experience (UX) · Evaluation methods · UX measurement

1 Introduction

The last years in Human Computer Interaction (HCI) literature have been characterized by an intense and rich exploration of user experience (UX) related concepts. Researchers have been investigating definitions and understanding of UX across different cultures and perspectives [1–3]; establishing concepts, frameworks and models for supporting design and development processes [4, 5]; and developing and evaluating methods, techniques, instruments and measures for evaluating UX [6–9]. Particularly, researchers have been calling attention to the relevance of and the need for a theoretical discussion around UX research and practice [10]. The theoretical roots used to develop different types of UX work are a broad work in progress, including a range of different types of theories, models and conceptual frameworks [1]. For instance, UX research has been based on psychological models and theories, formalist aesthetics and product semiotics, Gestalt theory, theories about communication, and theories inspired by the art and design fields [1]. Consequently, there are several possible understandings about

A. Marcus and W. Wang (Eds.): HCII 2019, LNCS 11586, pp. 49–67, 2019.
https://doi.org/10.1007/978-3-030-23535-2_4

the meaning of UX, each proposing different approaches for evaluating its qualities, which results in broadly different evaluation methods, techniques, and instruments.

Despite the established standards that define usability (ISO 9241-11:2018) and UX (ISO 9241-210), there is also a growing discussion intending to clarify the distinction between these often confused concepts [78: 104]. Because UX remains a rather vague concept, difficult to fully understand for both researchers and practitioners [11, 12], UX measurement is frequently confused with usability measurement while satisfaction, which is a component of usability [13, 14], is indistinctly treated as a UX quality - sometimes the one and only necessary to assess user experience [15]. Besides, the literature has demonstrated that several factors, including social and cultural changes, can directly interfere in the way UX is understood and hence practiced [12, 16].

The problem behind UX underdeveloped concept is the danger that user experience and its related concepts such as trust, loyalty, identity, and engagement will not be fully realized in studies of people and technology [17]. In this scenario, selecting a combination of UX evaluation methods commonly relies on individuals' experience and expertise rather than on information about the UX constructs that can be measured in empirical studies [18], and on which instruments can support UX measurement [19]. Although the literature has not yet established standard UX metrics and several philosophical arguments on UX measurement have been raised [18], evaluators should not conduct UX evaluations mostly based on their personal experience and a very restricted knowledge of the methods and instruments employed [20]. There is a huge need for UX professionals, researchers and HCI learners to make informed and conscious choices to select right instruments and methods when evaluating UX qualities [21].

Aiming to help fill in this gap, in this work we present the results of a systematic snowballing procedure [22] conducted to investigate the characteristics of the UX evaluation instruments that have been proposed and used by HCI community in the last years. Our main goal in this research is to compile a large quantity of knowledge covering a wide variety of types of UX evaluation instruments, updating the literature on UX evaluation, and to provide researchers and practitioners with a useful catalog of UX instruments. We present a compilation of 116 instruments to assist researchers and practitioners in making informed choices about which instruments can support UX data collection, according to their research goals. In addition to that, the data analysis provided a glance on how the initial list of instruments evolved, allowing us to contribute with a discussion on the directions the research on UX evaluation instruments is taking.

2 Related Work

For a long time, usability figured as the main HCI criterion on which researchers and practitioners relied for measuring the quality interactive systems interaction. According to Bargas-Avila and Hornbæk, usability focus on the efficiency and the accomplishment of tasks was one of the instigating factors for the development of user experience as a concept of quality of use that addresses hedonic qualities and emotional factors, in addition to the utility and pragmatic aspects commonly covered by usability [15].

Bevan, Nigel, and Miles presented an overview of tools developed to assess user performance, user satisfaction, cognitive workload, and analytic measures [23]. Some of the concepts they analyzed, such as perceived usability, were later understood as part of UX qualities.

Agarwal and Meyer conducted a survey to list existing instruments, motivated by the idea of identifying methods that went beyond usability, i.e., methods that more explicitly included emotions, relating directly to User Experience [24]. They identified verbal, nonverbal and physiological measurement tools and discussed that good usability metrics are often indicative of good user experience. Roto, Obrist and Väänänen-Vainio-Mattila categorized User experience Evaluation Methods (UXEM) for academic and industrial contexts, gathered in a special interest group session (SIG) [25]. They distinguished UX and usability methods based on pragmatic and hedonic model [26], and classified them according to their methodology.

Reviews have also been conducted to investigate UX evaluation in specific application domains and UX measurement. Ganglbauer et al. conducted an overall review about psychophysiological methods used in HCI, describing in details different psychophysiological methods, such as electroencephalography (EEG), electromyography (EMG), electrodermal activity (EDA) [27]. Nacke, Drachen and Göbel presented a classification of methods to measure Game Experience, presenting three categories of experience related to games: (1) quality of product, (2) quality of human-product interaction and (3) quality of interaction in a given social, temporal, spatial or other context. Yiing, Chee and Robert described categories for HCI qualitative methods that are used to evaluate interface of video-games, focusing in Affective User-Centered Design [28]. Aiming to guide professionals in choosing UX methods, they classified methods into user feedback and non-invasive categories. Hung and Parsons conducted a survey with emphasis on the Engagement construct, cataloging instruments self-reported instruments related to UX, engagement, communication, emotion, and other qualities, excluding later those that did not belong to the HCI field [29].

Although several important studies have been investigating different types of UX evaluation methods and instruments, Vermeeren's et al. list of UX evaluation methods was used as basis for our work, as it consists in one of the most complete and well-known compilation of methods presented in UX literature [30]. They collected data from workshops and SIGs, and also searched literature for previously categorizations of UX methods. As a result, they categorized 96 methods according to specific information, such as study type, development phase, requirements, type of approach and applications. In the present work, we chose to focus on practical and well defined evaluation instruments, instead of including UX frameworks, techniques, methods and models. Hence, we analyze the original instruments listed by Vermeeren's et al. under a different point of view, including new instruments.

3 Methodology

The present study classifies and catalogs a set of 116 UX evaluation instruments gathered from a snowball sampling [22], which consists in gathering research subjects through the identification of an initial subject which is used to provide other related

subjects. Our initial subject was a subset of the papers listed by Vermeeren et al., which is a seminal and highly cited paper in the area. This subset consists in the 39 papers that describe UX tools and instruments in Vermeeren's et al. list, since it also includes methods, models and frameworks, which are out of this work scope. The 39 papers - which describe 49 UX evaluation instruments - were used as start set in the snowballing technique. As a result, we obtained a final set of 116 instruments, which include updated versions of the ones originally listed, in addition to novel instruments proposed for different domains, such as the Internet of Things, and for specific audiences, for example, children.

Figure 1 provides a schematics of the methodology followed in this research, which have three main steps: (1) Selection of Initial Set of Instruments, (2) Snowballing [22] and (3) Instruments Cataloging.

The first step of this research methodology was to select a start set of papers to use for the snowballing procedure. Having chosen Vermeeren's et al. list as our basis, we analyzed the 86 UX evaluation artifacts[1] made available by authors in their site *allaboutux.com*. For each artifact, two researchers independently read the name, description, and intended applications. For purposes of selecting the papers to be included in the start set, we considered UX evaluation instruments as *planned and validated tools designed to systematically collect qualitative data/measure quantitative data related to UX constructs from a variety of participants, producing results based on psychometric properties in a format ready for analysis/interpretation*. The two sets selected by the researchers were later compared and consolidated, after being checked by an expert researcher, which resolved the inconsistencies. The inclusion criteria were: (a) the artifact must match the adopted definition of UX evaluation of instrument; (b) the artifact paper must be available in Portuguese, English or Spanish; and (c) the paper have to be available in a digital library. By the end of this phase, we had selected 39 papers as our start set.

Then, to execute the snowballing sampling, the 39 papers were distributed between two researchers, which applied independently a forward snowballing technique in order to find new instruments. Forward snowballing refers to identifying new papers based on those papers citing the paper being examined [5]. The citations to the paper being examined were studied using Google Scholar and, for each original paper, researchers verified how many times it had been cited by others. If the number of citations was greater than 100, they should select among them the 25 most relevant papers in addition to the 10 most recent articles, totalizing 35 new papers for each original papers with more than 100 citations. In case the original paper had less than 100 citations, the 25 most relevant papers were included. We acknowledge this procedure limited our capacity to catalog as many instruments as possible. Given our constraints, however, we adopted this procedure to make significant work more likely to be included, as well papers from authors that regularly publish in the area.

[1] Although the original research paper reports 96 methods, the online list at allaboutux.com, mentioned by the authors in the same paper, currently only presents 86 evaluation artifacts (among methods, frameworks, models and instruments).

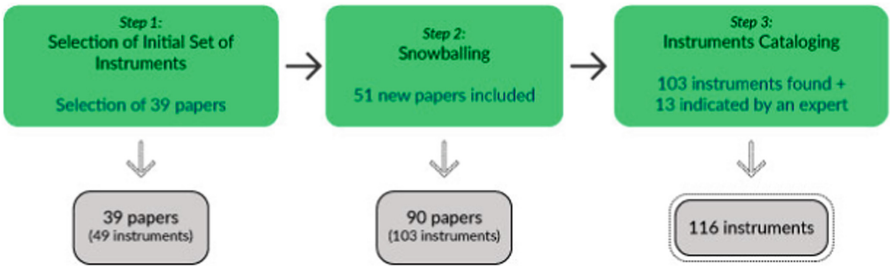

Fig. 1. Summary of methodology steps.

Given these criteria, each candidate paper citing the original paper was examined. The first screening was done based on the reading of paper title, abstract, and keywords. If this information was insufficient for a decision, the citing paper was studied in more detail and the place citing the paper already included was examined. If this was insufficient too, then the full text was studied to make a decision regarding the new paper. The goal was to identify any evidence that the citing paper proposed a new UX evaluation artifact or an update of an existing one. In this phase, from the 39 start set, 1001 citing papers were screened, 221 papers were read and analyzed, resulting in the inclusion of 51 papers. By the end of this phase, we had a set of 96 papers describing 103 UX evaluation instruments.

Finally, the instrument cataloging step consisted in the data extraction of the selected papers. In this step, 13 new papers were included after the indication of a senior researcher, helping to mitigate the limitation of our paper search process. Two researchers read the full text of 103 papers describing 116 UX evaluation instruments - as some papers described more than one instrument [e.g. 31 and 32], and cataloged them. The cataloging process consisted of extracting and tabulating the following data for each instrument: reference, publication year, instrument name, type of instrument (scales, psychophysiology, post-test pictures, two-dimensional graph area, other [21]), UX qualities (overall UX, affect, emotion, fun, enjoyment, aesthetics, hedonic, engagement, flow, motivation [15, 21]), type of approach (quantitative, qualitative or quali-quantitative), main idea, general procedure, applications, and target users.

The oldest instrument cataloged is from 1982 [33], and the newest are from 2018 [34, 35]. The complete categorization of the 116 UX evaluation instruments is available at https://bit.ly/2N7K2ly. We intend to keep periodically updating and expanding the information available.

4 Results

From the 116 UX evaluation instruments identified, 48 (41.38%) come from the start set of papers gathered from Vermeeren's et al. list, and 68 (58.62%) are instruments developed from 2011 onwards, identified using the methodology described before. The cataloged instruments reported addressing 29 different UX qualities, which can be evaluated by eight different types of instruments, as exemplified in Table 1.

Table 1. Examples of UX qualities evaluated by the different types of instruments

UX quality	Types of instrument	Ex.	UX quality	Types of instrument	Ex.
Aesthetics	Scales/Questionnaires	[39]	Human-robot trust	Scale/Questionnaire	[42]
Affect	Scales/Questionnaires; Psychophysiological; Post-test picture/object; Two-dimensional diagram/graph area; Others	[40]	Immersion	Scale/Questionnaire	[43]
Appraisal	Scales/Questionnaires	[41]	Intrinsic motivation	Scale/Questionnaire	[33]
Emotion	Post-test Picture/object; Scale/Questionnaire; Two-dimensional Diagrams/Graph area; Psychophysiology; Others	[31]	Stress	Software/equipment	[44]

Scales and questionnaires constitute 62.07% of the 116 instruments identified. The second most common type of instrument is classified as psychophysiology (10.34%), followed by two-dimensional diagrams/graph area (7.76%) and software/equipment (7.76%), and post-test picture/object (6.90%). Other types of instruments occurred less frequently, being usually developed for specific contexts, such as: diary templates [36, 37], scale combined with two-dimensional graph area [16, 38] and observational checklist [34]. These trends are suggestive of the directions research has taken in this field, and are further described in the remainder of this Section.

We classified scales and questionnaires in the same category ("scale/questionnaire"), although we acknowledge there is a conceptual difference between their definitions. Still, we grouped them together because often authors use the terms interchangeably and, in some cases, scales are developed for specific questionnaires [45]. Good questionnaires can be described as a well-defined and well-written set of questions to which an individual is asked to respond open-ended or closed-ended questions [46]. Scales are used in closed-ended questions to support an ordered response from a number of given choices, in some logical order [47].

The prevalence of self-reported UX data collection is clear in the 72 scales/questionnaires identified, which report to evaluate a range of 26 different UX qualities (Table 2). From these, seven (9.72%) evaluate general aspects of UX (i.e. the authors do not describe any specific UX quality) such as [48], seven (9.72%) evaluate specific sets of UX qualities [49], as shown in Table 3, and six (8.33%) evaluate satisfaction [50]. It is important to notice that is out of this research scope to analyze whether different terms employed by authors refer to a same UX quality.

The UX scales found target nine different types of application. Thirty seven (51.39%) are classified as "application-independent" (i.e. they are reportedly suitable to evaluate UX in three or more types of application), such as [51]. Thirteen (18.06%) aims to evaluate UX in games and virtual environments [52], eight (11.11%) are focused on online platforms [45], four (5.56%) are for mobile devices and three (4.17%) target intelligent systems, environments and objects [53].

Table 2. Examples of UX qualities evaluated by scales/questionnaires

Scales/Questionnaires			
UX quality	Examples	UX quality	Examples
Affect	[40]	Human-robot trust	[42]
Aesthetics	[39]	Immersion	[57]
Appraisal	[41]	Intrinsic motivation	[33]
Aspects of game experience	[54]	Mental effort	[44]
Cognitive absorption	[55]	Perceived usability	[58]
Cybersickness	[56]	Presence	[57]

Table 3. Examples of specific sets of UX qualities evaluated by scales/questionnaires

Scales/Questionnaires			
Set of UX qualities	Reference	Set of UX qualities	Reference
Challenge & control Fantasy Creative and constructive expressions Social experiences Body and senses	[59]	Attractiveness Perspicuity Efficiency Dependability Stimulation Novelty	[60]
Usability Trust Appearance Loyalty	[49]	Affect Efficiency Learnability Helpuness Control	[51]

The variety of UX qualities evaluated by scales and questionnaires is greater than in other types of instruments. While the 72 scales and questionnaires measure 26 different UX qualities, the remaining 44 instruments evaluate only 8 different qualities. Regarding the target users, 58 scales/questionnaires (80.56%) aim to evaluate user experience for all type of users [51], while eight (11.11%) are aimed at children [61], five (6.94%) were designed for users performing specific roles, such as journalists [48] and consumers [62], and one scale/questionnaire (1.39%) is aimed at people with disabilities [50]. Scales and questionnaire are more common than other types of instruments, nevertheless, this predominance seems to be decreasing over the last years, considering the cataloged instruments. Between 1982 and 1999, 14 out of 16 (87.50%) instruments are scales/questionnaires, while in the next 10 years (2000 to 2009) its quantity drops to 33 out of 48 (68.75%). Finally, from 2010 to 2018, it constitutes 25 out of 52 (48.08%) of the UX instruments identified (Fig. 2).

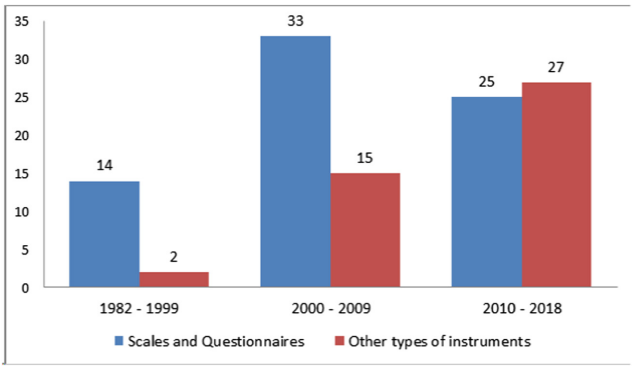

Fig. 2. Comparison between scales/questionnaires and other instruments by year.

4.1 Psychophysiological, Graphs, Software Instruments and Post-Test Pictures

The second most recurrent type of UX evaluation instrument cataloged is psychophysiological, in which user's physiological responses are recorded and measured, usually with sensors attached to the participant. The 12 psychophysiological instruments (10.34%) identified reported evaluating 4 different UX qualities: affect, emotion, generic user experience and specific sets of qualities (Table 4). Nine of the psychophysiological instruments found (75%) evaluate emotion [63]. One of them evaluates affect [64], another evaluates generic user experience [65] and the other targets a set of UX qualities: emotion and perception [66]. Most of these instruments are "application-independent" (91.67%), and one is specific to evaluate UX in audiovisual applications [65]. The most common purpose of the psychophysiological instruments found is to measure emotion in any types of applications (75%). All the 12 psychophysiological instruments aim to evaluate user experience for all types of users.

The third most common type of instrument identified is two-dimensional diagrams/graph area (7.76%). This category of type of instrument covers diagrams, charts, timelines and two-dimensional graph areas through which the users can report their experiences. We found nine instruments of this category, which evaluate three types of UX qualities. Four of these instruments (44.44%) evaluate emotion [63] and also four (44.44%) evaluates specific sets of UX qualities, such as attractiveness (appeal) of the product, ease of use and utility [67], and usability, challenge, quantity of play and general impression [19], while one (11.11%) evaluates affect [68]. In addition to application-independent instruments [63] - which were the most common target of this type of instrument (66.67%) - they target four different types of application: audiovisual [69], games and virtual environments [19] intelligent systems, environments and objects [70]. Eight of the two-dimensional diagrams/graph area instruments (88.89%) target all types of users, and one aim to evaluate UX specifically for children [19].

We also identified nine UX evaluation instruments developed as software or specific equipments. They evaluate seven different UX qualities: affect [71], aspects of

Table 4. Examples of UX qualities evaluated by psychophysiological, two-dimensional diagrams/chart area and software/equipment instruments.

Psychophysiological		Two-dimensional		Software/Equipment	
UX quality	*Examples*	*UX Quality*	*Examples*	*UX quality*	*Examples*
Affect	[64]	Emotion	[63]	Feelings	[74]
Emotion	[63]	Affect	[68]	Aspects of game experience	[72]
Generic user experience	[65]	Emotion and contextual details	[32]	Stress	[75]
Emotion and perception	[66]	Attractiveness (appeal) of the product, ease of use and utility	[67]	Behavior	[73]

game experience [72], behavior [73], emotion [35], feelings [74], stress [75] and generic user experience [76]. The most common target of software instruments are aspects of game experience (33.33%). The software/equipment group aims to evaluate games and virtual environments [77] and online platform [76], besides those that are application-independent [75]. With regard to the target users, eight (88.89%) software/equipment instruments evaluate user experience for all type of users and one (11.11%) is specific for product customers [74].

Eight of the 116 instruments are post-test pictures/objects. Among of these, three different UX qualities were revealed: emotion, evaluated by five (62.50%) instruments, affect, evaluated by two (25.00%) instruments and one (12.50%) instrument is to evaluate a specific set of UX qualities [78], which consists in: emotion, ease of use, usefulness and intention to use. Six (75.00%) of the post-test pictures/objects instruments evaluate UX holistically and two (25.00%) are for intelligent systems, environments and objects. One (12.50%) of these instruments are specifically to evaluate UX for childrens [79] and the seven others (87.50%) are suitable for all types of users.

4.2 Catalog of UX Evaluation Instruments

The set of UX evaluation instruments identified was organized as a catalog, systematizing and relating the data extracted from the papers describing each instrument. The catalog compiles 116 instruments intending to assist researchers and practitioners in making informed choices about which instruments can support UX data collection, according to their research goals. For now, the catalog is presented as a set of spreadsheets, but as this research goes on we will periodically update the information available. As future work, we are going to develop and make available an interactive version of the catalog. For each instrument, the catalog describes: reference, publication year, instrument name, main idea, general procedure, type of instrument, type of approach, UX quality, target users and applications.

Main idea and general procedure are summarized textual descriptions that provide the reader, respectively, with an overall understanding of the purpose of an instrument, and with information on how to conduct an evaluation using it. Instruments are categorized in six categories: scale/questionnaire, psychophysiological, software/equipment, two-dimensional diagrams/graph area, post-test pictures/objects and others. Each category groups two or more types of instruments (Table 5).

Table 5. Categories of types of instruments

Scales/Questionnaires			
Categories	Types of instruments included	Categories	Types of instruments included
Post-test picture/objects	Tangible manikins; shaped objects; post-test picture	Psychophysiological	Eyetracker-based; polygraphic data; multimodal sensor; psychophysiological
Two-dimensional diagram/chart area	Diagrams; charts; two-dimensional graph areas	Scale/questionnaire	Scales; questionnaires
Software/equipment	Auto text collector software; maps/geographic information; survey tool; telemetry tool; software	Others	Observational checklist; diary entry templates; combined scales and two-dimensional graphs

The type of approach can be either qualitative, quali-quantitative, or quantitative. Applications are divided in eight categories: (i) online platform, (ii) audiovisual, (iii) intelligent systems, environments and objects, (iv) games and virtual environments, (v) hardware and robotics, (vi) mobile devices, and (vii) e-learning. Besides these, there is also the application-independent category that describes instruments aiming to evaluate UX in three or more different types of applications. Each category includes two or more types of applications cited in instruments papers (Table 6). Finally, target users are categorized in four main groups: children, people with disabilities and role-specific, a category that characterize instruments developed to evaluate UX with persons performing specific functions or roles. Besides, the category "all types of users" describes instruments aiming to be used with any users.

All these aspects are presented in the catalog as filters, to help people interested in conducting UX evaluation in analyzing which instruments to choose, depending on the goals of their evaluation, the type of application, the UX qualities to be evaluated, and the target users. In addition to present the classification of each instrument, the catalog also shows the relationships between categories and instruments descriptors, as depicted in Fig. 3. The full version of the catalog can be accessed in this link https://bit.ly/2N7K2ly.

Table 6. Categories of types of applications

Category	Applications included	Category	Applications includes
Online platform	Web, shopping websites, and Social systems	Games and VE	Virtual environments; games; virtual reality
Application independent	Generic and those addressing three or more applications	Hardware and robotics	Robots; hardware
Audiovisual	360 videos; Generic audiovisual; Gameplay streaming	Mobile devices	Mobile; portable interactive devices (PIDs)
Intelligent systems, environ. and objects	Ambient intelligent environment; Internet of Things; home appliances; automated systems	E-learning applications	Educational software for children; Internet-based learning systems

Mobile Devices	Scale / Questionnaire	Generic User Experience	Attrak-Work Questionnaire; Emoji UX Questionnaire
		Success	[no name informed]
		Usability	Mobile Phone Usability Questionnaire
	Diary Template	Generic User Experience	Experience Diary
E-learning applications	Scale / Questionnaire	Usability	Pedagogically Meaninful Learning Questionnaire; [no name informed]
		Cognitive absorption	TAM
Online Platform	Scale / Questionnaire	Aesthetics	Aesthetics scale; VisAWI
		Service Experience	ServUX questionnaire
		Satisfaction	WAMMI; End-user computing satisfaction; WEBsite USability Evaluation Tool
		Usability	UFOS-V2
		Usability, Trust, Appearance and loyalty	SUPR-Q
	Software/equipment	Generic User Experience	UUX-Posts
Laboratory Experiments	Scale / Questionnaire	Intrinsic motivation	Intrinsic Motivation Scale

Fig. 3. Portion of the UX instruments catalog.

5 Discussion

The term "instrument" is traditionally associated with the measurability of UX qualities [18, 21]. Although it is far from our intention to define what characterizes a user experience instrument, in this research, UX instruments are addressed in a broader way. They are seen as evaluation artifacts designed to collect user data and to facilitate observation and/or measurement of UX qualities. In our understanding, given the nature of the user experience, qualitative and quantitative approaches have to be articulated to a thorough evaluation and deeper understanding of the UX qualities. Hence, our focus is on stimulating practical UX work by cataloging tools with diverse

approaches, designed to systematically collect data related to UX constructs from a variety of participants.

In the remainder of this section we discuss some insights, trends and concerns yielded during data analysis about the use and development of UX evaluation instruments and how they incorporate UX qualities.

5.1 (Re)Use of UX Evaluation Scales/Questionnaires

Overall, our findings show a consistent picture that indicates scales and questionnaires as the most common types of UX instruments, also addressing a greater variety of UX qualities than all the other types of instruments. This indicates that the trend once identified in the early 2010s [21], that scales are commonly used with most UX qualities, remains unchanged. However, a rising trend seems to be combining traditional techniques for capturing self-reported data with UX measurement, in quali-quantitative approaches.

Some parsimony is necessary in the development and utilization of UX evaluation questionnaires, since this type of instrument can either be structured, well-tested, robust, and result in data with a high level of validity, or poorly done, resulting in data of questionable validity [46]. This type of instrument is often used not because it is the most appropriate method but because it is the easiest method [46]. A clear example of this situation is the experience report presented by Lallemand and Koenig [7] in which they report a bad experience using a UX questionnaire that was supposed to be standardized and validated. The problem they faced came from the fact that scales are often considered validated after a single validation study which leads to conclusion that the scale psychometrics properties are good and can therefore be considered as valid.

Hence, before creating new UX scales we must consider if, given the great quantity and variety of existing instruments, it is really necessary to create new ones. Wouldn't these instruments be more robust if we focused our efforts on validating, translating, expanding and improving already existing scales? It would be an effective way to improve UX instruments and make them suitable for the widest range of users possible. This is the case of MemoLine [19], an adaptation for kids arised from UX Curve [67]; AttrakWork [48], that proposes to "support the evaluation of user experience of mobile systems in the context of mobile news journalism" which is based on AttrakDiff [79]; and TangiSAM [80], a Braille adaptation of the Self Assessment Manekin [40]. In this regard, researchers have been discussing about the holistic UX questionnaires trend to follow a "one size fits all" approach [7]. In this regard, we agree with Lallemand and Koenig [7] when they state that the development of more specific methods, targeted at particular application domains is necessary. We further add that more generic evaluation instruments that already exist should be used as basis for this development.

5.2 Different Perspectives on How to Consider UX Qualities

Several instruments propose evaluating specific UX qualities such Emotion, Affect, Presence and Immersion, or even a specific set of qualities, such as Aesthetics and Emotion combined [53]. The fact that in the last years more instruments are focusing on evaluating the subjective components of user interaction is positive, because it

demonstrates that researchers begun to reflect in a more deeply way about the specificities of user experience. Thus, a broad spectrum of UX qualities have been evaluated, addressing particular types of applications and users' characteristics. Some important examples of contributions designed to evaluate UX in specific and complex situations are a questionnaire developed for measuring emotions and satisfaction of Brazilian deaf users [50], and a scale developed to measure specific sets of UX qualities with preschoolers: (1) challenge and control, (2) fantasy, (3) creative and constructive expressions, (4) social experiences and (5) body and senses [61].

In a different direction, as listed in Sect. 4, some instruments have been proposed to evaluate UX without specifying which qualities are taken into account [37], following a more generalist UX evaluation approach. This can be a consequence of the lack of consensus about what User Experience means [12], since the different understandings of this concept impacts the effectiveness, development and even teaching of this discipline [81]. There are also instruments that define User Experience as a sum of qualities [e.g. 82 and 31]. Those were classified by us as instruments that measures specific sets of UX qualities. However, the set of qualities that characterize UX varies widely from one instrument to another, which is, again, a consequence of the lack of a shared UX definition.

These situations depict a scenario where the term User Experience seems to be used almost instinctively in some cases, making it hard to know what is assessed when an instrument claims evaluate UX. For instance, [83] and [67] are respectively a questionnaire and a two-dimensional graph area, both aimed at evaluating experience with a product/artifact focus, targeted at all types of users, and application-independent. However, the first one understands UX as usability, desirability, credibility, aesthetics, technical adequacy and usefulness, while the second one considers attractiveness of the product, ease of use and utility.

A similar scenario occurs for specific UX qualities, such as emotion, the most frequently evaluated UX quality, according to our results. For measuring emotion, [84] examines levels of desire, surprise, inspiration, amusement, admiration and satisfaction, while [85] measures valence, arousal and engagement, and [86] analyses anger, fear, happiness, and sadness. Although one may argue that they were constructed under the assumptions of different theoretical roots, often the reasoning behind the instruments psychometrics is not explicit. Consequently, perhaps the evaluator - specially in case of professionals - is not even aware of these differences when choosing a UX evaluation instrument or method.

However, there are some UX qualities that seem to be more established in the literature, such as affect. Most of the instruments that measure affect are based on or adapted from the PAD scale [87] and PANAS scale [88], evaluating a commonly defined set of aspects that describe affect. In this context, it is important to highlight that instruments with good psychometric properties in one culture may not have the same properties when translated for another culture, hence the relations between UX components have to be validated [89].

Although the concept of UX still needs to be better established, the commitment of researchers and practitioners in investigating definitions and improving the understanding of UX factors has been very constructive to the community. The joint efforts to develop effective evaluation methods have been resulting in a variety of instruments

for diverse application domains and groups of users. Also, psychophysiological measures like [64] provide the opportunity to cross self-reported and observation measures with psychophysiological information, enriching data. This type of instrument was the second most frequently found in our cataloging, which seems to be an indication that HCI community is following the UX research agenda proposed by Law and Van Schaik [89].

6 Conclusion

This work presented an analysis and compilation of a variety of types of UX evaluation instruments and qualities, providing researchers and practitioners with a systematized catalog of UX instruments. Although this research has some limitations previously discussed, our goal is to help supporting researchers and professionals in making informed decisions about the choice of instruments for UX evaluation in their every day work. We also shared some insights and concerns about the directions research on UX evaluation has been taking, that we expect to inspire the community. Our future work include expanding the collection and analysis of UX evaluation instruments, comprise more categories of analysis, types of applications and target users, and developing an interactive version of the catalog presented in this paper.

References

1. Obrist, M., Law, E., Väänänen-Vainio-Mattila, K., Roto, V., Vermeeren, A., Kuutti, K.: UX research: what theoretical roots do we build on–if any? In: CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp. 165–168. ACM (2011)
2. Law, E.L.C., Roto, V., Hassenzahl, M., Vermeeren, A.P., Kort, J.: Understanding, scoping and defining user experience: a survey approach. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 719–728. ACM (2009)
3. Ardito, C., Buono, P., Caivano, D., Costabile, M.F., Lanzilotti, R.: Investigating and promoting UX practice in industry: an experimental study. Int. J. Hum.-Comput. Stud. **72**, 542–551 (2014)
4. Pucillo, F., Cascini, G.: A framework for user experience, needs and affordances. Des. Stud. **35**(2), 160–179 (2014)
5. Wright, P., McCarthy, J., Meekison, L.: Making sense of experience. In: Blythe, M., Monk, A. (eds.) Funology 2, pp. 315–330. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-68213-6_20
6. Obrist, M., Roto, V., Väänänen-Vainio-Mattila, K.: User experience evaluation: do you know which method to use? In: CHI 2009 Extended Abstracts on Human Factors in Computing Systems, pp. 2763–2766. ACM (2009)
7. Lallemand, C., Koenig, V.: How could an intranet be like a friend to me?: why standardized UX scales don't always fit. In: Proceedings of the European Conference on Cognitive Ergonomics 2017, pp. 9–16. ACM (2017)
8. Rajeshkumar, S., Omar, R., Mahmud, M.: Taxonomies of user experience (UX) evaluation methods. In: 2013 International Conference on Research and Innovation in Information Systems (ICRIIS), pp. 533–538. IEEE (2013)

9. Bevan, N.: Classifying and selecting UX and usability measures. In: International Workshop on Meaningful Measures: Valid Useful User Experience Measurement, vol. 11, pp. 13–18 (2010)
10. Obrist, M., Roto, V., Vermeeren, A., Väänänen-Vainio-Mattila, K., Law, E.L.C., Kuutti, K.: In search of theoretical foundations for UX research and practice. In: CHI 2012 Extended Abstracts on Human Factors in Computing Systems, pp. 1979–1984. ACM (2012)
11. Hellweger, S., Wang, X.: What is user experience really: towards a UX conceptual framework (2015). arXiv preprint arXiv:1503.01850
12. Rajanen, D., et al.: UX professionals' definitions of usability and UX – a comparison between Turkey, Finland, Denmark, France and Malaysia. In: Bernhaupt, R., Dalvi, G., Joshi, A., K. Balkrishan, D., O'Neill, J., Winckler, M. (eds.) INTERACT 2017. LNCS, vol. 10516, pp. 218–239. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68059-0_14
13. de Normalisation, O.I.: ISO 9241-11 (2018). Saatavissa: https://www.iso.org/obp/ui/#iso:std:iso,9241(11)
14. Nielsen, J.: Usability Engineering. Elsevier, San Diego (1994)
15. Bargas-Avila, J.A., Hornbæk, K.: 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 System, pp. 2689–2698. ACM (2011)
16. Schubert, E.: Measuring emotion continuously: validity and reliability of the two-dimensional emotion-space. Aust. J. Psychol. **51**(3), 154–165 (1999)
17. Zaman, B.: Introducing a pairwise comparison scale for UX evaluations with preschoolers. In: Gross, T., et al. (eds.) INTERACT 2009. LNCS, vol. 5727, pp. 634–637. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03658-3_68
18. Law, E.L.C., Van Schaik, P., Roto, V.: Attitudes towards user experience (UX) measurement. Int. J. Hum. Comput. Stud. **72**(6), 526–541 (2014)
19. Vissers, J., De Bot, L., Zaman, B.: MemoLine: evaluating long-term UX with children. In: Proceedings of the 12th International Conference on Interaction Design and Children, pp. 285–288. ACM (2013)
20. Väänänen-Vainio-Mattila, K., Roto, V., Hassenzahl, M.: Towards practical user experience evaluation methods. In: Meaningful Measures: Valid Useful User Experience Measurement (VUUM), pp. 19–22 (2008)
21. Law, E.L.C.: The measurability and predictability of user experience. In: Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems, pp. 1–10. ACM (2011)
22. Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, p. 38. ACM (2014)
23. Bevan, N., Macleod, M.: Usability measurement in context. Behav. Inf. Technol. **13**(1–2), 132–145 (1994)
24. Agarwal, A., Andrew M.: Beyond usability: evaluating emotional response as an integral part of the user experience. In: CHI 2009 Extended Abstracts on Human Factors in Computing Systems. ACM (2009)
25. Roto, V., Marianna O., Väänänen-Vainio-Mattila, K.: User experience evaluation methods in academic and industrial contexts. In: Proceedings of the Workshop UXEM, vol. 9 (2009)
26. Hassenzahl, M.: The interplay of beauty, goodness, and usability in interactive products. Hum. Comput. Interact. **19**(4), 319–349 (2004)
27. Ganglbauer, E., et al.: Applying psychophysiological methods for measuring user experience: possibilities, challenges and feasibility. In: Workshop on User Experience Evaluation Methods in Product Development (2009)

28. Ng, Y., Khong, C.W., Nathan, R.J.: Evaluating affective user-centered design of video games using qualitative methods. Int. J. Comput. Games Technol. **2018** (2018)
29. Hung, Y.-H., Parsons, P.: Affective engagement for communicative visualization: quick and easy evaluation using survey instruments. In: Visualization for Communication (VisComm) 2018 (2018, to appear)
30. Vermeeren, A.P., et al.: User experience evaluation methods: current state and development needs. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, pp. 521–530. ACM (2010)
31. Poels, K., De Kort, Y. A.W., IJsselsteijn, W.A.: D3. 3: Game experience questionnaire: development of a self-report measure to assess the psychological impact of digital games (2007)
32. Karapanos, E., Martens, J.B.O.S., Hassenzahl, M.: On the retrospective assessment of users' experiences over time: memory or actuality? In: CHI 2010 Physiological User Interaction Workshop, Atlanta, GA, 10–15 April 2010, pp. 4075–4080. Association for Computing Machinery, Inc. (2010)
33. Ryan, R.M.: Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. J. Pers. Soc. Psychol. **43**(3), 450 (1982)
34. Almeida, R., Darin, T., Andrade, R., de Araújo, I.: Towards developing a practical tool to assist UX evaluation in the IoT scenario. In: Anais Estendidos do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web, pp. 91–95. SBC (2018)
35. Granato, M., Gadia, D., Maggiorini, D., Ripamonti, L.A.: Software and hardware setup for emotion recognition during video game fruition. In: Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good, pp. 19–24. ACM (2018)
36. Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., Stone, A.A.: A survey method for characterizing daily life experience: the day reconstruction method. Science **306**(5702), 1776–1780 (2004)
37. Gomez, R.E.: The evolving emotional experience with portable interactive devices. Doctoral dissertation, Queensland University of Technology (2012)
38. Russell, J.A., Weiss, A., Mendelsohn, G.A.: Affect grid: a single-item scale of pleasure and arousal. J. Pers. Soc. Psychol. **57**(3), 493–502 (1989)
39. Laviea, T., Tractinsky, N.: Assessing dimensions of perceived visual aesthetics of web sites. Int. J. Hum Comput Stud. **60**(3), 269–298 (2004)
40. Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994)
41. Scherer, K.R.: Appraisal considered as a process of multilevel sequential checking. Apprais. Process. Emotion Theory Methods Res. **92**(120), 57 (2001)
42. Schaefer, K.: The perception and measurement of human-robot trust (2013)
43. Van Damme, K., et al.: 360° Video journalism: experimental study on the effect of immersion on news experience and distant suffering. J. Stud. 1–24 (2018)
44. Meijman, T., et al.: The measurement of perceived effort. In: Contemporary Ergonomics, pp. 242–246 (1986)
45. Väänänen-Vainio-Mattila, K., Segerstähl, K.: A tool for evaluating service user experience (ServUX): development of a modular questionnaire. In: User Experience Evaluation Methods, UXEM 2009 Workshop at Interact (2009)
46. Lazar, J., Feng, J.H., Hochheiser, H.: Research Methods in Human-Computer Interaction. Morgan Kaufmann, Burlington (2017)
47. Dillman, D.: Mail and Internet Surveys: The Tailored Design Method. Wiley, New York (2000)

48. Väätäjä, H., Koponen, T., Roto, V.: Developing practical tools for user experience evaluation: a case from mobile news journalism. In: European Conference on Cognitive Ergonomics: Designing Beyond the Product—Understanding Activity and User Experience in Ubiquitous Environments, p. 23 (2009)
49. Sauro, J.: SUPR-Q: a comprehensive measure of the quality of the website user experience. J. Usability Stud. **10**(2), 68–86 (2015)
50. Sales, A., Reis, L., Araújo, T., Aguiar, Y.: Tutaform: a multimedia form for Brazilian deaf users. In: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, pp. 269–276. ACM (2018)
51. Kirakowski, J.: The software usability measurement inventory: background and usage. In: Usability Evaluation in Industry, pp. 169–178 (1996)
52. Lessiter, J., Freeman, J., Keogh, E., Davidoff, J.: A cross-media presence questionnaire: the ITC-Sense of Presence Inventory. Presence Teleoper. Virtual Environ. **10**(3), 282–297 (2001)
53. Bernhaupt, R., Pirker, M.: Methodological challenges of UX evaluation in the living room: developing the IPTV-UX questionnaire. In: PUX 2011 Program Committee, p. 51 (2011)
54. Calvillo-Gámez, E.H., Cairns, P., Cox, A.L.: Assessing the core elements of the gaming experience. In: Bernhaupt, R. (ed.) Game User Experience Evaluation. HIS, pp. 37–62. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15985-0_3
55. Saadé, R., Bahli, B.: The impact of cognitive absorption on perceived usefulness and perceived ease of use in online learning: an extension of the technology acceptance model. Inf. Manag. **42**(2), 317–327 (2005)
56. Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.G.: Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int. J. Aviat. Psychol. **3**(3), 203–220 (1993)
57. Witmer, B.G., Singer, M.J.: Measuring presence in virtual environments: a presence questionnaire. Presence **7**(3), 225–240 (1998)
58. Lewis, J.R., Utesch, B. S., Maher, D.E.: UMUX-LITE: when there's no time for the SUS. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2099–2102. ACM (2013)
59. Sim, G., Horton, M.: Investigating children's opinions of games: fun toolkit vs. this or that. In: Proceedings of the 11th International Conference on Interaction Design and Children, pp. 70–77. ACM (2012)
60. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (ed.) USAB 2008. LNCS, vol. 5298, pp. 63–76. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89350-9_6
61. Zaman, B.: Introduction and validation of a pairwise comparison scale for UX evaluations and benchmarking with preschoolers. In: INTERACT 2009. Springer, Uppsala (2009)
62. Horn, D., Salvendy, G.: Product creativity: conceptual model, measurement and characteristics. Theor. Issues Ergon. Sci. **7**(4), 395–412 (2006)
63. Lasa, G., Justel, D., Gonzalez, I., Iriarte, I., Val, E.: Next generation of tools for industry to evaluate the user emotional perception: the biometric-based multimethod tools. Des. J. **20** (sup1), S2771–S2777 (2017)
64. Stahl, A., Hook, K., Svensson, M., Taylor, A.S., Combetto, M.: Experiencing the affective diary. Pers. Ubiquit. Comput. **13**(5), 365–378 (2009)
65. Robinson, R.B.: All the feels: a twitch overlay that displays streamers' biometrics to spectators. Dissertation, UC Santa Cruz (2018)
66. Hussain, J., et al.: A multimodal deep log-based user experience (UX) platform for UX evaluation. Sensors **18**(5), 1622 (2018)

67. Kujala, S., et al.: UX curve: a method for evaluating long-term user experience. Interact. Comput. **23**(5), 473–483 (2011)
68. Broekens, J., Brinkman, W.P.: AffectButton: a method for reliable and valid affective self-report. Int. J. Hum.-Comput. Stud. **71**(6), 641–667 (2013)
69. Nagel, F., et al.: EMuJoy: software for continuous measurement of perceived emotions in music. Behav. Res. Methods **39**(2), 283–290 (2007)
70. Ntoa, S., Margetis, G., Antona, M., Stephanidis, C.: UXAmI observer: an automated user experience evaluation tool for ambient intelligence environments. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) IntelliSys 2018. AISC, vol. 868, pp. 1350–1370. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-01054-6_94
71. Betella, A., Verschure, P.F.M.J.: The affective slider: a digital self-assessment scale for the measurement of human emotions. PLoS ONE **11**, 2 (2016)
72. Kim, J.H., et al.: Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 443–452. ACM (2008)
73. Moura, D., El-Nasr, M.S., Shaw, C.D.: Visualizing and understanding players' behavior in video games: discovering patterns and supporting aggregation and comparison. In: Proceedings of the 2011 ACM SIGGRAPH Symposium on Video Games, pp. 11–15. ACM (2011)
74. Schütte, S., Almagro, L.M.: Development and application of online tools for kansei engineering evaluations. In: KEER2018, Go Green with Emotion. 7th International Conference on Kansei Engineering & Emotion Research 2018, Kuching, Malaysia, 19–22 Mar 2018, no. 146, pp. 20–28. Linköping University Electronic Press (2018)
75. Ayzenberg, Y., Hernandez Rivera, J., Picard, R.: FEEL. In: Proceedings of the 2012 ACM Annual Conference Extended Abstracts on Human Factors in Computing Systems Extended Abstracts - CHI EA 2012 (2012)
76. Mendes, M.S., Furtado, E.S.: UUX-posts. In: Proceedings of the 8th Latin American Conference on Human-Computer Interaction – CLIHC 2017 (2017)
77. Drachen, A., Canossa, A.: Analyzing spatial user behavior in computer games using geographic information systems. In: Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era on – MindTrek 2009 (2009)
78. Cavalcante, E., Rivero, L., Conte, T.: MAX: a method for evaluating the post-use user eXperience through cards and a board (2015)
79. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Mensch & Computer 2003, pp. 187–196. Vieweg+Teubner Verlag (2003)
80. Moreira, E.A., dos Reis, J.C., Baranauskas, M.C.C.: TangiSAM. In: Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems - IHC (2017)
81. Law, E., Roto, V., Vermeeren, A.P., Kort, J., Hassenzahl, M.: Towards a shared definition of user experience. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems, pp. 2395–2398. ACM (2008)
82. Lasa, G., Justel, D., Retegi, A.: Eyeface: a new multimethod tool to evaluate the perception of conceptual user experiences. Comput. Hum. Behav. **52**, 359–363 (2015)
83. Veeneklaas, J.N.: VisUX: a framework of user experience within data visualizations. MS thesis (2018)
84. Desmet, P.: Measuring emotion: development and application of an instrument to measure emotional responses to products. In: Blythe, M.A., Overbeeke, K., Monk, A.F., Wright, P.C. (eds.) Funology, pp. 111–123. Springer, Dordrecht (2003)

85. McDuff, D., Karlson, A., Kapoor, A., Roseway, A., Czerwinski, M.: AffectAura: an intelligent system for emotional memory. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 849–858. ACM (2012)
86. Isbister, K., Höök, K., Laaksolahti, J., Sharp, M.: The sensual evaluation instrument: developing a trans-cultural self-report measure of affect. Int. J. Hum.-Comput. Stud. **65**(4), 315–328 (2007)
87. Mehrabian, A.: Framework for a comprehensive description and measurement of emotional states. Genet. Soc. Gen. Psychol. Monogr. (1995)
88. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: the PANAS scales. J. Pers. Soc. Psychol. **54**(6), 1063 (1998)
89. Law, E.L.C., Van Schaik, P.: Modelling user experience—an agenda for research and practice. Interact. Comput. **22**(5), 313–322 (2010)
90. Andersen, E., Liu, Y.E., Apter, E., Boucher-Genesse, F., Popović, Z.: Gameplay analysis through state projection. In: Proceedings of the Fifth International Conference on the Foundations of Digital Games, pp. 1–8. ACM (2010)