Attitudes towards User Experience (UX) Measurement

Effie Lai-Chong Law[1], Paul van Schaik[2], Virpi Roto[3]

[1]Department of Computer Science, University of Leicester, UK

Telephone: +44 116 252 5341

Email: elaw@mcs.le.ac.uk

[2] School of Psychology, Teesside University, UK

[3] School of Arts, Design and Architecture, Aalto University, Finland

ABSTRACT

User Experience (UX), as a recently established research area, is still haunted by the challenges of defining the scope of UX in general and operationalising experiential qualities in particular. To explore the basic question whether UX constructs are measurable, we conducted semi-structured interviews with ten UX researchers from academia and one UX practitioner from industry where a set of questions in relation to UX measurement were explored (Study 1). The interviewees expressed scepticism as well as ambivalence towards UX measures and shared anecdotes related to such measures in different contexts. Interestingly, the results suggested that design-oriented UX professionals tended to be sceptical about UX measurement. To examine whether such an attitude prevailed in the HCI community, we conducted a survey - *UX Measurement Attitudes Survey* (UXMAS) - with essentially the same set of 13 questions used in the interviews (Study 2). Specifically, participants were asked to rate a set of five statements to assess their attitude towards UX measurement, to identify (non)measurable experiential qualities with justifications, and to discuss the topic from the theoretical, methodological and practical perspective. The survey was implemented in a paper-based and an online format. Altogether, 367 responses were received; 170 of them were valid and analysed. The survey provided empirical evidence on this issue as a baseline for progress in UX measurement. Overall, the survey results indicated that the attitude towards UX measurement was more positive than that identified in the interviews, and there were nuanced views on details of UX

measurement. Implications for enhancing the acceptance of UX measures and the interplay between UX evaluation and system development are drawn: UX modelling grounded in theories to link experiential qualities with outcomes; the development of UX measurement tools with good measurement properties, and education within the HCI community to disseminate validated models, and measurement tools as well as their successful applications. Mutual recognition of the value of objective measures and subjective accounts of user experience can enhance the maturity of this area.

1. INTRODUCTION

The exploration of the issue of user experience (UX) measurement was embarked on (e.g. Law 2011) after another, if not more, thorny issue of UX - its multiple definitions - had been examined (Law et al. 2009). In principle these two foundational issues should be solved in tandem. The recent efforts of deepening the understanding of the theoretical roots of UX (e.g. Obrist et al. 2011) can complement the earlier work on UX evaluation methods on the one hand (Vermeeren et al. 2010) and the current operatonalisation work for UX measurement on the other hand (e.g. van Schaik, Hassenzahl & Ling 2012).

The field of HCI in which UX is rooted has inherited theoretical concepts, epistemological assumptions and methodologies from a diversity of disciplines, ranging from engineering where measures are strongly embraced (cf. William Thomson's dictum "to measure is to know") to humanities where measures can be regarded as naïve or over-simplistic, especially when the concepts to be measured are ill-defined, leaving (too) much for interpretation (Bartholomew 2006). As UX subsumes a range of fuzzy experiential qualities (EQs) such as happiness, disgust, surprise and love, to name just a few, controversies and doubts about the measurability of UX are inevitable.

The literature on UX published since the turn of the millennium indicates that there are two disparate stances on how UX should be studied (i.e. qualitative versus quantitative) and that they are not necessarily compatible or can even be antagonistic. A major argument between the two positions is the legitimacy of breaking down EQs into components, rendering them to be measured. This tension is rooted in the age-old philosophical debate on reductionism versus holism. Indeed, a rather comprehensive review on the recent

UX publications (Bargas-Avila & Hornbæk 2011) shows that UX research studies have hitherto relied primarily on qualitative methods; the progress on UX measures has thus been slow. There have also been voices in HCI that challenge the need, value and even appropriateness of measuring UX constructs (e.g. Boehner et al. 2007; Forlizzi & Battarbee 2004; Höök 2010; Swallow, Blythe & Wright 2005). However, there is also an emphasis on structural and measurement models of UX (e.g. Law & van Schaik 2010), and on the significance as well as ease of measuring UX constructs, especially for industry (Wixon 2011).

Discussions in formal (e.g. Kaye et al. 2011; Roto et al. 2010) as well as informal settings (e.g. personal communications) suggest that UX professionals who have training in design or whose job is design-oriented tend to be sceptical or ambivalent about UX measurement. To explore whether such an attitude prevails in a wider HCI community has motivated us to conduct a study called ***UX Measurement Attitude Survey*** (UXMAS). To the best of our knowledge, a survey on this specific topic has never been conducted. Findings of the survey can validate the ostensible assumption that the HCI community is convinced about the *plausibility, necessity* and *utility* of UX measurement. In examining various stances on UX measures, some fundamental theoretical, methodological and practical issues hindering the progress of UX can be revealed. Insights, so gained, can refine and substantiate the work agenda of this emerging research area, which remains challenged by a list of thorny issues. Specifically, how HCI researchers and practitioners perceive the interplay between UX measures and the design and development of an interactive system is a focus of our work on UXMAS.

In summary, by studying the prevailing attitudes towards UX measurement with the tool UXMAS, which is the first survey on this topic highly relevant to the growing UX research, we aim to stimulate the HCI community to discuss UX measurement from different perspectives. Furthermore, results of our empirical studies can lead to a validated tool to assess attitude and behaviour on UX measures, thereby enhancing the acceptance of UX measures as well as their impacts on system development.

The structure of this paper is as follows. First, we present the related work, especially the debates over UX measures from the established measurement theories as well as contemporary views of UX professionals in

Section 2.1. Then we describe a review study on the recent empirical research work on UX measures in Section 2.2. Next, we present the design, implementation and results of UXMAS in Section 3, 4 and 5, respectively. Last, we conclude and draw implications for our future work in Section 6.

2. RELATED WORK

2.1 Overview on the debates over UX measures

A caveat should be issued that the limited space here does not do any justice at all to the enormously long and rich history of measurement, which can be traced back to the 17th and late 19th century for physical sciences and social sciences, respectively. Big volumes on measurement have been published (e.g. three volumes of *Foundations of Measurement* 1971-1990, Academic Press, cited in Hand (2004); four volumes of *Measurement;* Bartholomew 2006). Great scholars include William Thomson (Lord Kelvin), who established some major measurements in engineering and held a tremendously firm stance on the role of measurement in science, and S.S Stevens (1946), who developed the theory of scale types and imparted strong influences on measurement in social sciences such as intelligence tests. While these and other volumes argue for and show the indispensability of measurement, there is no lack of counter-arguments, based on socio-political, epistemological and other grounds (Bartholomew 2006). It is beyond the scope of this paper to delve thoroughly into the related histories. Instead, we highlight arguments that can help understand attitudes towards UX measures.

In this study, we adopt Hand's (2004, p.3) definition of measurement "*quantification: the assignment of numbers to represent the magnitude of attributes of a system we are studying or which we wish to describe*." We also augment Thomson's classic claim by stating that if you cannot *interpret* what is measured, you cannot improve it. Arguably one can measure (almost) anything in some arbitrary way. The compelling concern is whether the measure is *meaningful*, *useful* and *valid* to reflect the state or nature of the object or event in question. However, this concern is also applicable to the three well-established usability metrics – effectiveness, efficiency and satisfaction (ISO 9241; ISO 25010). While they have been widely adopted in usability research and practice, their impact on the system development process is not generally

recognized. How these measures are actually defined, taken and used can vary largely with contexts and the relationships among them remain unclear, rendering a usability summary measure disputable (Sauro & Lewis 2009). These issues have triggered much discussion from the late 1990s to mid-2000s (e.g. Hornbæk 2006) when the shift of emphasis to UX has visibly begun, though the debates on usability methods and measures remain (e.g. Hornbæk & Law 2007).

Given that UX has at least to some extent developed from usability, it is not surprising that UX methods and measures are largely drawn from usability (Tullis & Albert 2008). However, the notion of UX is much more complex, given a mesh of psychological, social and physiological concepts it can be associated with. Among others, the major concept is emotion or feeling (McCarthy & Wright 2004). Dated back to more than a century ago, the James-Lange Theory of Emotion (see review in Lang 1994) was developed to explicate the intricate relationships between human perception, action and cognition. Accordingly, emotion arises from our conscious cognitive interpretations of perceptual-sensory responses; UX can thus be seen as a cognitive process that can be modelled and measured (Hartmann, De Angeli & Sutcliffe 2008). Larsen and Fredrickson (1999) discussed measurement issues in emotion research with reference to the influential work of Ekman, Russell, Scherer and other scholars in this area. More recent work along this direction has been conducted (cited in Bargas-Avilas & Hornbæk 2011). These publications point to a common observation that measuring emotion is plausible, useful, and necessary. However, like most, if not all, psychological measurements, they are only approximations (Hand 2004) and should be considered critically. This reservation can be reflected in Kahneman's (2011) debatable statement: "*Many psychological phenomena can be demonstrated experimentally, but few can actually be measured*" (p. 123). Interestingly, Kahneman has involved in the work on measuring well-being since the 90s. Another rather ambivalent attitude towards UX measurement is reported in Roto and colleagues (2010): "*No generally accepted overall measure of UX exists, but UX can be made assessable in many different ways*." (p. 8).

UX researchers may roughly be divided into two camps, which can be named as "design-based UX research camp" and "model-based UX research camp" (Law 2011). The main cause for the tension between the two

camps in UX is their disparate appreciation towards the approaches that emphasize representing user experience in a certain, comparable and generalizable way and those that emphasize articulating rich embodied experiences with contexts (Boehner et al. 2007).

For instance, Forlizzi and Battarbee (2004) argued that: "... *emotional responses are hard to understand, let al.one quantify*." (p. 265). Similarly, Swallow and colleagues (2005) remarked that: *"... such approaches may be useful for experimental analysis but they can miss some of the insights available in accounts that resist such reduction ... qualitative data provides a richness and detail that may be absent from quantitative measures*." (pp. 91-92). In rebutting these stances, Hassenzahl (2008) argued that the uniqueness and variation of experiences with technology is much less than it is implied by the phenomenological approach. Tractinsky (in Roto et al. 2010, p.25) asserted that as a complex construct UX should be studied with scientific methods and that it is necessary to develop measures and measurement instruments to test and improve UX theories, which should eventually help in designing interactive systems for various experiences in different contexts. In contrast, some explicit statements against measurements and reductionism were voiced by Höök (ibid): "*The question is whether measuring the end-user experience as a few simplistic measurable variables is really helping us to do better design or to better understand the user experience. In my view, there are too many reductionists out there who harm research in this area by pretending that we can provide measurements and methods that will allow anyone to assess the UX-'value' of a designed system*" (p.17). Whether this pessimistic view on UX measurement is commonly shared by UX design researchers has been examined in this study.

Another cause of tension is the difference between industrial and academic needs such as instantly useful data for product development as opposed to meticulously analysed data for theory-building (Kaye et al. 2011). Norman claimed in a recent interview (2008): "*There is a huge need for UX professionals to consider their audience… We should learn to speak the language of business, including using numbers to sell our ideas*." Numbers of some sort are deemed useful, primarily because of their brevity and accessibility. A caveat is that such usefulness is contingent on *who* uses the measures for *what purpose* – a major concern for

understanding the interplay between UX evaluation and system development. Norman's advocacy is directed at top management executives who need to make (critical) decision on design and development issues within a (very) short period of time. While Norman puts emphasis on the plausibility of measures to convince the managerial staff, the validity of measures seems not of his major concern.

We explore the above views with reference to the empirical data gathered for this study. In particular, the aim of our study is to examine in detail the HCI community's attitude towards UX measurement. Results of analysing the arguments for and against UX measurement may inspire people to develop ideas as well as strategies to improve its quality, credibility and thus acceptance.

2.2 Review on publications on user experience measures

2.2.1 Method

With the goal to identify which and how UX constructs were measured in the recent UX research studies, we conducted a review by adapting the research protocol designed by Bargas-Avila and Hornbæk (2011; henceforth BAH), who systematically reviewed 51 publications on UX from 2005-2009. Several intriguing results were reported by BAH: (i) the methodologies used are mostly qualitative and commonly employed in traditional usability studies, especially questionnaires and scales; (ii) among others, emotions, enjoyment, and aesthetics are the most frequently measured dimensions; (iii) the products and use contexts studied are shifted from work to leisure and from controlled tasks to consumer products and art.

In comparison, the scope of our review was narrower than BAH's. The timeframe was also different. As BAH had already carried out a thorough review on the studies from 2005-2009, we focused on those from last three years, 2010-2012. Specifically, we followed the procedure described in BAH, searching the three scientific repositories: ACM Digital Library (DL), ISI Web of Knowledge (WoK), and ScienceDirect (ScD). However, the search words we used were "user experience" *and* "measure". In DL and ScD, we used *Advanced Search* to restrict the search within the three fields: *Title, Abstract*, and *Keywords*. In logical terms, the search is expressed as follows:

(Title:"user experience" OR Abstract:"user experience" OR Keywords:"user experience) AND (Title:measure OR Abstract:measure OR Keywords:measure)

The search returned 117 and 89 in DL and ScD, respectively. In WoK, as no search restriction in this way is enabled, the search was performed within 'all fields' and returned 310. We checked for duplicates among the search results of the three sources and eliminated them. Next, we applied the screening process as described in BAH (p. 2691). We included publications that are original full papers (thereby filtering out workshop papers, posters and non-refereed press articles), which speak in a broad sense of interactions between users and products/services and report primary empirical user data (i.e. reviews such as Hassenzahl et al., 2012 are not included). We also excluded out-of-scope papers addressing topics like telecommunications networks. However, we did not apply the criterion that publications should cite at least one of the authors who are deemed by BAH as 'key to UX research', because we find the list somewhat arbitrary.

A full list of **58** publications used for further analysis is referenced in a webpage[1]. A caveat is mentioned that our review is *not* meant to be an extension of BAH, because we have *not* replicated their approach in an exact manner and our goal was also different from theirs.

2.2.2 Measured UX constructs

For each of the 58 selected studies, we extracted information that was relevant to our goal of knowing which and how UX constructs had been measured. Furthermore, to examine the issue of interplay between UX evaluation and system development, we aimed to identify whether and how the UX measures were used by developers or designers.

All these studies measured UX in addition to other cognitive (e.g. learning efficacy for a speed-reading task; Mumm & Mutlu, 2011) and behavioural (e.g. task completion time) constructs. Eleven of the studies measured only one single UX construct (e.g. aesthetics, fun, enjoyability) or unspecified emotions/affects

[1] http://www.le.ac.uk/compsci/people/elaw/i-uxsed-references

(in this case we classified it as 'general' see Table 1). The number of UX constructs measured in a study ranged from one to fourteen (cf. Flavián-Blanco et al. 2012 measured different sets of experiential quality before, during and after interactions). Altogether 42 unique UX constructs were measured by the selected studies. Table 1 shows the twelve constructs with frequency higher than two. In contrast to BAH's observation, it seems that the multi-dimensional UX measurement is not uncommon. For instance, *flow*, the most commonly measured UX construct, could be assessed psychometrically along nine dimensions (van Schaik & Ling 2012a); *emotion* was measured along three (i.e. visceral, behavioural and reflective derived from Norman's 2004 work; Park et al. 2011) or six basic emotions identified by Paul Ekman. Unexpectedly, *frustration*, which is often measured in usability studies, was addressed by only one study.

*** Insert Table 1. UX Constructs measured in the recent empirical studies ***

All the 58 studies used questionnaires or scales, be they validated (e.g. AttrakDiff, Self-assessment Manikin, Game Experience Questionnaire, Flow State Scales, PANAS) or home-grown, to measure the constructs of interest; this observation can be corroborated by BAH. In five studies psycho-physiological measures such as heart rate, skin conductance and EEG, were taken and calibrated with self-reported measures. An interesting study aimed to correlate keystroke patterns with confidence, hesitance, nervousness, relaxation, sadness, and tiredness (Epp et al., 2011). Two of the studies (Olsson et al., 2012; Karapanos et al., 2010) analysed experience narratives to derive some quantitative measures of emotions. With regard to context of use, 16 of the selected studies were on video games, 2 on movies, 8 on mobile phones, 8 on specific applications (e.g. a speed-reading widget), and 22 on general products/services such as website homepages and e-commerce. This observation aligns with BAH's conclusion that the UX research tended to be conducted in non-work-related contexts.

Furthermore, of particular relevance to the interplay between user evaluation and system development is how the UX measures were or would be handled in the selected studies. Surprisingly, none of the studies report whether and how the UX measures have *actually* been used in the next cycle of the system development; therefore, the 'downstream utility' of the UX measures remains unknown. Nonetheless, 43 of

the studies described, although to various levels of elaboration (9 high, 20 moderate, and 14 low), how the UX measures could be used by developers or designers for improving the products, whereas 15 of the studies did not mention anything in this regard. This might be explained by the fact that most of the selected studies were academic research work for model validation as well as understanding the phenomenon of UX rather than industrial case studies. Another rather surprising observation is that 16 of the studies did not address the psychometric property of the measurement tools used, which are normally close-ended questionnaires. The other 42 discussed the issues of reliability and validity with three of them analysing the methodological issues on measurement in depth (Karapanos et al., 2012; Procci et al., 2012; van Schaik & Ling, 2012b).

In summary, the above review aims to illustrate the current state-of-the-art of UX measurement in practice. These behaviour-based data can be used to complement the findings about the attitudes of the HCI community towards UX measurement as gauged by our surveys, which are described in the following sections.

3. METHOD

3.1 Overview

A survey called *User Experience Measurement Attitude Survey* (**UXMAS**) was created and deployed in three different contexts:

- *Interview*: 11 interviews were conducted on an individual basis between October and November 2011. Participants were recruited via email invitations in a research institute in Finland and also via personal contacts of the first author.
- *Paper-based survey*: It was distributed to the participants of a one-day seminar on UX hosted by the SIGCHI Finland in October 2011. Out of approximately 100 participants 35 returned the completed survey.
- *Online survey*: It was widely distributed to relevant communities via mailing list, including SIGCHI, BCS-HCI, NordiCHI, some local UXPA (User Experience Professional Association) chapters and related research groups (e.g. TwinTide; allaboutux). Personal invitations were also sent to UX researchers known to the authors. The survey was launched between June and August 2012 and attracted 332 visits, but only 135 responses were useful for further analysis.

All participations were voluntary with no tangible reward.

3.2 Design of UXMAS

UXMAS consists of 13 questions grouped into three main parts. Part A comprises five background questions (Table 2).

*** Insert Table 2**.** Background questions ***

Part B comprises five questions on the measurability of UX qualities (Table 3). The purpose of Q6 is to understand if participants' interpretations align with any of the existing definitions of measurement. For Q7, the rationale underpinning each statement varies. The first one was derived from the classic justification for measurement advocated by Thomson (1891). The second and third ones were two rather extreme views against UX measures expressed in some informal contexts (e.g. group discussion in a

workshop). They were aimed to stimulate thoughts and should not be treated as scientific statements. The fourth and fifth statements represent views on the potential uses of UX measures. They were deliberately broad in scope to stimulate discussions.

The notion of **experiential qualities** (EQs) is central for Q8, Q9 and Q10. In the simplest sense, they are referred to as feelings. In the broadest sense, they are related to the concept of emotional responses, as defined in the Components of User Experience (CUE) model (Thüring & Mahkle 2007), which are influenced by instrumental (i.e. usability) and non-instrumental qualities (i.e. aesthetic, symbolic and motivational). We chose the CUE model for analysing experiential qualities, as it constitutes the most-comprehensive model of UX to date and it integrates usability and (other) aspects of UX. While CUE focuses more on evaluation, in the context of design, the notion of EQs is defined as articulations of major qualities in the use of a certain type of digital artefact intended for designers to appropriate in order to develop their own work (Löwgren 2007). To enable open discussion no definition was given to participants.

Part C comprises three questions aimed to simulate in-depth discussion (Table 4). Note that this part was not included in the paper-based survey, given that the time constraint of the event where it was administered. While all the 11 interviewees answered all the three questions of Part C, they were optional for the participants of the online survey.

*** Insert Table 3. Five main questions on UX measures ***

*** Insert Table 4. Questions for in-depth discussion ***

4. STUDY 1: INTERVIEW UXMAS

4.1 Participant and Procedure

An invitation to the interview was circulated by email to relevant research teams in Aalto University in Finland. Eight participants volunteered to take part in it. The other three participants were recruited via personal invitation. Their participations were also voluntary. There were altogether 11 participants, designated as *S*1, *S*2 and so on (NB: to differentiate from Study 2 where participants are coded as *P*). Seven

of them were female and four were male. Five aged between 31 and 40, another five between 41 and 50, and one above 50. All were researchers except *S*5, who was a practitioner. The job of eight of the participants was predominantly design-oriented, be it practical or theoretical, such as empathic design for house renovation, co-design for persuasive games, and design theories. The other three focused more on UX evaluation of interactive products such as mobile phone. Two of them have worked in UX for less than 1 year, three 1-3 years, five 3-5 years and one for than 5 years. All the interviews were primarily conducted on an individual basis by the first author in English. Shortly before the interview a digital copy of the list of the questions was sent to the participants. It was at their discretion how to make use of the list or do nothing with it at all. A printed copy was also available for reference throughout the interview. All the interviews were audio-taped and transcribed subsequently.

4.2 Results and Discussion

4.2.1 Definition of a Measure (Q6)

When participants were asked to describe what a measure is, they addressed the following facets of measures: *purpose* (e.g., comparison, reference), *property* (e.g., quantitative, variable, objective, dimensional, recognizable), *pre-condition* (e.g. definition, criteria), *process* (e.g., observation, judgment), *problem* (e.g. intangible, breaking down into components), and *example* (e.g., temperature, meter, reactions).

4.2.2 Five Statements on UX Measures (Q7)

Given the small sample size, no inferential statistics of the ratings are computed. Justifications for the ratings are more relevant to our understanding of their attitudes; the analyses are presented below.

UX measures lead to increase of knowledge (mean = 4.0, range: 2-5). When prompted to specify which kinds of knowledge would be increased, several were mentioned,

- references against which products can be compared;
- the extent to which the development goals is achieved;
- values to be delivered by certain design methods;

- information helpful for future projects;
- experience per se.

Ambivalence was observed, for instance: "There are ways to get knowledge about UX in a more meaningful way rather than using measures, but I still think that they are important." (*S*6). Besides, the need for including qualitative data as complementary knowledge was emphasized: "We should have both… qualitative is to know what the reason is for user experience and for the related design issue." (*S*8). Furthermore, conditions for benefiting from UX measures were specified: "It requires people using the measure, understand the measure and what it actually means… There might be people who are not trained to use UX measures, no matter how well we define the measures." (*S*5). This observation highlights the need for enhancing education and training in UX.

UX measures are insane (mean = 2.0, range: 1-4). A common view was that the insanity lies not in UX measures but rather in what claims to be made about them, especially when people do not understand such measure, intentionally misuse them, are unaware of their inherent limitations (e.g. incompleteness) or over-formalize them. There were also concerns whether UX measures can explain why people experience something or have any use for design, as remarked by *S*11 (a designer):

> "… for the purpose of design, measuring variables up to a very high degree and intricate level of measurement might not be that purposeful because you have to translate the numbers back to design requirements, and I am not sure whether that works."

UX measures are a pain (mean = 3.27, range: 1-5). Pain inflicted was psychological rather than physical. Reasons for such pain varied with the phase of UX measurement. In the preparation phase, defining valid and meaningful metrics, which entailed deep and wide knowledge of various matters, was cognitively taxing and thus painful. For data collection, participant recruitment and time constraint were a pain for researchers, as illustrated by *S*4's remark: "We would not use half-an-hour to measure something but rather get some qualitative data out of participants." On the other hand, the intrusiveness and lengthiness of the procedure could be pain for users. For data analysis, statistical analysis was deemed challenging by

four participants. This again is a clear implication for the training of UX. Interpretation of UX measures was another common concern: it could be an issue of lack of knowledge, confirmation bias, and attempts to draw implications from exact measures for design.

UX measures are important for design (mean = 4.0, range: 2-5). Participants' stance on this claim was ambivalent. They recognized that UX measures could help identify design constraints and justify design decisions by convincing developers and management, given that numbers could convey a sense of reliability. However, they stipulated the importance of UX measures in design with the need of combining with qualitative data, for instance:

> "I mean they are important, but I'd *not* base my design solely on UX measures... there are lot of things that I don't think that we can measure properly enough yet… it would cause too much work to get really really good measurement that would be our main basis for design… [UX measurement] would only be second; the first being an overall understanding of qualitative views we have found out from users." (*S*4)

> "If UX measures are clusters that are described through numbers or questionnaires, then they are *not* important for design, whereas if UX measures are, for instance, clusters of qualitative data and users' accounts, then they are important for design" (*S*11)

Some participants explicitly expressed their doubt about the role of UX measures in design, for instance:

> "I can see relatively little value of applying UX measures, because they don't really link to the product's attributes in most cases… they link it at an abstract level… it is hard to trace what the underlying causes for certain response. It is almost impossible if we just use UX measures without combining them with qualitative data" (*S*1)

> "They're only important where achieving certain experiences is part of the goal of design… I think goal design is a balance of achieving positive experiences and positive outcomes…I'd say typically in most design settings the outcomes are more important than experience." (*S*9)

Furthermore, one participant pointed out the differences between usability and UX measures:

> "… sometimes it is difficult to explain why we design like this even when we provide evidence. From usability point of view we can more easily give this measurement that it is better, but designing for UX is problematic. People with technical problems have problems making the difference between UI and UX. They think they are the same thing." (*S*3)

UX measures are important for evaluation (mean = 4.6, range: 2-5). On average the participants had a higher level of agreement on this claim and were somewhat less ambivalent. Similar supporting arguments were presented: justifying decisions, validating design goal, and giving reliability (cf. *S*2's remark: "If you only use the designer intuition, only use empathic interpretation, it is not very reliable for the rest of the world"). Some participants pointed out the time issue: in which development phase UX measures are taken and how much time the process of measuring is allowed, for instance:

> "you don't have a good chance for proper measurement …in industry-led cases they are more keen on fast phenomenon… the industrial people want to improve the design but not really want to provide input for the academic world in general" (*S*4)

There are also reservations about the role of UX measures in evaluation, for instance:

> "it's not been proven yet that they can make any difference to outcomes…. I mean, they *could* be; certainly if you include traditional usability measures, then persistent task failure for many designs is going to be something you want to know about. But I don't think they're automatically important; they're all hinges around design objects" (*S*11)

In summary, the interplay between UX measures, which are common evaluation outcomes, and (re)design, as perceived by the design-oriented researchers, is ambiguous.

4.2.3 Measurable and non-measurable experiential qualities (EQs)

The participants were asked to identify experiential qualities (EQ) that were of personal/professional relevance and their respective measurability (Q8), that were (almost) certainly measurable (Q9) and that were (almost) certainly non-measurable (Q10). We adopted and adapted the CUE model (Thüring & Mahlke 2007) (Figure 1) to group the responses elicited (NB: some of which are not EQs) from the three questions into four categories:

- *Instrumental qualities* (INQ) – "the experienced amount of support the system provides and the ease of use. Features, such as the controllability of the system behaviour and, the effectiveness of its functionality, fall into this category." (ibid, p. 916);
- *Non-instrumental qualities* (NIQ) – "the look and feel of the system and other system qualities that are not instrumental" (ibid). Features such as visual aesthetics, haptic quality and motivational qualities;
- *Short-term affective response* (STAR) (cf. experiential qualities) – a user's subjective feeling, motor expression or physiological reaction (Scherer 2005) occurs during or immediately after interacting with a system or a product. It broadens the scope implied by the original notion of 'emotional reactions' (Thüring & Mahlke 2007) to accommodate mildly affective responses with an artefact;
- *Long-term evaluative response* (LTER) (cf. system appraisal) – long-term effect of interacting with the system on a user's attitude, behaviour and cognition;

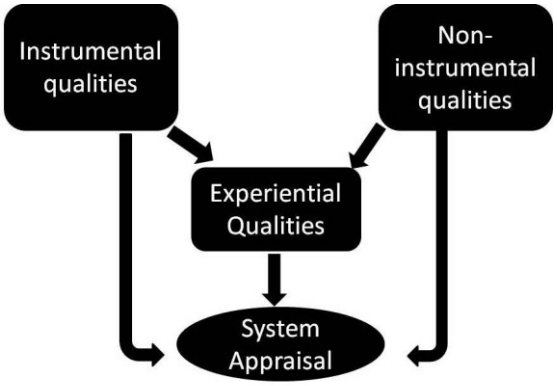

Figure 1: A simplified CUE model

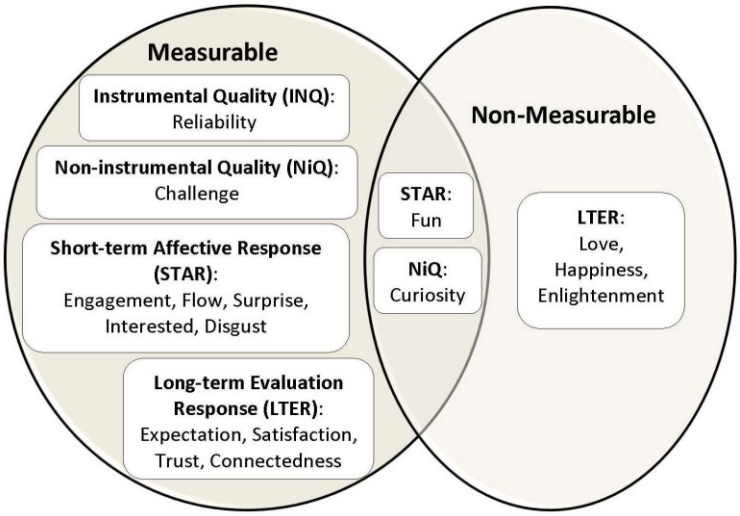

Figure 2. Measurability of qualities and constructs

Several intriguing observations are noted:

i) All three UX constructs considered as non-measurable fall into the category of LTER; it seems implying that long-term effects of interaction are considered not amenable to measurement;

ii) No instrumental and non-instrumental qualities were identified as exclusively non-measurable by the participants; this is not surprising as instrumental qualities are closely related to traditional software attributes that have explicitly been operationalised and operationlising non-instrumental qualities such as aesthetic and symbolic has been endeavoured in recent UX research efforts (e.g. Hassenzahl & Monk, 2010);

iii) Fun is the EQ that was dually considered as measurable as well as non-measurable. This is somewhat surprising because game experiences of which fun is an integral part have been one of the hot topics in UX research where different attempts to measure fun have been undertaken (e.g. Gross & Bongartz 2012). This observation underpinned *S*11's argument for the measurability of fun as it is a well-defined concept. In contrast, *S*1's counterargument referred to the complexity and multidimensionality of fun; reporting on overall fun after interaction seemed more plausible than on individual sub-constructs;

iv) Several high-level constructs were mentioned: 'hedonic quality' for measurability and 'long-term experience' and 'deep [sub]-conscious experience'; they do not fit into any of the categories.

Furthermore, the main argument for measurability is that the EQs of interest are well-defined and documented in the literature. Two participants, however, could not name any certainly measurable EQ because they considered that qualitative data were better for understanding feelings and that experiential concepts were in general fairly vague. In contrast, the key arguments for non-measurability are the epistemological assumption about the nature of certain experiences and lack of a unified agreement on what UX is. The five participants could not name any certainly non-measurable EQ. They, while assuming that everything can be measured, had the reservations for the validity, impact and completeness of UX measures. Specifically, *S*9 pointed out the issue of conflating meaningfulness with relevance:

> "I think anything can be measured in a meaningful way; it depends who the audience is… the issues with measurement … are well understood in the psychometric system whether you are really measuring what you think you are measuring. So, and, again you need to distinguish between meaningfulness and relevance… there are things that are irrelevant … but I don't think it's possible for things in this world to have no meaning… people are natural interpreters."

With regard to the question on how to measure EQ, the participants identified a range of known HCI methods, which can be categorized into three major types: *overt behaviour* (e.g., time-on-task, number of trials to goal); *self-reporting* (e.g. diary, interview, scale); and *psycho-physiological* (e.g. eye-tracking, heart rate). Obstacles for implementing measurement were also mentioned, including various forms of validity, individual differences, cultural factors, confidence in interpreting non-verbal behaviour, translating abstract concepts into concrete design property, and consistency of observed behaviour.

4.2.4 Anecdotal descriptions on the interplay between evaluation and development

In responding to the interview questions, some participants described intriguing cases that can well illustrate the challenges of enhancing the interplay between UX evaluation and system development. Subsequently we highlight the challenges and related anecdotes, which are grouped as theoretical (Q11), methodological (Q12) and practical issues (Q13).

Theoretical issues

- *Problem of measuring UX in a holistic way, and breaking down into components seems not an ideal solution:*

  *S*3: When we go through the issues with uses, we observe the whole expression, their comments on certain issues. If we have a lot of things to study, it is more difficult to run this kind of a holistic study; in a lab test where we only study some specific items. In an evaluation session when we study several issues, we can show users some of them and then the whole one. Holistic approach is the way to go, but measures about some specific details help as well.

*S*4: I'd say UX is holistic in nature, it is difficult to break it down into very small pieces. From the traditional scientific perspective, the way to measure something, to break it down and separate different factors … The value of the measurement gets lower if you break it down to small pieces.

- *Memorized experiences prone to fading and fabrication:*

  *S*5: the actual intensity of the moment fades very fast… So it is interesting to see how to recall and how we change the memory of the experience. When we ask people whether they like something or not it depends on the moment you are asking. iPhone, there is so much positive information of that product out there that even if you did not like it, your environment is so positive about it that you are positive as well. It is the same as with reconstructing the memories. … Most people as well as I myself are sure I have memories where I cannot make a difference between the reconstructed and actual memory.

- *UX measures are highly sensitive to timing and nature of tasks:*

  *S*2: When to measure depends the duration and complexity of the task. For a small task, we can let people complete it and take measures at the end. For the longer one may need to be interrupted….

  *S*8: Different measures in different phases of use complement each other. If you only measure momentary, you just get huge amount of positive and negative experiences, but you cannot know what can we do with design, which ones to address, prioritization is very difficult? Users have feelings up and down all day, what is the point and what to do next, which of those are influential and critical? Then you have to do momentary measures. You have to see what the influential factors are in the long run. … It is difficult to interpret psycho-physiological measures. You don't have exact measures for evaluating emotions at the moment. Very momentary information can be useful, but you also need other measures. Even though you can capture all the momentary emotional measures, you don't know how the user interprets the emotion. ... Psycho-physiological measurements can be useful e.g. in designing games. It would be very useful the exact point when

the person has a challenging or very dull experience. … Mobile phones are used in different contexts; it is difficult to measure the emotions in all of them.

Methodological Issues

- *Different preferences for qualitative and quantitative data by design- and engineering-oriented stakeholders:*

  S7: … we are not fond of measures … we have smart design work, something we have emphasized more on qualitative and inspirational aspect of UX. We have something to do with design perspective; kind of measurement only gives basic constraints and do not give directions. It depends where you apply the methods; how they should be interpreted and position the methods. Measures are good background knowledge but we have more unpredictable, qualitative data.

  *S*8: Qualitative data could cover everything, but then how to convince the engineers, that's why we need numbers. Also for research purpose, it could be interesting to find the relationships between factors. I have to measure somehow to find out which is more influential, hedonic or pragmatic quality, on customer loyalty… quantitative data are more convincing, but developers need qualitative data as well because they want to understand the reason for frustration. … It is important to measure both immediate experience and memorable experience. Practitioners are very thrilled by the idea that you can do it afterwards because it is so easy. So the companies are very interested in long-term UX or this kind of retrospective evaluation, they don't mind that, because they are convinced that memories are very important because they are telling stories to other customers; they are loyal to the companies based on the memories. Only the reviewers are criticising the validity of retrospective methods. Practitioners are very interested in it and like the idea.

  *S*10: You have to interpret psycho-physiological data and map these data to one of these experiential concepts and it is very hard to know whether you get it right. You can have a high heart rate because you really love it or you hate it. So may be it also depends on how many categories you have; the more

categories you have, the more difficult to find a good mapping. I have two UX components, good or bad or positive affect vs. negative affect, maybe it is easier to get it right; you have less chance of making error. But again, does it fit the purpose?

*S*11: To see the impact of the goal of the system, how people perceive it. I think that's fine. For the purpose of design, quantitative measures do not make sense. It is a wrong method for the purpose of design.

- *Resource-demanding evaluation with a large number of heterogeneous users:*

  *S*4: Our perspective is very design-oriented. My experience in measuring UX in design process is not so much. It is so easy and fast to make the participants fill out AttrakDiff, it really would not make sense *not* to do it. How we analyse the results and get out of it, that's still to be seen. We don't have so many participants that we could see what the different ways of using those results are. Like a backup, we get a general understanding of the situation to compare for making the second prototype, what things to change. When we have the second prototype and we use the same measurement, we can see where the design is going. As measurement depending so heavily on individual participants, it is difficult to make conclusion about the measurements… it is hard to say why there is a difference in the results because of different social groups.

- *Need of sophisticated prototypes for eliciting authentic user experiences:*

  *S*7: Difficult, especially housing business … we cannot build only one prototype and then ask people experience it, get feedback and then do it… we need good examples, media we can use to produce our tools, social media, TV, etc to show what kind of solution we might have.. the storytelling method like a movie;

Practical Issues

- *Lack of knowledge in exploiting feedback on UX for future system development:*

*S*5: Most people in industry, whether they have backgrounds in economics, engineers or marketing, for them handling qualitative information is very difficult and they even don't know how to use that or they would need that…. We've been criticising the UX evaluation, not about how we measure UX, but how we use the information it in industry. … But there is so much information that people don't bother to read or follow them. We need to make things simple and easy so that people don't have backgrounds they can understand. … This area of UX has the good side of interdisciplinary as well as the negative ones.

- *Lack of standard UX metrics renders redesign decisions prone to personal bias*es:

  *S*5: People make decisions based on their personal beliefs. They just pick from the UX measures the ones that support their existing belief, and ignore the other results that don't support. … We had noticed that the same icon did not work for various kinds of notification… We got feedback the people were annoyed… there was a very strong personality in the design team who said that he did not want the design changes because they look ugly… It is problematic that UX have no commonly agreed definition or no commonly agreed metrics. It allows people to use this kind of argumentation that "I believe that it is better UX". You don't need to justify, it can be a personal opinion even though there are tons of user feedback.

- *Packaging UX measures for decision makers and speaking their language:*

  *S*4: … social TV case we did Attrakdiff questionnaire and industry partner was very interested in that. They saw the potential in that when we had enough data, more convincing, more easily convince their superior of the organization to finance their projects, show the need for working on some aspects further; objective foundations.

  *S*5: It is not meaningless to measure moment-to-moment experience, but the question is how you use this information… But how to pack the thing and sell the thing to people making products or legislation decisions. … Strategy management what they are most interested in is that what are the elements that make users buy next devices from the same company as well and what can reduce the number of

helpdesk contacts. The first one is related to the future revenue of the company and the second one is related to the cost saving.

4.2.5 Philosophical arguments on UX Measurement

In discussing Q11-Q13, most of the interviewees anchored their responses in some real-life anecdotes, as presented in Section 4.2.4. However, some, especially *S*9, who is expert at design theory, addressed stimulating philosophical arguments on UX measurement, which are worth deeper reflections.

Radical constructivism versus scientific realism. Historically different philosophers have distrusted different things. There are those who like Plato distrust human perception and claim that there is a real world out there, which is distorted by our perceptions. In contrast, there are those who like Locke trust our perceptions more than anything else; accordingly the only true reality is the reality as we perceive it and the only thing we have access to are our own perceptions. Hence, the philosophical issue is to what extent our perceptions and feelings reflect a real world, whether a real world exists. For a non-realist, all experiences are valid. It tends to be scientific realism that discounts emotional experience. As all of our models are linguistic, the concepts that we choose and qualities that we choose to describe and understand our emotions are the ones that will shape the way we see them.

Reductionism versus selectivity. Arguably experience cannot be broken down. The problem with approaches that are not holistic is that the risk of over-interpreting a phenomenon. When we can choose to focus on a certain aspect of experience to measure, it is not a matter of reductionism but rather of selectivity. When we isolate a phenomenon of interest, which is deemed discreet enough to be an objective study, it is not a process of reduction; instead, it is a process of selection, as an integral part of a classic scientific method. The assumption underlying reductionism is that one actually has an understanding of the phenomenon as a whole and an ability to decompose it exhaustively into its principle components. But the current work in UX measurement seems suffering from the lack of such a holistic understanding.

Measurability and predictability. A critical implication of measuring is to enable prediction. However, the lack of plausible theories that connect experience measures as independent variables to design outcomes as dependent ones. People attempt to design measures in advance of their application. It can be a *fundamental* error because one *cannot* understand what is worth measuring until you understand the phenomenon of interest as a whole. The debate may not be reductionism versus holism. Instead, it can be much more partiality, premature commitment, hasty decision-making, unfounded assumptions that may threaten UX measures, because there is simply no point in measuring something until one understands why you want to measure it. If it is predictive, then a theory is required. While an experiential measure is correlated with a usage outcome, without a plausible theoretical framework we cannot identify the experience factor that *causes* the outcome. The major weakness of UX is the lack of theory.

4.3 Implications of the Interview Study

Most of the interviewees were ambivalent, if not sceptical, towards UX measures. A deeper understanding of the relationship between experience and memory, and of the temporality of UX are also required. While the utility and necessity of employing both quantitative and qualitative methods is commonly recognized, the concomitant issue of providing appropriate education and training in UX needs to be explored. Specifically, UX researchers and practitioners should be equipped with knowledge and skills to know why certain UX measures are taken and how to use and interpret them in order to inform design and development decisions.

Insights into the issues of UX measures have been gained from the interviews. The study has raised more questions than it can answer. As the number of participants was relatively low with most of them originating from one country, the views expressed might not be representative. Given this drawback, we have been motivated to expand the investigation on UX measurement with a larger scale survey. Results thereof are reported in Section 5.

5 STUDY 2: PAPER-BASED AND ONLINE UXMAS

5.1 Participants

As mentioned in Section 3.1, 35 out of about 100 workshop participants completed the paper-based UXMAS. The online UXMAS was run for about one month in July 2012 and attracted 332 responses of which 135 were complete. Altogether 170 responses were analysed. Table 5 shows the corresponding demographic data. One participant did not indicate his age range and some participants checked more than one box for their job (e.g. practitioner cum researcher), amounting the total to greater than 170. A relatively high percentage of participants in both the paper-based and online survey were practitioners. The work of eleven (out of 35) participants in the paper-based survey was design-oriented, whereas (with missing data) the distribution of 135 online survey respondents in terms of their job's relatedness to design is as follows: 28 (> 20% and <= 40%), 31 (> 40% and <= 60%), 17 (> 60% and <= 80%) and 31 (> 80%).

Six participants in the online survey and four in the paper-based survey indicated 'never' having done UX work; and their work area was described as web engineering, services science or simply HCI. Of the 170 participants, 48% had 'more than 5 years' of experience in UX work, 19% 'between 3 and 5 years', 17% 'between 1 and 3 years', 11% 'less than 1 year', and 6% 'never'. So, overall the participants can generally be regarded experienced in UX work.

The data of the paper-based survey and the online survey were analysed together, as research has demonstrated that the findings of surveys that are computer-based and paper-based do not differ (e.g. Lonsdale et al. 2006). Moreover, there is evidence that computer-based surveys have the advantage of better data quality in terms of fewer missing responses and higher internal consistency (Hanscom et al., 2002; Lonsdale et al., 2006).

5.2 Results and Discussion

Most of the data captured in this study are qualitative. For analysing them, we developed coding schemes for individual questions from Q6 to Q13, using thematic analysis (Boyatzis 1998) and the CUE model (Section 4.2.3; Figure 1). For the demographic data (Q1-Q5) and quantitative data of the five attitude

statements (Q7), correlation analysis was performed. Several constraints about our datasets should be clarified. While there were altogether 170 participants contributing to the pool of data, not all responded to every single question, especially those in Part B and Part C (not available in the paper-based survey). Hence, the number of data points per question varied.

*** Insert Table 5. Demographic data of all the participants ***

5.2.1 Reliability of Coding

The first two authors jointly developed a coding scheme for each of Q6 and Q8-13 (Tables 6-12). They then independently coded the responses, which when containing multiple arguments were segmented. Reliability of coding was evaluated using Krippendorff's alpha, Cohen's kappa and Scott's pi, producing consistent results. Landis and Koch (1977) distinguish the following brackets for kappa: 0.00-0.20 (slight agreement), 0.21-0.40 (fair agreement), 0.41-0.60 (moderate agreement), 0.61-0.80 (substantial agreement) and 0.81-1 (almost perfect agreement). On average reliability values were 0.67 for all three measures. Kappa values were 0.44 for Q6, 0.56 for Q8a, 0.92 for Q8c (answer: 'Yes'), 0.69 for Q9a, 0.94 for Q9b, 0.66 for Q9c, 0.65 for Q10a, 0.60 for Q10b, 0.60 for Q11, 0.59 for Q12, and 0.76 for Q13). Accordingly, three questions showed moderate agreement, six substantial agreement and two almost perfect agreement. Responses that were coded differently were fully reviewed and an agreement was reached on each difference. There were 19 instances of 'Q8c- I don't know' (i.e. the measurability of EQ most relevant to one's work; Table 3); only 4 had justification for which it was not worth to develop a coding scheme. The nine instances of 'Q8c- No' were coded using the scheme for Q10b (i.e. justification for non-measurable EQ).

5.2.2 Definition of a Measure (Q6)

Different facets were mentioned when participants were asked to describe what a measure is (Table 6). Basically, they focused on 'why', 'what' and 'how'; none addressed 'when'. While most of the facets are included in standard definitions of measurement (Hand 2004), the reference in the responses to the application of UX in product development here is noteworthy, indicating that measurement may be seen as relevant to design.

*** Insert Table 6. Categories of defining a measure ***

5.2.3 Attitude towards UX Measurement (Q7)

Principal components analysis of Q7, with cut-off value of 1 for eigenvalues, produced a one-factor solution (eigenvalue = 2.17), explaining 43% of variance in the items. Factor loadings were .61, -.68, -.53, .73, and .74, in order of items. Reliability was acceptable, with Cronbach's alpha = .65. Therefore, a mean attitude score was calculated per participant, scaled from -2 to 2 where 0 represents a neutral value, and used in subsequent analysis. Overall, attitude was positive, mean = 0.97, SD = 0.59, $CI(mean)_{.95}$ = [0.88; 1.06]. Correlations between attitude and participants' demographics were not statistically significant with small or negligible effect size. Correlations ranged from -.15 for company size to -.08 for practitioner (yes/no) to -.05 for age to -.06 for experience in UX work to -0.01 for extent of design-related work to .for 04 student (yes/no) to .03 for researcher (yes/no).

Analysis of individual items showed that four out of five items indicated a favourable attitude towards UX measurement; the confidence interval of the mean of the 'Evaluation' item exceeded 1 and those of the items 'Knowledge', 'Insane' (with reversed polarity) and 'Design' included 1, and none of these intervals included 0. For example, regarding 'Knowledge', both theoretical (e.g. hypothesis testing: "Evaluation typically reveals something new or verifies a hypothesis", P12) and practical (e.g., informing good design: "It can help to understand other aspects required to build good software design", P29) benefits were reported. Regarding 'Design', benchmarking was mentioned as an aid to design: "Defining a measure (definition & benchmark) tells us how we should approach design" (P97). Regarding 'Evaluation', the benefit of providing empirical evidence was stressed: "without measurement never ending yes/no (personal opinion based) debates are inevitable" (P152).

5.2.4 Measurability of Experiential Qualities (Q8-10)

The participants were asked to identify EQs that were of personal or professional relevance and their respective measurability (Q8a), that were (almost) certainly measurable (Q9a) and that were (almost)

certainly non-measurable (Q10a). The same categorisation scheme described in Section 4.2.3 is applied here. We recap the four categories:

- *Instrumental qualities* (INQ)
- *Non-instrumental qualities* (NIQ)
- *Short-term affective response* (STAR);
- *Long-term evaluative response* (LTER);

Note that some EQs could be categorized as either STAR or LTER; where necessary, we made a decision based on the context provided by responses to other parts of Q8-10. We pooled all the three sources of Q8a, Q9a and Q10a (Figure 3). For one of these questions, several participants named two or more EQs which we separated for coding. After eliminating unintelligible responses (e.g., 'a', 'hmm'), there were 180, 129 and 97 instances of EQs for Q8a, Q9a and Q10a, respectively. Some of the EQs were mentioned twice or even multiple times. While there are 123 unique EQs, 65 were mentioned only once, including uncommon qualities such as *explorability*, *serendipity* and *spirituality* and common ones such as *sadness* and *confusion*. To enhance the clarity and accessibility of the overall results on the EQ measurability, we exclude those EQs with frequency of one. If a participant identified an EQ for Q8a without explicitly declaring its measurability (i.e. Q8c-'I don't know'), then this EQ or its frequency is not included into the overall results; there are 25 such cases. If a participant named the same EQ for Q8c-'Yes' and Q9a (measurable) or for Q8c-'No' and Q10a (non-measurable), then it was counted once to avoid code inflation. Figure 3, as a form of Venn diagram, depicts 58 EQs and their associated frequencies of being mentioned as measurable, non-measurable (underlined numbers) or both (in the "overlap" area).

Several findings are noteworthy:

(i) The number of constructs (qualities or responses) falling in the category INQ is unexpectedly high (16 unique ones as compared with 22 in the case of STAR). It may imply that some participants tend to associate the term 'experiential quality' with traditional HCI concepts and that the distinctiveness of UX is not yet well-established. There could also be influences from some of the existing definitions of

UX such as "The user experience is the totality of end-users' perceptions … include effectiveness … efficiency … emotional satisfaction… " (Kuniavsky 2010);

(ii) Constructs that are exclusively regarded as non-measurable are rare: none for INQ and only one for each of the other three categories. For instance, *enchantment* in STAR one of the two participants explained: "I don't think an enchanted person transformed from what they were before is in a good position to put a number on the transformation. They don't have a stable position or perspective with respect to the experience." (P287);

(iii) The number of constructs that are consensually considered as measurable is highest in INQ, as most of those qualities such as *efficiency* are well-defined in practice and standards. In STAR, *frustration* – a concept commonly used in usability – was frequently mentioned;

(iv) Constructs falling in the 'overlap' of the four categories are intriguing: In INQ, *ease of use*, *usability* and *usefulness* are generally seen as measurable qualities, but some participants think otherwise. In STAR, *fun* splits the opinions of the participants evenly, whereas *enjoyment* and *engagement* are lopsided towards being measurable. In NIQ, *aesthetic appeal/beauty* was considered more as non-measurable; it seems inconsistent with the work published on quantifying this quality (e.g. Hassenzahl & Monk 2010). In LTER, *satisfaction* – as one of the canonical three usability metrics – was mentioned by 22 participants with only one treating it as non-measurable, whereas *happiness* was rather regarded more as non-measurable.

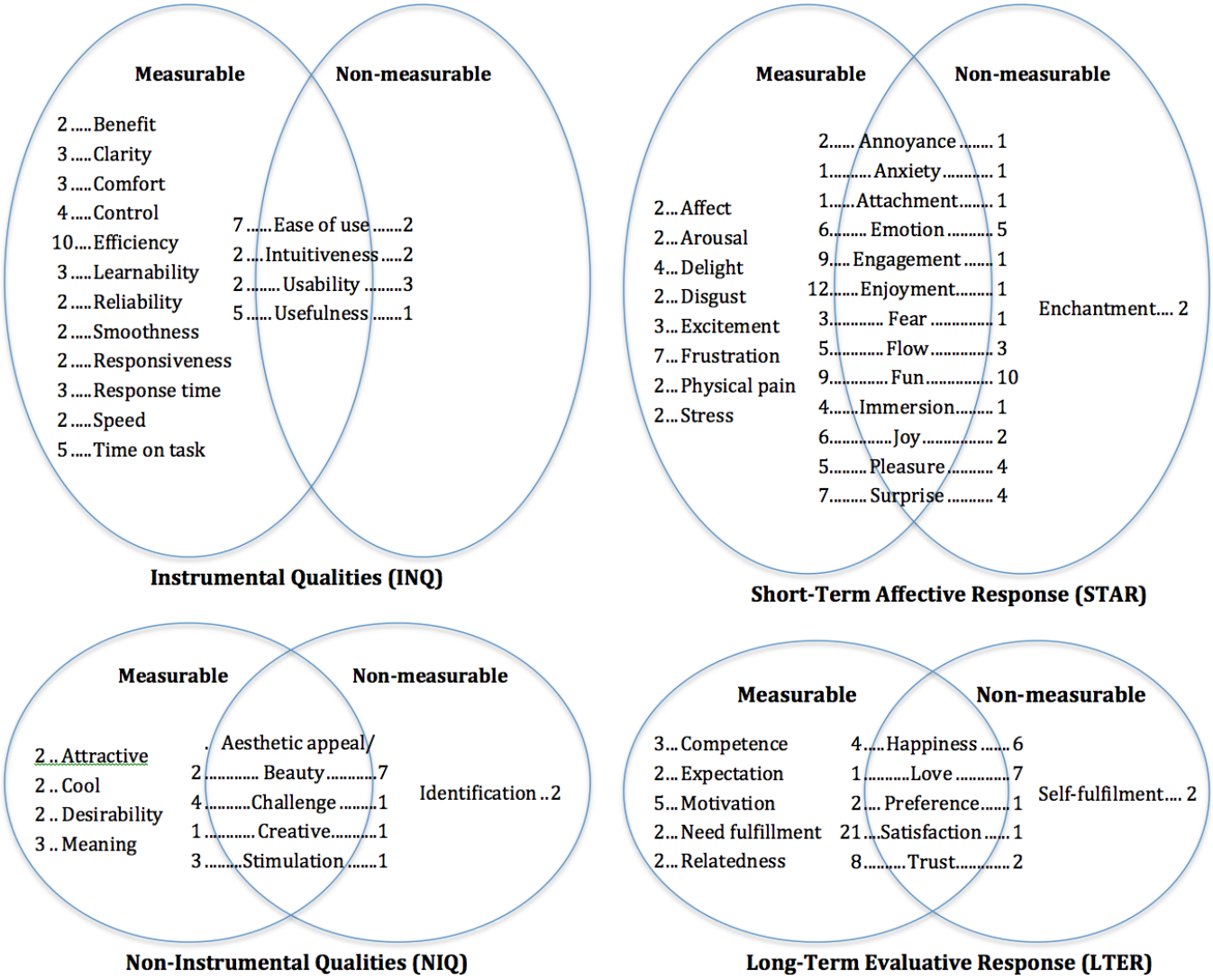

Figure 3. Measurable and non-measurable experiential qualities (Q8, Q9 and Q10)
Note: The qualities in the middle shared opinions, e.g. 2 respondents mentioned Annoyance as an example of a clearly measurable and 1 as a clearly non-measurable quality of experience

*** Insert Table 7. Categories for justifying construct measurability ***

To explore the rationale for the perceived measurability, responses to Q9c (why can) and Q8c-'No' and Q10b (why can't) were analysed. Table 7 and 8 display the codes. For the measurable EQs named, participants were asked to describe how they can be measured (i.e. responses to Q8c-'Yes' and Q9b). Prevailing measurement approaches were mentioned, which can be categorised into three main types: *self-report* (n=132), *overt behaviour* (n=86), and *biological* (n = 24). As expected (cf. Vermeeren et al. 2010), self-report methods were predominant, but not exclusively chosen.

*** Insert Table 8. Categories for justifying construct non-measurability ***

5.2.5 Arguments for and against UX measurement (Q11-13)

Here we report the results from online and paper-based data collection (for the related results of the interviews, see Section 4.2). When asked about theoretical arguments for or against UX measurement (Table 9), some participants provided responses apparently addressing practical issues. In particular, the theoretical and practical utility and necessity of UX measurement reflects the perceived inherent need for measurement in order to make progress in UX. Besides, the need for particular conditions to be met as a basis for good UX measurement was highlighted.

When asked about methodological arguments for or against UX measurement (Table 10), some participants addressed apparently practical issues, both benefits and problems. At a fundamental level, some believed that experience is not amenable to measurement and others that UX measurement suffers from a lack of definition of the concept of UX. Ambiguity of the causes of UX was also reported as an argument against. Of particular interest is the argument that lack of education hampers UX measurement; this view was also repeatedly mentioned in the interviews (Section 4.2).

*** Insert Table 9. Categories of theoretical argument for and against UX measurement ***

*** Insert Table 10. Categories of methodological arguments for and against UX measurement ***

The main stated practical argument for UX measurement was relevance to design and marketing (Table 11). There were complementary arguments with respect to resources and knowledge required. One of the fundamental issues against was the inherent nature of UX measures such as context-dependence and subjectivity. A practical argument against was factors in corporate culture (organizational climate) hindering UX work. Another consideration was that conditions (in an organization or project) can influence UX measurement.

*** Insert Table 11. Categories of practical arguments for and against UX measurement ***

6. GENERAL DISCUSSION

Prior to discussing individual issues pertaining to UX measurement, it is deemed useful and necessary to clarify how UX theories are related to UX models. Generally speaking, the term 'model' can be understood as a description of applying a theory in a particular context with a set of specific initial and boundary conditions. Put simply, a UX model is an abstract representation or approximation of an underlying UX theory. Furthermore, the relationship between theory and model can also be understood as the link between conceptual realm and empirical realm (Jaccard & Jacoby 2010). In this paper, we focus on the empirical realm whilst emphasizing the urgency to explore deeper the conceptual one.

6.1 Comparison of the Two Studies

Basically, participants of Study 1 and Study 2 were asked to respond to the same set of questions, which were delivered orally with a printed copy for reference and in a written format (printed or digital), respectively. Their ratings for the five statements of Question 7 were comparable ($mean_{Study1}$= 3.86 vs. $mean_{study2}$ = 3.97), suggesting that they had similar attitudes towards UX measurement (NB: no inferential statistics are computed, given the different data collection procedures). However, their qualitative responses, which are much more elaborated in Study 1, indicate that the participants of Study 1 were relatively more sceptical about UX measurements than those of Study 2. The scepticism was related to the participants' background training and work experiences such as organizational constraints. With regard to the (non-)measurability of experiential qualities (EQs), apparently the number of unique EQs mentioned in Study 1 is much lower than that of Study 2, due to the corresponding differences in the number of participants. Two of 16 EQs in Study 1, *curiosity* and *enlightenment*, are not covered in Study 2. Furthermore, for Study 1 and Study 2, no non-measureable instrumental quality (INQ) is identified. Whereas for Study 1 there is no non-measurable non-instrumental quality (NiQ) or short-term affective response (STAR), for Study 2 the corresponding EQs are *identification* and *enchantment*. In Study 1, *happiness* and *love* are considered as non-measurable long term evaluative responses (LTER), but in Study 2, four out of ten and one out of eight participants, respectively, regarded them as measurable.

6.2 Evaluation and Development of the UX-attitude Scale

A single UX-attitude dimension (i.e. Q7 with the five items, Table 3) was found to have factorial validity and acceptable reliability at this stage of development. Most important, the overall favourable attitude was not correlated with demographics, including practitioner or researcher status or experience in UX work. In addition to the encouraging psychometric results (factor structure and reliability), the categorised responses to Q11-Q13 provide a further basis for evaluating the content of and developing of the UX-attitude survey into a scale. To varying degrees, these responses provided evidence for the 5 item-statements. In particular, 'Increase in knowledge' (Item 1) was supported by the category 'Theoretical utility and necessity of UX measurement' (Q11). 'Important for design' (Item 4) was related to 'Relevance to design and marketing' (Q13) as well as perhaps to 'Practical utility and necessity of UX measurement' (Q11) and 'Practical benefits' (Q12). 'Insane' (Item 2) may be related to 'Research approaches against measurement/Theoretical objections' (Q11), 'Incompatibility between the nature of measurement and inherent characteristics of experience' (Q12), and 'Scepticism about UX measurement' (Q13). 'Pain' (Item 3) may be related to 'Practical objections' (Q11), 'Practical problems' (Q12) and 'Resources required are high' (Q13). 'Important for evaluation' (Item 5) may be related to 'Practical utility and necessity of UX measurement' (Q11) and 'Practical benefits' (Q12).

In addition, several other aspects of UX measurement emerged from the coded responses. Hence, all in all, a single psychometrically measured dimension provided a concise representation of some of the arguments for and against UX measurement. For further development of explanatory and predictive research into people's thoughts regarding UX measurement it may be useful to consider a theoretical framework that links attitude to behaviour (the actual use of UX measures by UX practitioners and researchers) such as the technology acceptance model (Venkatesh & Bala 2008). One useful consideration may be the inclusion of behavioural beliefs (people's beliefs about the consequences of UX measurement). This work could pinpoint the drivers and inhibitors of behaviour and thereby help in prioritizing aspects of UX measurement that should be addressed by research to increase its acceptance. Item 1 of the existing survey

is a good example of a behavioural belief, whereas Item 2 is a 'pure' statement of attitude (and to a lesser extent perhaps Item 3). In this sense, despite the encouraging psychometric results, the existing instrument may be seen to be measuring a mixture of attitude and behavioural belief for separate study in future research.

6.3 UX Constructs

In the four categories – instrumental quality (INQ), non-Instrumental quality (NIQ), short-term affective response (STAR), and long-term evaluative response (LTER) – most of the UX constructs named were deemed measurable or the opinions were divided. A few such constructs were unanimously deemed not to be measurable. Main arguments for this measurability were feasibility of measurement and availability of specific measures. Main arguments against were subjectivity of UX and lack of definition of UX.

While the survey revealed various UX constructs of which a large number are classified as experiential qualities (EQs; see Figure 1), the following question remains: is 'stamp-collecting' (accumulating without underlying theory) of EQs useful? In order to be useful, EQs should be grounded in higher-order outcomes, such as high-level design goals (e.g., users' loyalty to an online service) or preferably – but not necessarily only – objective outcomes that improve people's productivity or health (Sheldon et al. 2001). For example, *flow* experience (STAR) has been found to have a positive impact on task performance (van Schaik & Ling 2012a). Stamp-collecting will not necessarily achieve this grounding and even if it does we will not know for certain without modelling UX. Fortunately, this grounding can be established by model building (Jaccard & Jacoby 2010). This involves relating 'upstream' predictors (e.g. INQ, NIQ, STAR) to 'downstream' outcomes (e.g., LTER). Through necessary simplification, models can help specify the variables that have the largest impact on outcomes to be measured, in order to make approximate predictions of these outcomes, in agreement with Voltaire's proverb "The perfect" (prediction) "is the enemy of the good" (approximation). The extent of simplification is a matter of choice, depending on, for example, resources available and ambition of the project. For this purpose, appropriate existing models (e.g. the CUE model) can be adopted and adapted. Otherwise, new models can be constructed, with useful guidance through a

toolbox of model-building techniques (ibid). Thus, UX modelling can help focus the selection of EQs and establish whether the selected EQs are in fact predictive of higher-order outcomes.

6.4 Arguments about UX measurement

Conceptually, the coded responses to Q11-Q13 cover three broad types of argument regarding UX measures. Theoretically, ongoing debates on reductionism versus holism will persist. Ultimately this is a philosophy of science question that may never be settled (cf. Section 4.2.5). Nonetheless, a key question is whether we adopt an all-or-none approach (e.g. exclusively qualitative) or an integrated one. Some participants of the study, while holding a 'sympathetic' attitude towards measurement, remain hesitant to adopt UX measures. For instance, one asserted that "My approach to experience is holistic… experience as something interpreted rather than measured, design as participative, and evaluation as storied … I have no objection at all to other approaches to user experience and I can see their value. They are just not what I do" (P287). In contrast, some participants believe firmly the uses of measurement: "There are many design aspects that require the use of physical measures as well as social and psychological measures. Sure, it is possible to design with limited reference to measures, but this will restrict the type and the quality of that which is designed. Evaluation is severely limited without measures" (P1). We assume that this gap of acceptance of UX measures can be bridged by stronger efforts in theorizing UX. In Kuutti's words: "empirical data is blind without a theory, but any theory without connection to empirical data is vacuous" (in Roto et al. 2010, p.22). This view is resonated: "Even if we have a compelling theory, it needs measurement for support" (P11). Based on the existing work on UX and relevant arguments gathered in this study, it is deemed necessary to deepen understandings about the relationships between experience and memory (i.e. temporality of UX), the dynamics of motivation and emotion, and how cognition, affect and disposition interact.

Furthermore, we believe the ongoing development of methodologically sound and practically useful UX measures, with good measurement properties can also help further increase the acceptance and use of UX measurement. Specifically, work should focus on how UX measurement can be made practical by

demonstrating its usefulness in design and marketing, its timeliness and its advantages in terms of resource savings. It is expected that – just as in the history of the usability through usability engineering – as methodological strengths and practical benefits as well as education in UX measurement develop further, theoretical arguments for UX measurement can be accepted more readily.

6.5 Attitude and Behaviour on UX measurement

In Section 2.2 we report our review of 58 publications on UX empirical research studies. While it is not fair to compare the results of the actual UX measurement taken in practice, given that the survey respondents might not be those who had conducted the studies reviewed (some overlap could exist, though), it is interesting to note some similarities and discrepancies between the two sets of findings. Comparing Table 1 with Figure 3, *frustration*, which is often assessed with a questionnaire/scale in usability testing, is rarely mentioned by the survey respondents or measured in the selected research studies reviewed. In contrast, *flow*, which is most commonly measured EQ in the selected studies, but the survey respondents did not consensually agree on its measurability (5 'yes' vs. 3 'no'). Considering *aesthetics,* as shown in our review as well as Bargas-Avila and Hornbæk's (2011), it is a commonly measured UX construct in practice, but among the nine survey respondents seven regarded it as non-measurable. A deeper analysis of such a discrepancy between attitude and behaviour with reference to UX practices and theories may shed some further light onto the issues of UX measurement.

6.6 Limitations

UXMAS comprises mostly open-ended questions, which typically undermine response rate. While the number of responses could have been higher, the ideas shared were deemed valuable to gain further understanding. Another typical drawback of the web-based survey is self-selected sample. While it was very difficult for us, if not impossible, to capture all different opinions in the field, the responses did represent a spectrum of attitudes. Further, the questions were deliberately broad to stimulate thoughts, this might lead to misinterpretation. When such cases were detected in our analysis, although there were few, they were excluded.

7. CONCLUSION

UX, as a recently established research area, is still haunted by the challenges of defining the scope of UX in general and operationalizing experiential qualities in particular. In summary, the attitude towards UX measurement shown by the interviewees in Study 1 could generally be described ambivalent and sceptical. This contrasted with the more positive attitude expressed by the survey respondents in Study 2, although there were nuanced views on details of UX measurement. Overall, a significant implication of both studies is that it is necessary for UX professionals to identify plausible means for compromising the difficulties of evaluating UX in a holistic manner with the limitations of adopting the reductionist approaches. Employing quantitative measures to the exclusion of qualitative accounts of user experiences, or vice versa, is too restrictive and may even lead to wrong implications. Specifically, it is essential to understand why certain UX measures are taken and how to use and interpret them in order to inform design and development decisions. In summary, the contribution of this paper is to provide empirical evidence regarding the HCI community's attitude towards UX measurement as a base line for progress in UX measurement.

As mentioned earlier in Section 2, one of the contentious arguments about UX measurement is their implications to the redesign of the system under evaluation. UX metrics, when grounded in robust theoretical models and are operationalized properly, should be able to indicate types of affective and emotional responses as well as their extent (e.g. psycho-physiological measures; self-reporting scales) and the context where they are elicited (including people and their activities in which they engage). Such informational values of metrics are acknowledged by the participants of both Study 1 and Study 2. Nonetheless, metrics, be they usability or UX, are typically used for summative evaluation for benchmarking a system against competitive ones and for validating the improvement on the previous version. Apparently, metrics are also useful for formative (or diagnostic) evaluation to identify 'UX problems' (cf. usability problems). However, as remarked by the participants of our studies, more explicitly by those in Study 1, quantitative evaluation feedback might not be as useful as the qualitative ones for deriving alternative design ideas. Nonetheless, numeric values have their strength of being simple, precise

and neat. As described by some participants in both studies, UX measures can have the power to persuade decision-makers to modify the problematic design as indicated by the measures, although how the modifications should be implemented may not be sufficiently informed by such measures. All in all, we tend to conclude that UX measurement can play an important role in sustaining the interplay between evaluation and redesign by driving the changes required, but the debate whether it is more or less important than qualitative feedback remains.

We propose the following agenda of a plausible approach to enhancing the acceptance of UX measures: (a) UX modelling grounded in theories to link experiential qualities with outcomes. Specifically, we argue that theoretical frameworks examining the tripartite relationship between affect, action, and cognition are much more relevant than those focusing only on one of these three aspects; (b) the development of UX measurement tools with good measurement properties and (c) education within the HCI community to disseminate validated models and measurement tools and their successful application. All in all, revisiting the motive of this study that there seems a widening gap between the two major groups of UX professionals (i.e. one emphasizing objective measures and the other subjective accounts of experiential qualities), we see the need to amplify the earlier call (Boehner et al. 2007) for mutual recognition of strengths and weaknesses of the related approaches and values. This, in our view, will not only advance the emerging UX research and but also strengthen the interplay between UX evaluation and system development.

REFERENCES

Bargas-Avila, J.A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience. In *Proc. CHI'11*

Bartholomew, D. J. (2006) (Ed.). *Measurement Vol. I-IV*. Sage Benchmarks in Social Research Methods

Boehner, K., DePaula, R., Dourish, P., & Senger, P. (2007). How emotion is made and measured? *International Journal of Human-Computer Studies*, 65(4).

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage.

Epp, C., Lippold, M.,& Mandryk, R.L. (2011). Identifying Emotional States using Keystroke Dynamics. In *Proc. CHI* 2011.

Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. In *Proc. DIS '04.*

Gross, A.,& Bongartz, S. (2012). Why do I like it? Investigating the Product-Specificity of User Experience. In *Proc. NordiCHI 2012* (pp.322-330).

Hand, D.J. (2004). *Measurement theory and practice*. Wiley-Blackwell.

Hanscom, B., Lurie, J.D., Homa, K., Weinstein, J.N., 2002. Computerized questionnaires and the quality of survey data. Spine, 27, 1797-1801.

Hartmann, J., De Angeli, A. & Sutcliffe, A. (2008). Framing the user experience: information biases on website quality judgment. In *Proc. CHI* '08 (pp. 855–864).

Hassenzahl, M. (2008). User experience (UX): towards an experiential perspective on product quality. In Proc. *IHM'08* (pp.11-15).

Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human-Computer Interaction*, *25(*3), 235-260.

Hassenzahl, M., Heidecker,S., Eckoldt,K., Diefenbach, S.,& Hillmann U. (2012). All You Need is Love: Current Strategies of Mediating Intimate Relationships through Technology. ACM Transactions on Computer-Human Interaction, Vol. 19(4).

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Man-Machine Studies, 64*, 79-102.

Hornbæk, K., & Law, E. L-C. (2007). Meta-analysis of correlations among usability measures. In *Proc. CHI'07.*

Jaccard, J. & Jacoby, J. (2010). *Theory construction and model-building skills*. Guilford Press.

Kahneman, D. (2011). *Thinking fast and slow*. Penguin.

Karapanos, E., Zimmerman J., Forlizzi J., Martens, J-B.(2010). Measuring the dynamics of remembered experience over time*. Interacting with Computers , 22*, 328–335

Kaye, J. Buie, E.A., Hoonhout, J., Höök, K., Roto,V., Jenson,S. & Wright, P. (2011). Designing for user experience: academia & industry. In *CHI EA 2011* (pp. 219-222)

Kuniavsky, M. (2010). *Smart things: ubiquitous computing. User experience design*. Morgan-Kaufman,

Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lang, P. J. (1994). The varieties of emotional experience: A meditation on James-Lange theory. *Psychological Review*, *101*(2), 211-221.

Larsen, R. J., & Fredrickson, B.L. (1999). Measurement issues in emotion research. In D. Kahneman, E. Diener & N. Schwarz (Eds.), *Well-being*. Sage.

Law, E. L-C, Roto, V., Hassenzahl, M., Vermeeren, A., & Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. In *Proc. CHI '09* (pp. 719–728).

Law, E. L-C., & Schaik P. van (2010). Modelling user experience – an agenda for research and practice. *Interacting with Computers, 22*(5), 313–322.

Law, E. L-C., Schaik, P. van., & Roto, V. (2012). To measure or not to measure UX: An interview study. In *Proc. Workshop I-UxSED* 2012, NordiCHI'12.

Law, E.L-C. (2011). The measurability and predictability of user experience. In *Proc. EICS 2011* (pp 1-10).

Lonsdale, C., Hodge, K., Rose, E.A., 2006. Pixels vs. paper: Comparing online and traditional survey methods in sport psychology. Journal of Sport and Exercise Psychology, 28, 100-108.

Löwgren, J. (2007). Fluency as an experiential quality in augmented spaces. *International Journal of Design, 1*(3), 1-10.

McCarthy, J., & Wright, P. (2004). *Technology as experience*. MIT Press.

Mumm,J., & Mutlu, B.(2011). Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. Computers in Human Behavior, 27 (2011) 1643–1650

Norman, D. (2008). www.montparnas.com/articles/don-norman-on-user-experience-design/

Obrist, M., Law, E.L-C., Väänänen-Vainio-Mattila, K., Roto, V., Vermeeren, A., & Kuutti, K. (2011). UX research- which theoretical roots do we build on – if any. In *Extended Abstract CHI'11.*

Olsson, T., Salo, M. (2012). Narratives of Satisfying and Unsatisfying Experiences of Current Mobile Augmented Reality Applications. In *Proc. CHI 2012*.

Park,D. Leen, J-H., & Kim, S. (2011). Investigating the affective quality of interactivity by motion feedback in mobile touch screen user interfaces. *Int. J. Human-Computer Studies, 69*, 839–853

Procci, K., Singer, A. R., Levy,K.R. & Bowers, C. (2012). Measuring the flow experience of gamers: An evaluation of the DFS-2. *Computers in Human Behavior, 28*, 2306–2312

Roto, V., Law, E., Vermeeren A, Hoonhout, J. (2010). *Demarcating User eXperience. Dagstuhl Seminar*. http://drops.dagstuhl.de/opus/volltexte/2011/2949/pdf/10373_AbstractsCollection.2949.pdf

Sauro, J. & Lewis J.R. (2009). Correlations among Prototypical Usability Metrics: Evidence for the Construct of Usability. In *Proc. CHI'09.*

Schaik, P. van, Hassenzahl, M., & Ling, J. (2012). User experience from an inference perspective. *Transaction on Human-Computer Interaction*, 19(2), Article 11.

Schaik, P. van., & Ling, J. (2012a). An experimental analysis of experiential and cognitive variables in web navigation. *Human-Computer Interaction, 25*(3), 199-212.

Schaik, van P., & Ling. P. (2012b). A cognitive-experiential approach to modelling web navigation. *Int. J. Human-Computer Studies, 70*, 630–651

Scherer, K. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*(4), 695-729.

Sheldon, K. M., Elliot, A.J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology, 80*(2), 325-339.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science,* 103(2684).

Swallow, D., Blythe, M., & Wright, P. (2005). Grounding experience: relating theory and method to evaluate the user experience of smartphones. In *Proc. EACE'05* (pp. 91-98).

Thomson, W. (1891). *Popular Lectures and Addresses, Vol. I*. (p.80). MacMillan.

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human-technology interaction. *International Journal of Psychology, 42*(4), 253-264.

Tullis, T., & Albert, W. (2008). *Measuring the user experience*. Morgan Kaufman.

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Science, 39*(2), 273–315

Vermeeren, A. P.O.S. Law, E. L-C., Roto, V., Obrist, M., Hoonhout, J., Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. In *Proc NordiCHI 2010* (pp. 521-530).

Wixon, D. R. (2011). Measuring fun, trust, confidence, and other ethereal constructs: it isn't that hard. *Interactions, 18*(6): 74-77.

Tables 1-11

Measured UX construct	Frequency
Flow: general (8); multi-dimensional(4)	12
Aesthetic/beauty	9
Emotion: multi-dimensional(4); general (3)	7
Enjoyment	5
Affect: general (3); multi-dimensional (2)	5
Arousal/valence	4
Hedonic quality	4
Intrinsic motivation	4
Presence	4
Engagement: general (2); multi-dimensional (2)	4
Attractiveness	3
Satisfaction	3

Table 1. UX Constructs measured in the recent empirical studies

Q1. Gender: Female, Male
Q2. Age: <=20, 21-30, 31-40, 41-50, >50
Q3. I am a: Practitioner, Researcher, Student, Other
Q4. How long have you worked in the area of UX? (Never, <1 year, 1-3 year, 3-5 year, >5 year). Please describe the topic and related work.
Q5. How much of your work/study is related to design? (<=20%, >20% and <=40%, >40% and <=60%, >60% and <=80%, >80%). Make a comment on your choice.

Table 2. Background questions

Q6. What is a 'measure'?
Q7. (a) Please rate your agreement with each of the following statements (5-point Likert scale); (b) Explain your ratings ▪ UX measures lead to increase of knowledge ▪ UX measures are insane ▪ UX measures are a pain ▪ UX measures are important for design ▪ UX measures are important for evaluation
Q8. (a) Name a specific experiential quality (e.g., fun, surprise) that is most relevant to your work; (b) Explain the relevance; (c) Do you think the named quality can be measured: If 'yes', describe how; If 'no', describe why.
Q9. (a) Name an experiential quality that you are (almost) certain is measurable; (b) How can it be measured and when (before/during/after interaction)? (c) Why are you so (almost) certain about its measurability? What is your reservation, if any?
Q10. (a) Name an experiential quality that you think (almost) impossible to measure; (b) Why do you think so? What is your reservation, if any?

Table 3. Five main questions on UX measures

Q11. Which theoretical arguments (e.g. reductionism) are for or against UX measurement?
Q12. Which methodological arguments (e.g. validity) are for or against UX measurement?
Q13. Which practical arguments (e.g. cost) are for or against UX measurement?

Table 4. Questions for in-depth discussion

Data Source	Gender*		Age*				Job		
	F	M	<=30	>30 and <=40	>41 and <=50	>50	Practitioner	Researcher	Student
Paper (n=35)	19	16	16	15	3	1	21	9	4
Online (n=135)	44	90	21	58	29	25	69	80	16
Total	63	106	37	73	32	26	90	89	20

Table 5. Demographic data of all the participants

*One online respondent did not provide the information on gender and another one no information on age

Category name with Instances	**N**
Type of data: "A measure is what you use to determine the value of a specific variable you are interested in (either qualitative or quantitative)" (P91)	64
Comparison and evaluation: "A set of measures are a way to see if we are achieving the levels of quality we had planned for." (P62)	48
Objectivity: "On the basis of evidence/data that is independent of individual opinion." (P204)	33
Relation with concepts or qualities: "Measures are most useful when treated as tools for critical reflection regarding the concept." (P55)	31
Data collection: "A measure is the specification of a specific data-collection which describes a process-phenomenon." (P93)	24
Application in product development: "Measures are only useful if their use enables meaningful decisions to be made about the development of a design or the direction of an enquiry." (P204)	12
Quality of measurement: "It should be reliable and valid." (P134)	5
Data analysis: "They need statistics to arrive at generalizable results." (P124)	2

Table 6. Categories of defining a measure

Category name with instance	**n**
Feasibility - users' observable behaviour, consciousness of experience or ability to respond: "Challenge defines the difficulty of an operation, which seems to influence only its accomplishment, error rate and time taken." (P10)	36
Availability of validated or commonly used measurement methods/instruments: "Medical studies have a long tradition in studying and assessing physical pain.." (P124)	24
Importance: "Because pleasure is, I believe, a core element of UX." (P81)	3
*Interpretation: "*It is an easy interpretation of human behavior. There are of course cultural differences" (P297)	2

Table 7. Categories for justifying construct measurability

Category description	n
Subjectivity: "Experience is not measurable in the way distance or weight is. We need to rely on subjective interpretation (either by the observer or the subject)." (P319)	22
Definition: "Very difficult because there is even no common definition of surprise. Can be positive or negative, intense or not, rapidly or slowly occurring. Hard to define a common measure." (P127)	12
Practicality: "It's expensive to do reliably as it requires at least months of longitudinal research. Hard to justify." (P290)	9
Utterance: "I think it is difficult for people to express their general experience of a product. … explaining why they enjoy it is pretty tough ..." (P91)	8
Scepticism: "Any measure will be so reductive to be meaningless." (P109)	8
Uniqueness: "This [elation] rarely happens during interaction and as an extreme emotion would be difficult to quantify" (P265)	8
Context: "Also emotional responses will often, we think, be significantly influenced by the usage context (social, physical etc)." (P204)	7
Response Bias: 'Also, I assume that most people will hesitate to self-report this feeling [cool or hip] out of fear of being judged as feeling smug." (P213)	6
Multidimensionality: "Satisfaction is so complex and multifaceted. It's the sum of many parts … only some of the parts can (or will) be tested." (P121)	6
Multi-causality: "There are so many factors that influence fun" (P10)	5
Quality of measures: " it involves lot of imprecise video/face emotion reasoning techniques" (P87)	4
Organization: "I don't think our company has a deep enough understanding of what [desirability] is" (P19)	3
Obtrusiveness: "The fact that you have to observe somebody and let that person know that you observe her influences all emotional measures significantly." (P183)	3
Reference: "The problem is calibrating the scale: what is the highest pleasure score possible?" (P37)	2

Table 8. Categories for justifying construct non-measurability

Category name with instance	n
Theoretical arguments for UX measurement	
Research approaches in favour of measurement: "self-determination theory revised theory, flow theory, two factors theory" (P129)	15
Theoretical utility and necessity of UX measurement: "Increased understanding of the complexity of UX; Increased understanding of the relations between the different aspects of UX; Increased understanding of the impact of context on UX" (P134)	7
Practical utility and necessity of UX measurement: "Eventually everything gets measured in the bottom line. it's intellectual laziness to wait until that moment. you have to do it earlier, closer to what you can manipulate/improve" (P65)	7
Theoretical arguments against UX measurement	
Research approaches against measurement/ Theoretical objections: "Against: some people consider UX too phenomenological to be measured, it is an overall feeling that is difficult to break to pieces (the overall experience is more than the pieces." (P115)	14
Practical objections: "Arguments against: Increased complexity, which may cause the establishment of too complex measurement tools." (P134)	5
No theoretical arguments for or against UX measurement	
Conditional stance on UX measurement: "First we need a definition of User Experience, then we can develop measures for Experiences. I cannot see any theoretical argument against defining the term "Use Experience". There are of course many arguments with respect to the chosen measurement method." (P76)	7
Disinterest in UX theory: "I have not considered theoretical arguments either for or against UX measurements." (P309)	4

Table 9. Categories of theoretical argument for and against UX measurement

Category names with instances	n
Methodological arguments for UX measurement	
Practical benefits: "Need to measure in order to study effects on UX." (P134)	15
Good measurement properties: "Statistical validity, reproducibility, methodological clarity." (P124)	12
Methodological arguments against UX measurement	
Limitations of measurement methods: "People may say different things that they think for example for social acceptance reasons. People may not be able to articulate their feelings in a measurable way." (P115)	11
Incompatibility between the nature of measurement and inherent characteristics of experience: "Reliability and subjectivity are barely given through subjectivity of experiences" (P68)	7
Lack of definitional/theoretical frameworks: "What UX needs is a valid operational definition to be able to measure properly." (P140)	4
Practical problems: "Complex measures may be difficult to implement and use in practical UX work." (P134)	3
No methodological arguments for/against UX measurement	
Complementary concepts and methods: "Mixed and qualitative methodologies can support UX studies" (P116)	5
Lack of education: "most researchers in HCI don't know the most basic techniques to ensure a minimum of quality in their measures!" (P17)	2

Table 10. Categories of methodological arguments for and against UX measurement

Category names with instances	**n**
Practical arguments for UX measurement	
Relevance to design and marketing: "… if you need to gain understanding of the direction to which your design is going, then you should measure" (P33)	14
Resources required are low: "Cost is fairly minimal both in terms of money as well as time so measurement can be easily combined with user lab" (P152)	6
Positive effects on development costs: "avoiding additional costs after launching a system" (P160)	6
Relevance to research and education on UX: "the main argument in favor of UX is … an educational one. It would certainly make some students (e.g., engineers...) feel safer to deal with some measures of UX instead of with a less tangible reality of an UX." (P128)	5
Practical arguments against UX measurement	
Resources required are high: "Time, cost, and expertise are three practical arguments against UX measurement. Taking any type of measurement … takes a great deal of time to set-up, requires expensive, complicated equipment, and researchers need a high level of expertise to interpret the results." (P81)	18
Lack of standard methods and definition: "Lack of well-defined techniques for measuring UX" (P301)	5
Inherent nature of UX measures: "UX is a conglomeration of factors that may or may be impossible to measure. As such, UX may be impossible to measure as well..." (P25)	4
Corporate culture: "[…] project managers, product managers, etc. value other aspects of system development higher sometimes and therefore try to argue against it" (P182)	4
Scepticism about UX measurement: "good UX is the consequence of good usability engineering and does not need explicit treatment." (P210)	3
Lack of knowledge: "The costs are high because of the lack of knowledge on which parameters to include and which not." (P233)	2

Table 11. Categories of practical arguments for and against UX measurement