# Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires

A. Baki Kocaballi
Australian Institute of Health
Innovation, Macquarie University
*baki.kocaballi@mq.edu.au*

Liliana Laranjo
Australian Institute of Health
Innovation, Macquarie University
*liliana.laranjo@mq.edu.au*

Enrico Coiera
Australian Institute of Health
Innovation, Macquarie University
*enrico.coiera@mq.edu.au*

**User experience (UX) has become an important aspect in the evaluation of interactive systems. In parallel, conversational interfaces have been increasingly used in many work and everyday settings. Although there have been various methods developed to evaluate conversational interfaces, there has been a lack of methods specifically focusing on evaluating user experience. This study reviews the six main questionnaires for evaluating conversational systems in order to assess the potential suitability of these questionnaires to measure various UX dimensions. We found that (i) four questionnaires included assessment items, in varying extents, to measure hedonic, aesthetic and pragmatic dimensions of UX; (ii) two questionnaires assessed affect, and one assessed frustration dimension; and, (iii) enchantment, playfulness and motivation dimensions have not been covered sufficiently by any questionnaires. We recommend using multiple questionnaires to obtain a more complete measurement of user experience or improve the assessment of a particular UX dimension.**

*User experience, conversational agents, dialogue systems, voice interfaces, evaluation, standardised questionnaires*

## 1. INTRODUCTION

User experience (UX) has become an important aspect of interactive system evaluations in the last two decades. Although there is an increasing adoption and acknowledgement of the need to understand and evaluate user experience, two extensive reviews have not found a consensus in defining and evaluating user experience (Law et al. 2009; Bargas-Avila and Hornbæk 2011).

According to ISO (2010), user experience is a "*person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service*". According to a survey study of 275 UX researchers and practitioners (Law et al. 2009), this definition is in line with the views of most respondents about the subjectivity of UX. There are also three notes added to this definition. The first emphasises the extensive range of UX dimensions to be considered at different stages of using a product: "*User experience includes all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use*". The second note draws our attention to the various brand-related, system-related, user-related, and context-related factors shaping UX: "*User experience is a consequence of brand image, presentation, functionality, system performance, interactive behaviour and assistive capabilities of the interactive system, the user's internal and physical state resulting from prior experiences, attitudes, skills and personality, and the context of use*". Finally, the third note aims to clarify the relationship between usability and UX: "*Usability, when interpreted from the perspective of the users' personal goals, can include the kind of perceptual and emotional aspects typically associated with user experience. Usability criteria can be used to assess aspects of user experience*". Although the ISO's main definition and the notes provide a valuable understanding of UX and its broad scope, from a more practical viewpoint, it is still not clear what constitutes UX for a particular user interacting with a particular product in a particular context. There is a lack of information explaining what UX dimensions need to be considered for different types of projects.

In their review study Bargas-Avila and Hornbæk (2011) identified some frequently assessed UX dimensions. Although the list of UX dimensions is not definitive, it provides a useful starting point to consider UX factors in designing and evaluating interactive systems. The list includes: generic UX, affect/emotion, enjoyment/fun, aesthetics/appeal, hedonic quality, engagement/ flow, motivation, enchantment, and frustration. The authors noted

© *2018* Baki et al.. *Published by BCS Learning and Development Ltd. Proceedings of British HCI 2018. Belfast, UK.*

there are more UX dimensions, but they are less frequently assessed. In our study, we will use the list of UX dimensions offered by Bargas-Avila and Hornbæk (2011) as a basis for our assessment.

It is important to situate the focus of this study within the larger landscape of UX approaches. Battarbee and Koskinen (2005) usefully summarised UX approaches in three categories: the measuring approach, the emphatic approach, and the pragmatist approach. The measurement approach focuses on the aspects of user experience that can be measured directly by physical reactions of bodies or by subjective reporting. The emphatic approach is based on developing a rich understanding of users' needs, desires, dreams, and motivations through various formative methods involving visual and textual data, and creative tasks in design phase. This approach aims to project future user experience and inspire designers rather than assess a current user experience with a system. The pragmatic approach, which is informed by pragmatist philosophy (Dewey 1934), provides a holistic view of user experience focusing on understanding interactions between users, technologies and environment as the indivisible constituents of experience. The studies employing pragmatic approach tends to be theoretical and does not provide practical guidance on design and evaluation. Rather, they focus on increasing an awareness of and sensitivity to the irreducibility and embodied nature of experience. This study is situated within the measurement approach. The focus is upon questionnaires used for evaluating user experience and/or subjective user satisfaction in conversational interfaces.

To understand what the current standardised questionnaires for evaluating conversational systems can offer in measuring user experience, we analysed the assessment items listed in the six main questionnaires and coded them according to their association with UX dimensions frequently assessed in the UX literature. This paper presents work focusing on understanding the extent in which UX-related items are present in these questionnaires. Our intention is to use the questionnaires' coverage of UX dimensions as a preliminary step towards assessing their suitability to measure user experience. Our assessment is based solely on the presence of UX-related items in the questionnaires. Therefore, what this study offers is an initial assessment of the potential suitability of the commonly used questionnaires to measure UX.

## 2. USER EXPERIENCE IN CONVERSATIONAL INTERFACES

McTear, Callejas and Griol (2016) define conversational interface as "*the technology that supports conversational interaction with virtual personal assistants by means of speech or other modalities*" (p. 11). They explain that the rising popularity of conversational interfaces has been facilitated by a renaissance in Artificial Intelligence, the development of powerful processors supporting deep learning algorithms, and advances in the Semantic Web making available large amount of knowledge online.

A conversational system comprises various modules including Automatic Speech Recognition, Natural Language Understanding, Dialogue Management, Natural Language Generation, and Speech Synthesis (López-Cózar et al. 2011). There are specific measures defined for each module. For example, the word error rate is used for Automatic Speech Recognition, and the rate of out of vocabulary words is used for Natural Language Generation (López-Cózar et al. 2011). In terms of UX, there is no specific module directly responsible for UX; rather, all system modules play a role in shaping UX. There is very little work on evaluating UX in conversational interfaces. One approach specifically focusing on assessing user experience of multimodal dialogue systems is SUXES proposed by Turunen et al. (2009). It is a complete evaluation procedure with four phases involving three questionnaires. Although the authors of SUXES provided their user experience questionnaire's constructs, the actual questionnaire statements were not presented, making their proposal difficult to evaluate.

There are many studies explicitly mentioning UX in conversational systems literature in which UX has been variably defined as: usability (Bijani et al. 2013; Tchankue et al. 2010), something going beyond usability (De Carolis et al. 2010), an aspect of usability (Soronen et al. 2009), user satisfaction (Goulati & Szostak 2011; Xu et al. 2013), and a combination of ease of use, overall feeling and user satisfaction (Wulf et al. 2014; Lee & Choi 2017). Therefore, in parallel to the lack of a clear understanding of UX in the field of HCI, there are large variations between the conceptions of UX in conversational systems as well. In general, the main tendency appears to be conceiving UX as a design and evaluation factor going beyond usability and closely associated with user satisfaction.

In the next section, we will introduce some major standardised questionnaires used in conversational interfaces literature. In addition, we will present some other frequently used non-standardised questionnaires.

## 3. STANDARDISED QUESTIONNAIRES

A standardised questionnaire involves an established procedure for collecting and presenting the measurement, and psychometric qualification (Lewis 2016). Standardised questionnaires provide higher levels of reliability (Sauro & Lewis 2009; Hornbæk & Law 2007) and facilitate an easier comparison between similar studies (Hornbæk 2006).

To determine a list of questionnaires for our assessment, we examined recent review studies (Dybkjaer etl al. 2004; Kühnel 2012; Larsen 2003; Lewis 2016; McTear, Callejas & Griol 2016; Wechsung 2014; Wechsung & Naumann 2008; Wechsung et al. 2012) and identified the major questionnaires for evaluating conversational interfaces. Then, we compiled a list including the standardised questionnaires on evaluating user experience or subjective user satisfaction in conversational interfaces. Our final list included the AttrakDiff, the Subjective Assessment of Speech System Interfaces (SASSI), the Speech User Interface Service Quality (SUISQ), the Mean Opinion Scale (MOS), the Paradigm for Dialogue Evaluation System (PARADISE), and the System Usability Scale (SUS). We have eliminated some other well-known questionnaires including the Software Usability Measurement Inventory (SUMI), the Questionnaire for User Interaction Satisfaction (QUIS), SUXES, TRINDI Tick List and DISC Dialogue Management Grid. Because, the SUMI and QUIS were focused heavily on graphical user interfaces, and the TRINDI, DISC, SUXES did not have a validated questionnaire. We have included the non-standardised PARADISE and SUS questionnaires into our final list as both of them have been widely used as measurement tools.

### 3.1 AttrakDiff

The AttrakDiff (Hassenzahl et al. 2003) is one of the most frequently used standardised questionnaires in the HCI field to measure hedonic qualities. Although it has an explicit focus on hedonic qualities, it also measures pragmatic qualities and overall appeal of a product. The Attrakdiff has a strong theoretical basis informed by Hassenzahl's (2003) model of user experience. The model proposes that a product can have two main qualities: hedonic and pragmatic. Hedonic qualities refer to the capacity of a product to "*support the achievement of 'be-goals,' such as 'being competent,' 'being related to others,' 'being special'*" (Hassenzahl et al. 2008, p. 473). Pragmatic qualities refer to the capacity of a product to "*support the achievement of 'do-goals,' such as 'making a telephone call,' 'finding a book in an online-bookstore' or 'setting-up a webpage'*" (Hassenzahl et al. 2008, p. 473). While hedonic qualities support stimulation, communicate identity and provoke memory, pragmatic qualities support instrumental and task-related features of a product, ensuring effective and efficient means to perform a task (Hassenzahl 2003). In addition to hedonic and pragmatic qualities, the AttrakDiff measures overall appeal of a product as a result of its hedonic and pragmatic qualities. The AttrakDiff contains 28 items in three categories: pragmatic quality, hedonic quality, and attractiveness. It is important to note that the theoretical model behind the AttrakDiff does not try to measure emotions such as fun, satisfaction, joy, or anger as they are considered as '*consequences of a cognitive appraisal process'* (Hassenzahl 2003, p. 483). As we will discuss later, AttrakDiff may need to be complemented by another measuring tool to capture affect-related dimensions of UX.

### 3.2 SASSI

The Subjective Assessment of Speech System Interfaces (SASSI) (Hone & Graham 2000) questionnaire is one of the most commonly used standardized measuring tools for assessing subjective satisfaction with speech-based interfaces (Larsen 2003). SASSI focuses on the speech input quality while excluding the speech output. Although it is a limitation, a wide-range of factors related to user experience are assessed by the SASSI's 34 items in six categories: system response accuracy, likeability, cognitive demand, annoyance, habitability, and speed (Hone & Graham 2000). Different from the other questionnaires, the SASSI also assesses habitability, which '*refers to the extent to which the user knows what to do and knows what the system is doing*' (Hone & Graham 2000, p. 300). Hone and Graham (2000) explain that speech systems needed a design quality similar to the visibility used in graphical user interfaces. Since the term visibility was not suitable for voice interfaces, they preferred to use the term habitability to assess the degree of a match between the user's conceptual model and the actual system and its behaviour.

### 3.3 SUISQ

The Speech User Interface Service Quality (SUISQ) questionnaire is a measuring instrument developed to assess service quality of speech interfaces (Polkosky 2005, 2008). The SUISQ has a total of 25 items in four categories validated by a principle component analysis (Polkosky 2005): user goal orientation, speech characteristics, verbosity, and customer service behaviour. Polkosky (2008) found that all four categories significantly correlated with customer satisfaction. Compared to the SASSI, SUISQ has five items in the speech characteristics category assessing speech output quality. One unique category of the SUISQ is the customer service behaviour that assesses the '*the extent to which the system's behaviour is similar to the expectations of human service providers'* (Polkosky, 2008, p.48). This category covers the aspects of system such as politeness, friendliness, and professional attitude. Polkosky's (2008) research shows that expectations associated with human conversations, customer service, and interpersonal interaction play a role in people's judgement of speech user interfaces.

### 3.4 MOS-X

The Mean Opinion Scale (MOS) (Schmidt-Nielsen 1995) is a widely used measurement tool for evaluating the quality of synthetically created speech. It is a Likert-style questionnaire assessing intelligibility and naturalness of the synthetic speech by seven items: (i) Global Impression, (ii) Listening Effort, (iii) Comprehension Problems, (iv) Speech Sound Articulation, (v) Pronunciation, (vi) Speaking Rate, and (vii) Voice Pleasantness. In order to assess a larger range of voice characteristics, MOS-Expanded (MOS-X) has been developed with additional eight assessment items: (i) Voice Quality, (ii) Emphasis, (iii) Rhythm, (iv) Intonation, (v) Trust, (vi) Confidence, (vii) Enthusiasm, and (viii) Persuasiveness (Polkosky & Lewis 2003). MOS-X with its fifteen items covering the aspects of naturalness, intelligibility, prosody, and social impression is a very valuable instrument for assessing synthetic voice and speech quality that are important constituents of user experience of conversational interfaces. However, MOS-X with its very specific focus on voice quality by itself is not sufficient to evaluate core usability and many other user experience dimensions of such systems.

### 3.5 PARADISE

The Paradigm for Dialogue Evaluation System (PARADISE) is a general framework involving a model for predicting user satisfaction and a user satisfaction survey. PARADISE's model is based on a weighted linear combination of task success measures, dialogue costs, and a user satisfaction survey with eight questions on usability aspects of interacting with a system. Despite being widely used, PARADISE has been criticised for not describing how user satisfaction is actually measured and the omission of psychometric validation of its user satisfaction survey (Hone & Graham 2001). Another limitation is that PARADISE's model is based on the assumption that minimising dialogue cost and maximising task success would maximise user satisfaction. Although dialogue cost and task success play a key role in improving the usability of a system, this formulation of user satisfaction is too reductive, ignoring various experiential and emotional aspects of user-system interactions.

### 3.6 SUS

The System Usability Scale (SUS) (Brooke 1996) is the most well-known and widely-used questionnaire to evaluate the usability of interactive systems. It is a very simple ten-item Likert scale assessing the perceived ease of use and learnability of using a system. Each questionnaire item corresponds to a statement expressed in a very general form by the perspective of a user. The non-specific formulation of the statements has allowed the SUS to be used in many different contexts to evaluate many different systems. Many variations of the SUS have replaced the term 'system' in the original statements with the terms 'website' or 'mobile app' to make them suitable for their own application domain. The SUS has been employed in evaluating many conversational systems such as (Hoque et al. 2013; DeVault et al. 2014).

### 3.7 Others

The TRINDI Tick-List (Bos et al. 1999) and DISC Dialogue Management Grid (Bernsen et al. 1999) are two similar non-standardised measuring tools focusing on evaluating conversational capabilities of a system. TRINDI with its seventeen assessment items covers a larger range of factors than DISC with its nine items. Their questions are not targeted to users; rather, they are formulated for system designers to be used as a heuristic assessment tool. Some example dialogue capabilities to be assessed include: "*Is the utterance interpretation sensitive to dialogue context?"* (TRINDI), "*Can the system deal with answers to questions that give different information than was actually requested?"* (TRINDI), and "*Can indirect speech acts be handled?* (DISC). All the assessment items of both TRINDI and DISC are associated with the instrumental aspects of conversational interfaces. Neither of them is a complete evaluation tool, but they provide a useful set of questions for system designers to think about and test the conversational competence of the systems they are designing.

## 4. ASSESSMENT OF THE QUESTIONNAIRES

In this section, we will first explain our method to assess the six questionnaires. Then, we will present the results with a table and six radar charts.

### 4.1 Method

To understand what these questionnaires can offer in measuring user experience, the first step was to identify the main dimensions of UX. However, as discussed earlier, there are no commonly accepted UX dimensions or factors. Therefore, we decided to use the most commonly assessed UX dimensions in literature as identified by the systematic review study of Bargas-Avila and Hornbæk (2011). In their study, they identified nine UX dimensions: generic UX, affect/emotion, enjoyment/fun, aesthetics/appeal, hedonic qualities, engagement/flow, motivation, enchantment, and frustration. As instrumental factors are considered part of user experience (Hassenzahl 2003; ISO 2010), we included

**Table 1**. A User Experience Assessment Scheme for Standardised Questionnaires

| **UX Dimensions**[a] | **Explanation** | **Sample Relevant Attributes** |
|---|---|---|
| **Generic UX** | There is no specific focus on any UX aspects. It refers to a general impression or sentiment of the overall use or experience with the system (Bargas-Avila & Hornbæk 2011). | General statements such as *"I think would like to use this system"* (Brooke 1996) or "*The system gave me a good feeling about being a customer of this business*." (Polkosky 2005) |
| **Affect/Emotion** | Affect refers to psychological states including emotions, feelings, impressions and moods. Emotion emerges as a result of interacting with a product or system (Turner 2017). Bargas-Avila and Hornbæk (2011) stated that affect and emotion are treated mostly synonyms in UX research as an explanation for placing them in the same dimension. | A person's feelings such as happy, pleased, satisfied, contended, hopeful, relaxed, stimulated, excited, frenzied, jittery, wide-awake, aroused, controlled, influenced, in control, awed, submissive, and guided (Mehrabian & Russell 1974). |
| **Enjoyment/Fun** | This dimension refers to playful interactions characterised by perceptions of pleasure and involvement (Webster et al. 1993). How enjoyable, fun, or playful using a system is the focus. | A person's feelings such as spontaneous, imaginative, creative, happiness, original, and innovative (Lavie & Tractinsky 2004). |
| **Aesthetics/Appeal** | This dimension refers to classical aesthetics emphasizing clean and orderly design, and expressive aesthetics associated with the qualities of creativity and novelty (Lavie & Tractinsky 2004). Physical and sensory features of a product or an interaction resulting in attractiveness or appeal (Hassenzahl 2001). | Product or interaction attributes such as pleasant, good, aesthetic, inviting, attractive, sympathetic, motivating, and desirable (Hassenzahl et al. 2008). |
| **Hedonic Quality** | This dimension refers to "*the product's perceived ability to support the achievement of 'be-goals,' such as 'being competent,' 'being related to others,' 'being special'*" (Hassenzahl et al. 2008, p. 473). Hedonic qualities support stimulation, communicate identity and provoke memory (Hassenzahl et al. 2008). | Product or interaction attributes such as interesting, costly, exciting, exclusive, impressive, original, and innovative (Hassenzahl et al. 2008) |
| **Engagement/Flow** | Engagement refers to *"a category of user experience characterized by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control*" (O'Brien & Toms 2008, p. 941). Flow refers to "*a state in which people are so involved in an activity that nothing else seems to matter; the experience itself is so enjoyable that people will do it even at great cost, for the sheer sake of doing it*" (Jackson & Marsh 1996, p. 4). | *Engagement*: A person's feelings, or product or interaction attributes such as challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, motivation, and perceived user control (O'Brien & Toms 2008).<br>*Flow*: A person's feelings such as control, attention focus, curiosity, and intrinsic interest (Jackson & Marsh 1996). |
| **Motivation** | This dimension refers to "*internal factors that impel action and to external factors that can act as inducements to action."* (Lock & Latham 2004, p.388). | Product or interaction attributes such as motivating, and discouraging (O'Brien & Toms 2008), or any expressions of rationale behind using or interacting with a product (Jacobson & Pirinen 2007). |
| **Enchantment** | This dimension refers to a state in which a sense of disorientation co-exists with heightened levels of perception and attention (McCarthy et al. 2006). | A person's feelings, or product or interaction attributes such as sensuousness, immersion, playfulness, paradox, openness, ambiguity, and transformational (McCarthy et al. 2006). |
| **Frustration** | This dimension refers to the disliked aspects of a product or an interaction (Blythe et al. 2006; Hone & Graham 2000). | Product or interaction attributes such as repetitive, boring, irritating and frustrating (Hone & Graham 2000). |
| **Pragmatic Quality**[b] | This dimension refers to "*the product's perceived ability to support the achievement of 'do-goals,' such as 'making a telephone call,' 'finding a book in an online-bookstore' or 'setting-up a webpage'*". (Hassenzahl et al. 2008, p. 473). Pragmatic qualities correspond to instrumental and task-related features of a product providing effective and efficient means to perform a task (Hassenzahl 2001). | Product or interaction attributes such as efficient, effective, supporting, useful, controllable (Hassenzahl 2001), easy to use, and error tolerant (Quesenbery 2014). |

[a] Adapted from Bargas-Avila and Hornbæk (2011)
[b] Added because instrumental and ergonomic aspects considered part of UX dimensions (Hassenzahl 2001) and ISO (2010)

pragmatic quality into our assessment scheme as an additional dimension. However, it is possible to exclude pragmatic quality from our final assessment to understand the non-instrumental UX dimensions of a questionnaire.

After determining the UX dimensions, we performed an extensive literature search to obtain relevant attributes characterising each UX dimension. We started with the references cited in Bargas-Avila and Hornbæk's (2011) study for determining a few attributes, and then used these attributes as a guide to compile a set of relevant attributes for each UX dimension. It is important to note that the set of attributes we identified are not complete; rather, they are representative and indicative.

The aim of having a set of attributes was to use them as a frame of reference to understand the overall scope and characteristics of a particular UX dimension. Table 1 shows the final assessment scheme with the UX dimensions, definitions, and corresponding lists of attributes. The next step involved a coding process with two researchers coming together and aligning their understandings of each UX dimension and their attributes. This was followed by the actual coding activity in which the two researchers performed coding of all the items in each questionnaire independently. After completing the coding, the researchers came together again to compare their assessments. In the cases of conflicting assessments involving differently coded items, the researchers worked together to agree on a final assessment for each item by discussing the rationale behind their initial coding. A few sample questionnaire items with their associated UX dimensions are listed in Table 2.

**Table 2**. Sample questionnaire items with their associated UX dimension(s)

| Questionnaire Items | UX Dimension(s) |
|---|---|
| I felt tense using the system.<br>(SASSI) | Affect/Emotion |
| The product is:<br>Discouraging ↔ Motivating<br>(AttrakDiff) | Motivation,<br>Engagement/Flow |
| The system's voice sounded like people I hear on the radio of television.<br>(SUISQ) | Hedonic Quality,<br>Aesthetics/Appeal |
| Did the voice appear to be trustworthy?<br>(MOS-X) | Hedonic Quality |
| From your current experience with using the system, do you think you'd use the system regularly when you are away from your desk?<br>(PARADISE) | Generic UX |
| I found the system unnecessarily complex.<br>(SUS) | Pragmatic Quality |

*Notes*. Double- or triple-coding of an item was possible because one item could be associated with multiple UX dimensions.

### 4.2 Results

The results of our assessment are available in Table 3 including the actual number of items associated with each UX dimension, and in Figure 1 showing radar charts of the questionnaires' coverage of UX dimensions. Overall, the top three questionnaires with the highest coverage of UX dimensions included the SASSI, AttrakDiff and SUISQ. The SASSI provided assessments items in almost all UX dimensions except for motivation and enchantment. It also provided the highest number of assessment items in three dimensions: affect/emotion (8 items), pragmatic quality (23 items), and frustration (5 items). It was also the only questionnaire with one assessment item in enjoyment/fun and five assessment items in frustration. The AttrakDiff comprised the highest number of items in assessing hedonic (14 items), and engagement/flow (16 items) aspects and second highest in aesthetic (7 items) aspects. It was the only questionnaire assessing motivation (1 item). One important dimension that the AttrakDiff lacked was affect/emotion where the other two questionnaires provided a larger coverage. The SUISQ provided assessment items in six categories with the highest number of items in the aesthetics/appeal (9 items) and in generic UX category (3 items).

The other three questionnaires, the MOS-X, SUS and PARADISE, have provided very little coverage of UX dimensions. Amongst the three, the MOS-X was the only questionnaire with the assessment items in hedonic (4 items) and aesthetics/appeal (6 items) category. The SUS and PARADISE provided majority of their assessment items in pragmatic quality (9 and 7 items respectively) while having a single item in generic UX and engagement/flow.

Across all questionnaires, enchantment, motivation and enjoyment/fun were the UX dimensions most neglected whereas the pragmatic quality dimension was the most commonly assessed. Another important missing dimension was affect/emotion which was only assessed by the SASSI with eight items and the SUISQ with three items. The dimension of enchantment was not assessed by any questionnaires at all. This is possibly due to the ambiguous and complex nature of the concept. No studies in Bargas-Avila & Hornbæk's (2011) systematic review reported how enchantment could be measured, suggesting enchantment is possibly a UX dimension more suitable for more qualitative interview-style approaches.

**Table 3.** The number of items associated with each UX dimension in the six questionnaires

| UX Dimensions | Questionnaires | | | | | |
|---|---|---|---|---|---|---|
| | **AttrakDiff** | **SASSI** | **SUISQ** | **MOS-X** | **PARADISE** | **SUS** |
| **Generic UX** | | 1 | 3 | | 1 | 1 |
| **Affect/Emotion** | | 8 | 3 | | | |
| **Enjoyment/Fun** | | 1 | | | | |
| **Aesthetics/Appeal** | 7 | 2 | 9 | 6 | | |
| **Hedonic Quality** | 14 | 5 | 6 | 4 | | |
| **Engagement/Flow** | 16 | 7 | 5 | 4 | 1 | 1 |
| **Motivation** | 1 | | | | | |
| **Enchantment** | | | | | | |
| **Frustration** | | 5 | | | | |
| **Pragmatic Quality** | 7 | 23 | 9 | 7 | 7 | 9 |
| **Total number of items** | 28 | 34 | 25 | 15 | 8 | 10 |

*Notes*. The sum of the items per questionnaire is greater than a questionnaire's actual total number of items, because one item can be associated with several UX dimensions.

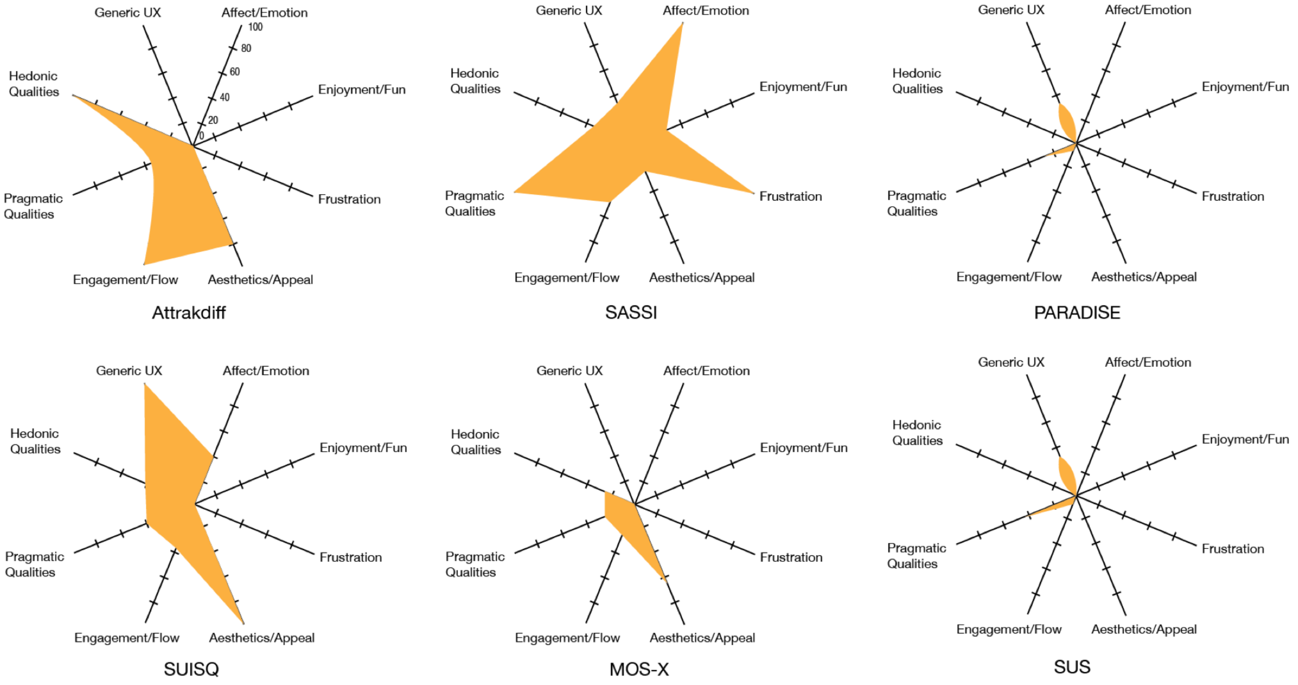

**Figure 1**. The radar charts for the six questionnaires' coverage of UX Dimensions.

*Notes*. The dimension of Enchantment is excluded because none of the questionnaires contained any relevant items. The values on each UX dimension have been normalised to a range between 0 and 100.

## 5. DISCUSSION

The results indicate that there is no questionnaire providing sufficient coverage across all UX dimensions. It is understandable that having assessment items for every dimension would require placing a lot of items in a single questionnaire, making it a very long and a less practical questionnaire, and a very daunting task for users to respond. Bradley and Lang (1994), the authors of the Self-Assessment Manikin (SAM), has raised a similar point on practicality. They explained that the Semantic Differential Scale (SDS) devised by Mehrabian and Russell (1974) has 18 bipolar adjective pairs to be rated along a 9-point scale, requiring a heavy investment of time and effort (Bradley & Lang 1994). Their study showed that the SAM with its much simpler three major affective dimensions correlated highly with ratings of the SDS, suggesting using a simpler assessment method could effectively and efficiently provide an understanding of the overall affective states of users. Similarly, it was argued that expressive aesthetics and hedonic quality are strongly overlapping constructs (Hassenzahl & Monk 2010), and a further consolidation of these constructs is attainable (Bargas-Avila & Hornbæk, 2011).

### 5.1 Recommended Questionnaires

In terms of selecting which questionnaire(s) to use, it is possible to recommend a few alternatives based on the needs of different cases. Our main rationale behind recommending a particular questionnaire is based on a questionnaire's coverage of the relevant UX dimension. The questionnaire with the highest number of items associated with a UX dimension is recommended. Table 4 shows the recommended questionnaires to use according to the desired UX dimension and the preference on using single or multiple questionnaires. While the second column includes the questionnaires with the highest coverage of the relevant UX dimension assessed by this study, the third column offers some other widely used and sufficiently generic questionnaires to complement the single questionnaire:

- Self-Assessment Manikin (SAM) to measure arousal, pleasure, and dominance on non-verbal scales (Bradley & Lang 1994);
- Adult Playfulness Scale (APS) to measure adults' playfulness (Glynn & Webster 1994);
- Flow State Scale (FSS) to measure engagement and flow experience (Jackson & Marsh 1996);
- Intrinsic Motivation Inventory (IMI) to measure intrinsic motivation (McAuley et al. 1989).

While the items in the SAM, APS, and FSS can be used without much modification, the IMI's items need to be modified slightly to fit specific activities in focus.

To illustrate how our recommendation can be understood, if a study is interested in the affective states of people when interacting with a conversational agent, the SASSI would be recommended as a single questionnaire of choice; or, the SASSI and the SAM as a combination of multiple questionnaires. Here, the reason of using the SASSI as a single questionnaire is that it has the highest number of items in affect/emotion, and as well, many other items assessing other major UX dimensions. Therefore, in addition to having a large coverage in a desired UX dimension, the SASSI provides an assessment of other potentially relevant UX dimensions.

If feasible, using multiple questionnaires could be very useful. For example, in a study to understand the effects of an individual's motivational orientation and particular product attributes on the perceived value of interactive products, Hassenzahl et al. (2008) used the SAM's affective measurements to complement the AttrakDiff's assessment of hedonic, aesthetics, and pragmatic qualities. In this respect, their strategy was to obtain a more complete assessment that covers a larger range of UX dimensions rather than focusing on a particular UX dimension. Moreover, in some cases, having a questionnaire with a large number of items in one dimension may not be needed as the same assessment might be performed with a fewer number of items by another questionnaire. Practicality and putting less demand on users are

**Table 4**. Some recommended questionnaires when a single questionnaire or a combination of questionnaires is to be used

| UX dimension to focus on | Using single questionnaire | Using multiple questionnaires |
|---|---|---|
| Affect/Emotion | SASSI | SASSI + SAM |
| Enjoyment/Fun | SASSI | SASSI + APS |
| Aesthetics/Appeal | AttrakDiff | AttrakDiff + SUISQ |
| Hedonic Quality | AttrakDiff | AttrakDiff + SUISQ |
| Engagement/Flow | AttrakDiff | AttrakDiff + FSS |
| Motivation | AttrakDiff | AttrakDiff + IMI |
| Enchantment | - | - |
| Frustration | SASSI | SASSI + SAM |
| Pragmatic Quality | SASSI | SASSI + TRINDI |
| Overall UX* | SASSI | SASSI + AttrakDiff |

* Overall UX corresponds to the combination of all other UX dimensions.

important concerns when conducting user testing and evaluation studies. Therefore, depending on a study's needs and focus, the questionnaire with fewer items on a desired UX dimension might be preferred if the validity of the shorter questionnaire for the desired factor is available.

If what is feasible is to use only one questionnaire for assessing the overall UX, then the SASSI is potentially the most suitable choice with its assessment items associated with a large range of UX dimensions. However, coupling it with the AttrakDiff would provide even a larger coverage of the UX dimensions with an increased focus on hedonic, engagement and aesthetics qualities. Depending on the availability of resources, these two could be further complemented by the MOS-X with its items assessing voice- and speech-focused aspects of user experience.

In terms of assessing mainly instrumental aspects of a conversational interface, again the SASSI would be recommended as a single measurement tool. However, the 10-item SUS with its effective and efficient assessment of usability could be a more economical and practical alternative. In addition, the TRINDI Tick List and DISC with their specific set of items focusing on conversational competence would provide useful ways to assess some core functionalities needed in a typical conversational system.

### 5.2 Combining Subscales of Different Standardised Questionnaires

Instead of using multiple questionnaires separately to obtain a more complete assessment of UX, combining the subscales of some of the standardised questionnaires into a single questionnaire might be an alternative solution. For example, Polkosky and Lewis (2003) worked on the MOS questionnaire to improve its reliability and added new items to the questionnaire to extend its coverage. Similarly, Lewis and Hardzinski (2015), worked on the SUISQ questionnaire and produced a shortened version of the SUISQ referred to as the SUISQ-R. In both cases, the authors retested the validity and reliability of the new questionnaires, and they obtained comparable results with the original questionnaires. Therefore, a similar approach can be employed across the subscales of the standardised questionnaires presented in this paper to construct a new questionnaire with a larger coverage of UX dimensions. However, this alternative has potentially some problems in relation to internal reliability, content validity, and construct validity. Although the subscales of the original questionnaires are reliable and validated, they need to be revalidated and their reliability needs to be retested as part of a new questionnaire. In addition, the subscales of these questionnaires are not fully aligned with the UX dimensions identified in this paper. In some cases, the questionnaires' subscales include a few items that are not relevant to measuring the desired UX dimension, and using a subscale with some irrelevant items can violate the content validity (Haynes, Richard & Kubany 1995). There are also other difficulties associated with the different types of scales employed by different questionnaires. For example, the AttrakDiff uses a semantic differential scale, but the SASSI and SUISQ use a Likert scale. Therefore, some kind of scale conversion is required. In addition, the differences between the degrees of specificity of semantic differential scale items and Likert scale items can be confusing for the questionnaire participants. As a result, there are some challenges against combining the subscales of the current standardised questionnaires. However, using these questionnaires' subscales as a starting point can shorten the development time of a new standardised questionnaire, which requires a substantial amount of work (Lewis & Hardzinski 2015).

### 5.3 Further UX Considerations and Dimensions

An evaluation consideration that can potentially gain more importance in relation to UX is the personality and attitude of a conversational agent. In this respect, the SUISQ is unique with its eight items from a customer service behaviour perspective. A few sample SUISQ statements include:

- The system seemed professional in its speaking style.
- The system seemed polite.
- The system used terms I am familiar with.

We believe UX measurement instruments for conversational interfaces can benefit from engaging with social-communicative theory and studies focusing on interpersonal communication and customer service behaviour. Polkosky's (2008) research provides a useful starting point.

Another potentially important concept as a specific UX dimension for conversational interfaces is the concept of habitability defined as '*the extent to which the user knows what to do and knows what the system is doing*' (Hone & Graham 2000, p. 300). Although it has not received enough attention in the field of spoken language systems, with the recent increased interest towards speech interfaces, habitability will likely to play an instrumental role in design and evaluation of such systems. Because, it allows the creation of a fundamental design construct that can define the "visibility" in voice user interfaces. Hone and Baber (2001) proposed a conceptualisation of habitability in relation to semantic, syntactic, lexical, dialogue, and recognition constraints operating over user utterances. While their proposal is useful, we believe the scope of habitability to extend beyond the notions of constraints and visibility, and include some other factors such as familiarity, emotional connection, and sense of agency.

Our study has focused on evaluating user experience from a measurement-oriented perspective. However, there are also other more emphatic and pragmatic user experience evaluation perspectives involving qualitative methods such as conversational analysis (Porcheron et al. 2018) and in-depth interviews with users (Luger & Sellen 2016). While conversational analysis studies can allow us to understand the evolution of user experience over a period of time in an indirect way without actually asking any questions to users, in-depth interviews can provide some richer data to understand the various factors shaping users' experience with conversational interfaces.

## 6. CONCLUSION

In this paper, we have briefly reviewed UX approaches in the field of HCI and presented varied definitions of UX in conversational systems. We used the most frequently assessed UX dimensions and their relevant attributes to obtain an assessment scheme. Then, we employed the assessment scheme to understand the questionnaires' coverage of UX dimensions as a preliminary step towards assessing their suitability to measure UX. We found that (i) four questionnaires included assessment items, in varying extents, to measure hedonic, aesthetic and pragmatic dimensions of UX; (ii) two questionnaires assessed affect, and one assessed frustration dimension; and (iii) enchantment, playfulness and motivation dimensions have not been covered sufficiently by any questionnaires. Our assessment has suggested that the SASSI with its large coverage spanning over eight UX dimensions would be a suitable questionnaire for overall UX measurement. Using multiple questionnaires may prove useful for obtaining a more complete measurement of user experience or improving the assessment of a particular UX dimension. Future work involves assessing the actual performances of these questionnaires in measuring the relevant UX dimensions.

## 7. REFERENCES

Bargas-Avila, J.A. and Hornbæk, K., 2011. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. *In Proc. Conference on Human Factors in Computing Systems*. ACM.

Battarbee, K. and Koskinen, I., 2005. Co-experience: user experience as interaction. *CoDesign*, *1*(1), pp. 5-18.

Bernsen, N.O., Dybkjær, L. and Heid, U., 1999. Current Practice in the Development and Evaluation of Spoken Language Dialogue Systems. In *Sixth European Conference on Speech Communication and Technology*.

Bijani, C., White, B.K. and Vilrokx, M., 2013. Giving voice to enterprise mobile applications. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services* (pp. 428-433). ACM.

Blythe, M., Reid, J., Wright, P., & Geelhoed, E. (2006). Interdisciplinary criticism: analysing the experience of riot! a location-sensitive digital narrative. *Behaviour & Information Technology*, 25(2), 127-139.

Bos, J., Larsson, S., Lewin, I., Matheson, C. and Milward, D., 1999. Survey of existing interactive systems. *Trindi (Task Oriented Instructional Dialogue) report*, (D1), p.3.

Bradley, M.M. and Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), pp. 49-59.

Brooke, J., 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), pp. 4-7.

De Carolis, B., Mazzotta, I., Novielli, N. and Pizzutilo, S., 2010. Social robots and ECAs for accessing smart environments services. In *Proceedings of the International Conference on Advanced Visual Interfaces* (pp. 275-278). ACM.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M. and Lucas, G., 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* (pp. 1061-1068).

Dybkjaer, L., Bernsen, N.O. and Minker, W., 2004. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, *43*(1-2), pp. 33-54.

Glynn, M.A. and Webster, J., 1992. The adult playfulness scale: An initial assessment. *Psychological reports*, *71*(1), pp. 83-103.

Goulati, A. and Szostak, D., 2011, August. User experience in speech recognition of navigation devices: an assessment. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (pp. 517-520). ACM.

Hassenzahl, M., 2001. The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction*, *13*(4), pp. 481-499.

Hassenzahl, M., 2003. The thing and I: understanding the relationship between user and product. In *Funology* (pp. 31-42). Springer Netherlands.

Hassenzahl, M., Burmester, M. and Koller, F., 2003. AttrakDiff: A questionnaire to measure perceived hedonic and pragmatic quality. In *Mensch & Computer* (pp. 187-196).

Hassenzahl, M., Schöbel, M. and Trautmann, T., 2008. How motivational orientation influences the evaluation and choice of hedonic and pragmatic interactive products: The role of regulatory focus. *Interacting with Computers*, *20*(4-5), pp. 473-479.

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

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological assessment*, 7(3), 238.

Hone, K.S. and Graham, R., 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, *6*(3-4), pp. 287-303.

Hone, K.S. and Baber, C., 2001. Designing habitable dialogues for speech-based interaction with computers. *International Journal of Human-Computer Studies*, *54*(4), pp. 637-662.

Hone, K.S. and Graham, R., 2001. Subjective assessment of speech-system interface usability. In *Seventh European Conference on Speech Communication and Technology*.

Hoque, M.E., Courgeon, M., Martin, J.C., Mutlu, B. and Picard, R.W., 2013, September. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (pp. 697-706). ACM.

Hornbæk, K., 2006. Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies*, *64*(2), pp. 79-102.

Hornbæk, K. and Law, E.L.C., 2007. Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 617-626). ACM.

Jacobson, S. and Pirinen, A., 2007. Disabled persons as lead users in the domestic environment. In *Proceedings of the 2007 conference on Designing pleasurable products and interfaces* (pp. 158-167). ACM.

Jackson, S.A. and Marsh, H.W., 1996. Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of sport and exercise psychology*, *18*(1), pp. 17-35.

John, D., 1934. Art as experience. *New York: Minton, Balch, and Company*.

Kühnel, C., 2012. *Quantifying quality aspects of multimodal interactive systems*. Springer Science & Business Media.

Larsen, L.B., 2003. Assessment of spoken dialogue system usability-what are we really measuring? In *Eighth European Conference on Speech Communication and Technology*.

Lavie, T. and Tractinsky, N., 2004. Assessing dimensions of perceived visual aesthetics of web sites. *International journal of human-computer studies*, *60*(3), pp. 269-298.

Law, E.L.C., Roto, V., Hassenzahl, M., Vermeeren, A.P. and Kort, J., 2009. Understanding, scoping and defining user experience: a survey approach. In *Proc. Conference on Human Factors in Computing systems* (pp. 719-728). ACM.

Lee, S. and Choi, J., 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies*, *103*, pp. 95-105.

Lewis, J.R., 2016. Standardized Questionnaires for Voice Interaction Design. *Voice Interaction Design*, *1*(1).

Lewis, J. R., & Hardzinski, M. L. (2015). Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire. *International Journal of Speech Technology*, 18(3), 479-487.

Locke, E.A. and Latham, G.P., 2004. What should we do about motivation theory? Six recommendations for the twenty-first century. Academy of management review, 29(3), pp. 388-403.

López-Cózar, R., Callejas, Z., Espejo, G. and Griol, D., 2011. Enhancement of Conversational Agents by Means of Multimodal Interaction. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices: Techniques and Effective Practices*, pp. 223-252.

Luger, E. and Sellen, A., 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proc. Conference on Human Factors in Computing systems* (pp. 5286-5297).

McAuley, E., Duncan, T. and Tammen, V.V., 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport*, *60*(1), pp. 48-58.

McCarthy, J., Wright, P., Wallace, J. and Dearden, A., 2006. The experience of enchantment in human–computer interaction. *Personal and ubiquitous computing*, *10*(6), pp. 369-378.

McTear, M., Callejas, Z. and Griol, D., 2016. *The conversational interface*. Springer.

Mehrabian, A., & Russell, J. A., 1974. *An approach to environmental psychology*. the MIT Press.

O'Brien, H.L. and Toms, E.G., 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the Association for Information Science and Technology*, *59*(6), pp. 938-955.

Polkosky, M. D., 2005. *Toward a social-cognitive psychology of speech technology: Affective responses to speech-based e-service*. Unpublished doctoral dissertation. University of South Florida.

Polkosky, M. D. (2008). Machines as mediators: The challenge of technology for interpersonal communication theory and research. In E. Konjin (Ed.), *Mediated interpersonal communication* (pp. 34–57). New York: Routledge.

Polkosky, M.D. and Lewis, J.R., 2003. Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology*, *6*(2), pp. 161-182.

Porcheron, M., Fischer, J.E., Reeves, S. and Sharples, S., 2018. Voice Interfaces in Everyday Life. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'18)*.

Quesenbery, W., 2014. The five dimensions of usability. In *Content and complexity* (pp. 93-114). Routledge.

Sauro, J. and Lewis, J.R., 2009. Correlations among prototypical usability metrics: evidence for the construct of usability. In *Proc. Conference on Human Factors in Computing systems* (pp. 1609-1618). ACM.

Schmidt-Nielsen, A. (1995). Intelligibility and acceptability testing for speech technology. In A. Syrdal, R. Bennett, and S. Greenspan (Eds.), *Applied Speech Technology*. Boca Raton: CRC Press.

Soronen, H., Pakarinen, S., Hansen, M., Turunen, M., Hakulinen, J., Hella, J., Rajaniemi, J.P., Melto, A. and Laivo, T., 2009, September. User experience of speech-controlled media center for physically disabled users. In *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era* (pp. 2-5). ACM.

ISO 9241-210:2010. Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.

Tchankue, P., Vogts, D. and Wesson, J., 2010. Design and evaluation of a multimodal interface for in-car communication systems. In *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists* (pp. 314-321). ACM.

Turunen, M., Hakulinen, J., Melto, A., Heimonen, T., Laivo, T. and Hella, J., 2009. SUXES-user experience evaluation method for spoken and multimodal interaction. In *Tenth Annual Conference of the International Speech Communication Association*.

Turner, P., 2017. Aesthetics. In *A Psychology of User Experience* (pp. 109-130). Springer.

Webster, J., Trevino, L.K. and Ryan, L., 1993. The dimensionality and correlates of flow in human-computer interactions. *Computers in human behavior*, *9*(4), pp. 411-426.

Wechsung, I., 2014. *An evaluation framework for multimodal interaction.* Springer International. doi, 10, pp. 978-3.

Wechsung, I., Engelbrecht, K.P., Kühnel, C., Möller, S. and Weiss, B., 2012. Measuring the Quality of Service and Quality of Experience of multimodal human–machine interaction. *Journal on Multimodal User Interfaces,* 6(1-2), pp. 73-85.

Wechsung, I. and Naumann, A.B., 2008. Evaluation methods for multimodal systems: A comparison of standardized usability questionnaires. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems* (pp. 276-284). Springer, Berlin, Heidelberg.

Wulf, L., Garschall, M., Himmelsbach, J. and Tscheligi, M., 2014. Hands free-care free: elderly people taking advantage of speech-only interaction. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (pp. 203-206). ACM.

Xu, Q., Li, L. and Wang, G., 2013. Designing engagement-aware agents for multiparty conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2233-2242). ACM.