Construction of UEQ+ Scales for Voice Quality

Measuring User Experience Quality of Voice Interaction

Andreas M. Klein
Faculty of Technology
University of Applied Sciences
Emden/Leer
Emden, Germany
andreas.klein@hs-emden-leer.de

Andreas Hinderks
Department of Computer
Languages and Systems
University of Seville
Seville, Spain
andreas.hinderks@iwt2.org

Martin Schrepp
SAP SE
Walldorf, Germany
martin.schrepp@sap.de

Jörg Thomaschewski
Faculty of Technology
University of Applied Sciences Emden/Leer
Emden, Germany
joerg.thomaschewski@hs-emden-leer.de

ABSTRACT:

The UEQ+ is a modular framework for the construction of UX questionnaires. The researcher can pick those scales that fit his or her research question from a list of 16 available UX scales. Currently, no UEQ+ scales are available to allow measuring the quality of voice interactions. Given that this type of interaction is increasingly essential for the usage of digital products, this is a severe limitation of the possible products and usage scenarios that can be evaluated using the UEQ+. We describe in this paper the construction of three specific scales to measure the UX of voice interactions. Besides, we discuss how these new scales can be combined with existing UEQ+ scales in evaluation projects.

KEYWORDS :

User Experience, Usability, Voice Systems, Voice Interaction, Voice User Interfaces, Measurement, estionnaires, UX, VUI

1 INTRODUCTION

The impression of a user about the user experience (UX) of a product results from his or her perception of many distinct quality aspects, for example, efficiency of use, stimulation, trust, or visual aesthetics. The importance of such quality aspects for the UX impression varies between products supporting different tasks and use cases [1]. For example, intuitive use is mandatory for an infrequently used self-service (users forget how to use it between two usage points). Simultaneously, an unnecessary click does not hurt much, as efficiency is not so crucial. For a very often-used business application, intuitive use is not essential (some learning is accepted). Due to the high usage frequency, each unnecessary click hurts much, so high efficiency is a key requirement. The many UX quality aspects and their varying importance for different products caused the creation of many different UX questionnaires. Each of these questionnaires realizes by its selection of scales a different set of measured UX quality aspects and thus fits to a certain group of products [2]. Of course, none of these questionnaires contain all UX quality aspects discussed in research literature, since this would increase the length of the questionnaire above any reasonable limit. For a UX researcher evaluating a concrete product, this can be an issue. If he or she has a clear view of which UX aspects are important and should be thus measured in an evaluation project, it can easily happen that none of the published UX questionnaires really fit these requirements. Sometimes, it is possible to combine several UX questionnaires to cover all relevant aspects, but this also causes practical problems since different questionnaires often have different item and answer formats [3, 4]. The UEQ+ [3, 4] is a modular framework that tries to address this issue. It consists of 16 scales that can be combined to form a concrete questionnaire. Thus, the researcher can decide which of these scales are important for the product that should be evaluated. Then he or she can simply pick those scales and combine them for the evaluation. For a detailed description of the idea behind UEQ+ and scale construction please refer to [3, 4]. The material to conduct a study with the UEQ+ can be found free

of charge on https://ueqplus.ueq-research.org/. Currently the UEQ+ framework does not provide scales that measure the quality of voice interaction. But in recent years voice and speech recognition software has become the leading-edge interface technology for a wide range of applications, for example in healthcare, automotive, authentication and identification, voice commerce and customer service, and smart home sectors [5]. Current studies expect that global sales of voice and speech recognition software will increase from about $1.3 billion in 2019 to nearly $7 billion in 2025 [5]. Thus, an increasingly important category of products cannot be evaluated by questionnaires built with the UEQ+ framework, for example popular voice assistants (VAs) such as Google Assistant (Google), Siri (Apple) or Alexa (Amazon). We try to fill this gap by constructing new UEQ+ scales that cover the UX aspects associated with voice interactions. This paper describes the scale construction and how these new scales can be used in the UEQ+ framework [3, 4] to measure the UX of systems based on voice interaction. There are a few questionnaires [6, 7, 8, 9] that measure the usability of voice systems. These questionnaires concentrate solely on the task-related aspects of an interaction and ignore non-task related or hedonic aspects of voice interactions. In addition, they mix UX aspects of voice interaction with other more general UX aspects and cannot, therefore, simply be reused within the UEQ+ framework.

2 THE UEQ+ FRAMEWORK

The UEQ+ is not a UX questionnaire in the sense that it can directly by applied to measure the UX of a specific product. It is a modular catalogue with 16 UX scales that can be combined to form a UX questionnaire. The name UEQ+ was chosen because the six scales of the UEQ (*Attractiveness*, *Efficiency*, *Perspicuity*, *Dependability*, *Stimulation*, *Novelty*) were used as a starting point. Besides these six scales several others were designed and added: *Trust* [10], *Haptics* and *Acoustics* [11], or *Aesthetics*, *Adaptability*, *Usefulness*, *Intuitive Use*, *Value*, *Trustworthiness of Content* and *Content Quality* [3,4]. The UEQ+ and an Excel tool for evaluation are available free of charge at https://ueqplus.ueq-research.org/. Since UEQ+ scales can be combined arbitrarily, a special modular scale format is used. Each scale consists of a short introductory sentence (this sentence sets the context for the items), followed by four items in the form of a semantic differential with a seven-point Likert-scale. In addition, the importance of the UX aspect represented by the scale is asked directly below the scale. As an example, we show the scale *Efficiency* [4]:

To achieve my goals, I consider the product as

slow	o o o o o o o	*fast*
inefficient	o o o o o o o	*efficient*
impractical	o o o o o o o	*practical*
cluttered	o o o o o o o	*organized*

The product property described by these terms is for me

completely irrelevant	o o o o o o o	*highly relevant*

The scales of the UEQ+ that are selected for a specific situation are simply displayed one below the other. In a product-related questionnaire only a limited number of scales (for example 6 scales) should be used in order to keep the effort required to complete the questionnaire manageable. Further information on the selection of scales can be found in the UEQ+ handbook [12]. The UEQ+ scales can be grouped into three different types:

Scale type 1 describes user interaction with the product. For example, the user holds the product in his hand (*Haptics*) or the product emits sounds during use (*Acoustics*). Aesthetic plays a role when interacting with graphical user interfaces (GUI), for example, whether the user interface appears nice and appealing.

Scale type 2 summarizes fundamental and psychological qualities. While applying the product, the general user needs are to be captured here. Examples are *Perspicuity*, whether the product is easy to understand and learn, or *Efficiency*, whether its goals can be achieved with minimal effort.

Scale type 3 addresses specific needs of utilization and the resulting consequences. For example, *Attractiveness*, whether the user evaluates the product positively or rejects it summarily. Additional subjective impressions are *Trust*, whether the user places its input data in safe hands or *Value*, whether the product is professional and of high quality.

If there is interaction with specific product types, scale type 1 is selected, whereas scale types 2 and 3 can be used independently of the interaction, this is, they can be combined with scale type 1 as desired. For Voice User Interfaces (VUIs), scale type 1 lacks the appropriate scales, and this article aims to close this gap.

3 ASPECTS OF VOICE INTERACTION

Depending on use cases, different additional UX criteria may be relevant for voice interaction. According to an American study [13], examples of use cases for smart speakers are found primarily in the areas of information (e.g. weather reports) and entertainment (listening to streaming music services). What Siri and Alexa are for consumers, SAP CoPilot will be for business users [14]. Key features of SAP CoPilot include natural language communication using a dialogue-oriented user interface and support for business contexts and machine learning based on predefined business rules [14]. Hassenzahl and Tractinsky [15] identify three components that significantly influence UX in human-computer interaction (HCI): first, the user with expectation and motivation; second, the system with characteristics such as functionality and purpose; and third, the context with the organizational/social environment and the purpose of the activity. Cohen, Giangola, Balogh [16] describe three elements (Goals/Context, User and Application) has to be understood, to define criteria for measuring quality and improvements in the design process of a VUI application. In VA applications, the user expects a natural and trustful interaction, that the system fulfils the user's intention and that the context recognizes the user's intent without particular formulations. From these considerations, three UX aspects can be derived that are significant for speech interaction [17].

Response behaviour: Users expect that a voice system communicates like a human conversationalist. Thus, responses should be respectful, patient, polite, and trustworthy.

Response quality: The responses of the voice system cover the user's information needs. Thus, answers are perceived as clear, distinct, and up-to-date; the queries match the context; and the user's intention is fulfilled.

Comprehensibility: The user has the impression that the VA correctly understands his or her instructions and questions using natural language. The intention of the user is recognized without forcing him or her to use an unnatural way of speaking.

4 STUDY FOR SCALE CONSTRUCTION

The empirical study described below shows the construction of the scales that represent the three UX aspects described above. Therefore, an online survey was conducted with a German-language questionnaire containing the scales with a selection of bipolar items in the future called *candidate items*.

4.1 Participants

The online questionnaire was sent to several email distribution lists of students and members of the University of Applied Sciences in Emden/Leer (Germany). A total of 96 persons participated voluntarily. The average age of the participants (59 male, 35 female, 2 no answer) was 35 years (SD 12).

4.2 Material

We created a selection of candidate items for each of the UX aspects described above for VAs. The concrete items can be found in the tables of the next section. The online questionnaire contains all candidate items listed one below the other after an introductory sentence that sets the appropriate context.

4.3 Procedure

The survey by online questionnaires took place between 7 and 24 January 2020. The participants were advised in the introductory email not to continue the survey if they had no experience with voice interaction. In the beginning, when querying for the socio-demographic data, the participants could choose which VA to use. Then the scales followed in the order *Response behaviour*, *Response quality* and *Comprehensibility* with the corresponding candidate items in the order shown in Tables 1 to 3. The study was done in the German language. Detailed data analysis and screenshots of the online study pages are available in the research protocol [18].

4.4 Results

The following VAs were rated by the participants (number of participants that have chosen this system in brackets): *Alexa* (35), *Siri* (27), *Google Assistant* (26), *Others* (8).

The factorial analysis of all three candidate item sets shows (Kaiser-Guttman criteria and analysis of the scree plot) that a single factor represents the data sufficiently well (see [18]). In addition, a common analysis of all 30 items together with principal component analysis (Varimax rotation) confirmed the assumption of three factors. We show in the following the items used and the loadings on the corresponding factor, which is the basis for the selection of the four items that represent the scale.

Table 1: Set of candidate items for response behaviour[1]

No.	Items		Loadings
1	technical	human	0.66
2	artificial	natural	0.80
3	unfamiliar	familiar	0.66
4	unusual	usual	0.25
5	slow	fast	0.48
6	unpleasant	pleasant	0.75
7	unlikeable	likable	0.81
8	unfriendly	friendly	0.66
9	boring	entertaining	0.68

Table 1 shows the items and loadings for the factor corresponding to the scale *Response behaviour*. Items 2, 6, 7, and 9 were selected with an introducing sentence as follows:

In my opinion the response behaviour of the voice assistant is

artificial o o o o o o o *natural*
unpleasant o o o o o o o *pleasant*
unlikable o o o o o o o *likable*
boring o o o o o o o *entertaining*

Table 2: Set of candidate items for response quality[2]

No.	Items		Loadings
1	incomprehensible	understandable	0.14
2	illogical	logical	0.55
3	inappropriate	suitable	0.74
4	useless	useful	0.76
5	not helpful	helpful	0.82
6	laborious	simple	0.42
7	uninteresting	interesting	0.41
8	unintelligent	intelligent	0.61
9	unclear	clear	0.47
10	indistinct	exacting	0.53
11	outdated	current	0.49

[1] Response behaviour (German original items): 1. '*technisch/menschlich*', 2. '*künstlich/natürlich*', 3. '*fremd/vertraut*', 4. '*ungewöhnlich/gewöhnlich*', 5. '*langsam/schnell*', 6. '*unangenehm/angenehm*', 7. '*unsympathisch/sympathisch*', 8. '*unfreundlich/freundlich*', 9. '*langweilig/unterhaltsam*'.

[2] Response quality (German original items): 1. '*unverständlich/verständlich*', 2. '*unlogisch/logisch*', 3. '*unpassend/passend*', 4. '*nutzlos/nützlich*', 5. '*nicht hilfreich/hilfreich*', 6. '*umständlich/einfach*', 7. '*uninteressant/interessant*', 8. '*unintelligent/intelligent*', 9. '*unklar/klar*', 10. '*undeutlich/deutlich*', 11. '*veraltet/aktuell*'.

Table 2 shows all items and loadings for the scale *Response quality*. The highest loadings are found in items 3, 4, 5, and 8. The introductory sentence is as follows:

The answers and questions asked by the voice assistant are

inappropriate	o o o o o o o	*suitable*
useless	o o o o o o o	*useful*
not helpful	o o o o o o o	*helpful*
unintelligent	o o o o o o o	*intelligent*

Table 3: Set of candidate items for comprehensibility[3]

No.	Items		Loadings
1	complicated	simple	0.84
2	inaccurate	accurate	0.78
3	nonsensical	apt	0.72
4	unambiguous	ambiguous	0.82
5	illogical	logical	0.75
6	incomprehensible	understandable	0.78
7	unexpected	expected	0.68
8	unclear	clear	0.79
9	enigmatic	explainable	0.75
10	difficult	easy	0.72

Table 3 shows the corresponding results for the scale *Comprehensibility*. The three highest loadings are found in items 1, 4, and 8. Item 8 was not selected because it showed an overlap with item 9 of Response quality (see table 2). Items 2 and 6 with the identical value of 0.78 move up. Since item 6 overlaps with items of the existing scales *Perspicuity* and *Quality of Content* of the UEQ+, it is replaced by item 9 with a slightly lower load. The focus here is on speech comprehensibility in the exemplary sense of: "*Does the VA give enigmatic answers because it does not understand me?*". The UEQ+ scales *Perspicuity* and *Quality of Content* represent UX aspects for voice assistance systems that are prospectively used often in combination with the new scales, as described in the following section. The final scale is as follows:

In my opinion the voice assistant has understood my voice commands

complicated	o o o o o o o	*simple*
unambiguous	o o o o o o o	*ambiguous*
inaccurate	o o o o o o o	*accurate*
enigmatic	o o o o o o o	*explainable*

5 USING SCALES IN THE UEQ+ FRAMEWORK

The selection of relevant scales for creating a product-related questionnaire depends on various sources of information. Winter, Hinderks, Schrepp and Thomaschewski [1] recommend that product-specific UX aspects should be considered first to be followed by other criteria. These can also be UX aspects that are essential for marketing and product placement but not vital for the user. Further information on the scale selection and creation of a product-related questionnaire can be found in the UEQ+ handbook [12].

We conclude with two examples showing how the new scales for voice interaction can be combined with other UEQ+ scales for concrete evaluation projects. We assume that we want to evaluate, for example, the application of smart home VAs used for general tasks, such as asking questions or online-shopping. In this case, the three voice scales can be combined with scales highly relevant for the information search in the web [1], that is, *Perspicuity*, *Trust* and *Quality of Content*. If we want to evaluate the UX of a VA customer service, other criteria are relevant [1] since the main focus of the user here is to get his or her request or task done. In this case, the classical scales *Efficiency*, *Dependability*, and *Perspicuity* may be good candidates besides the new voice scales. Of course, the specific use case and goal of the evaluation determine which scales should be combined.

6 SUMMARY

This article describes the construction of voice interaction scales for the UEQ+ framework. The modular concept of the UEQ+ is based on various scales, which allows the measurement of product-specific UX aspects [9]. The extension of the UEQ+ scale type 1 by voice interaction closes a gap in the UEQ+ and demonstrates a new method for the flexible evaluation of voice assistance systems. In this empirical study, three new scales were developed; and it was shown how relevant UX aspects for voice interaction could be represented. The data were evaluated using factor analysis and they presented the factors '*Response behaviour*', '*Response quality*' and '*Comprehensibility*' with four candidate items each. Two compact examples of possible questionnaires finally demonstrate how the UEQ+ scales can be combined with the new voice interaction scales. The validation of the new voice interaction scales is planned in a further study that will include the creation and application of questionnaires for voice system evaluation to obtain benchmarks.

REFERENCES

[1] Winter, D., Hinderks, A., Schrepp, M. & Thomaschewski, J., (2017). Welche UX Faktoren sind für mein Produkt wichtig? In: S. Hess & H. Fischer (Eds.), *Mensch und Computer 2017—Usability Professionals*. Regensburg: Gesellschaft für Informatik e.V. (pp. 191–200).

[2] Schrepp, M. (2018). User Experience mit Fragebögen messen [Measure user experience with questionnaires]. Amazon Kindle Direct Publishing, ISBN: 9781986843768.

[3] Schrepp, M., & Thomaschewski, J. (2019). Eine modulare Erweiterung des User Experience Questionnaire. Konferenz: Mensch und Computer 2019. DOI: 10.18420/muc2019-up-0108

[4] Schrepp, M., & Thomaschewski, J. (2019). Design and validation of a framework for the creation of user experience questionnaires. *International Journal of Interactive Multimedia and Artificial Intelligence*. DOI: 10.9781/ijimai.2019.06.006.

[5] https://www.tractica.com/newsroom/press-releases/voice-and-speech-recognition-software-market-to-reach-6-9-billion-by-2025/, accessed January 23, 2020.

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

[3] Comprehensibility (German original items): 1. '*kompliziert/einfach*', 2. '*ungenau/genau*', 3. '*unsinnig/sinnig*', 4. '*nicht eindeutig/eindeutig*', 5. '*unlogisch/logisch*', 6. '*unverständlich/verständlich*', 7. '*unerwartet/erwartet*', 8. '*unklar/klar*', 9. '*rätselhaft/erklärbar*', 10. '*schwierig/leicht*'.

[7] 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, NY: Routledge.
[8] Polkosky, M. D., & 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.
[9] Bos, J., Larsson, S., Lewin, I., Matheson, C., & Milward, D. (1999). Survey of existing interactive systems. Trindi (Task Oriented Instructional Dialogue) report, (D1), 3.
[10] Hinderks, A. (2016). Modifikation des User Experience Questionnaire (UEQ) zur Verbesserung der Reliabilität und Validität. Unveröffentlichte Masterarbeit, University of Applied Sciences Emden/Leer.
[11] Boos, B. & Brau, H., (2017). Erweiterung des UEQ um die Dimensionen Akustik und Haptik. In: Hess, S. & Fischer, H. (Hrsg.), Mensch und Computer 2017 – Usability Professionals, Regensburg: Gesellschaft für Informatik e.V., S. 321 – 327.
[12] Schrepp, M. & Thomaschewski, J. (2019). Handbook for the modular extension of the User Experience Questionnaire. - All you need to know to apply the UEQ+ to create your own UX questionnaire. DOI: 10.13140/RG.2.2.15485.20966.
[13] https://voicebot.ai/2019/03/12/smart-speaker-owners-agree-that-questions-music-and-weather-are-killer-apps-what-comes-next, accessed January 30, 2020.
[14] SAP CoPilot. https://blogs.sap.com/2016/10/06/the-human-touch-sap-introduces-a-digital-assistant-for-the-enterprise/, accessed February 5, 2020.
[15] Hassenzahl, Tractinsky, User experience—a research agenda, Behaviour & Information Technology, Vol. 25, No. 2, March-April 2006, 91 –97.
[16] Cohen, Michael H., Giangola, James P. & Balogh, Jennifer, (2004). Voice user interface design. Boston: Addison-Wesley Professional.
[17] Klein, A. M., Hinderks, A., Schrepp, M., & Thomaschewski, J., (2020). Measuring User Experience Quality of Voice Assistants. 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), Sevilla, Spain, 2020, pp. 1-4, DOI: 10.23919/CISTI49556.2020.9140966.
[18] Klein, A. M., Hinderks, A., Schrepp, M., & Thomaschewski, J., (2020). Protocol for. Measuring User Experience Quality of Voice Assistants. DOI: 10.13140/RG.2.2.12816.35848