Article in Press IJIMAI journal

A Benchmark for the UEQ+ Framework: Construction of a Simple Tool to Quickly Interpret UEQ+ KPIs

Anna-Lena Meiners[1]*, Martin Schrepp[2], Andreas Hinderks[3], Jörg Thomaschewski[1]

[1] University of Applied Sciences Emden/Leer (Germany)
[2] SAP SE (Germany)
[3] University of Sevilla (Spain)

Received 31 May 2022 | Accepted 12 April 2023 | Early Access 9 May 2023

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Questionnaires are a highly efficient method to compare the user experience (UX) of different interactive products or versions of a single product. Concretely, they allow us to evaluate the UX easily and to compare different products with a numeric UX score. However, often only one UX score from a single evaluated product is available. Without a comparison to other measurements, it is difficult to interpret an individual score, e.g. to decide whether a product's UX is good enough to compete in the market. Many questionnaires offer benchmarks to support researchers in these cases. A benchmark is the result of a larger set of product evaluations performed with the same questionnaire. The score obtained from a single product evaluation can be compared to the scores from this benchmark data set to quickly interpret the results. In this paper, the first benchmark for the UEQ+ (User Experience Questionnaire +) is presented, which was created using 3.290 UEQ+ responses for 26 successful software products. The UEQ+ is a modular framework that contains a high number of validated user experience scales that can be combined to form a UX questionnaire. Currently, no benchmark is available for this framework, making the benchmark constructed in this paper a valuable interpretation tool for UEQ+ questionnaires.

Keywords

Benchmark, Measurement, UEQ+ Framework, User Experience, UX Questionnaires.

DOI: 10.9781/ijimai.2023.05.003

I. Introduction

User experience (UX) is a key factor for the success of interactive products. It helps to attract new users and fosters loyalty [1]. Loyal customers are less likely to switch to a competitor – or to terminate their contract to switch to a competitor in the case of subscriptions –and more likely to buy a successor product from the same brand or manufacturer. Loyalty takes time to develop, so it is important that a product offers a constantly high level of UX. Thus, it is important to rigorously measure the UX of a product over a longer period of time. UX questionnaires, especially online questionnaires, are a popular measurement method for this purpose [2], since they allow to reach larger groups of users with low effort.

UX covers a large number of different task-related and non-task-related quality aspects concerning the interaction of a user and a product [3]–[5]. For a good UX impression, the product should be easy to learn or even intuitive to use, it should react to the user's commands as she or he expects, have a visually aesthetic user interface, and be fun to use, among other criteria. Therefore, to guarantee a high level of UX, several semantically distinct UX quality aspects must be considered.

How important such distinct UX aspects are for the overall UX impression of a product depends on personal preferences and on demographic attributes of the user – and, even more importantly, on the type of product [6]–[8]. For instance, efficiency is a key UX requirement for a business software that is used frequently in a typical workday. However, an intuitive interaction is not really expected in this case, since complex products typically require some learning phase, a factor which is considered in the design. Things look different for a web service that is used more sporadically, e.g. a tool to book concert tickets online: In this case, efficiency is nice, but it is not a key factor, whereas an intuitive interaction is absolutely required, since users will not accept any learning time for the simple tasks involved. This is a relatively simple example. For a more detailed analysis of the dependency between product type and the importance of UX aspects, see [6]–[8].

Of course, the number of questions that can be used in an online questionnaire is limited. Users cannot be expected to spend much time answering evaluation questions. At the same time, it is critical to measure the most important UX aspects, which will be different for each product. Measuring all the right aspects thus cannot be achieved using standardized UX questionnaires [3], [9], which contain a fixed set of items and scales.

The UEQ+ is a new modular framework that addresses this issue [9]. It contains a catalog of scales that can be combined to form a concrete UX questionnaire. Thus, a researcher can define the UX aspects that are important for a product and pick the UEQ+ scales that measure those aspects.

* Corresponding author.

E-mail address: anna-lena.meiners@ux-researchgroup.com

Please cite this article in press as:
A. L. Meiners, M. Schrepp, A. Hinderks, J. Thomaschewski. A Benchmark for the UEQ+ Framework: Construction of a Simple Tool to Quickly Interpret UEQ+ KPIs, International Journal of Interactive Multimedia and Artificial Intelligence, (2023), http://dx.doi.org/10.9781/ijimai.2023.05.003

Questionnaires built with the UEQ+ framework can already be used to compare different products concerning UX or to measure how the UX of a product develops over time [10]–[11]. However, no benchmark is currently available for the UEQ+. This is problematic since it is difficult to interpret the UX score of a single product without an appropriate comparison.

The goal of this paper is to define a first benchmark for the UEQ+ that will help UX researchers and practitioners to interpret standalone UEQ+ results.

II. The UEQ+ Framework

The UEQ+ [9] is a catalog of UX scales. It extends the UEQ [12]–[13] with additional scales. Each of these scales describes a special semantic aspect of the interaction between a user and a product. As already mentioned in the introduction, the UX aspects measured in a product evaluation depend on the product type and the specific research question. The modular structure of the UEQ+ addresses this requirement. Researchers can pick exactly those scales from the available UEQ+ scales that are most important for their study and can thus set up a questionnaire that optimally fits their research questions.

All UEQ+ scales share the same format, meaning that they can be combined in any order. The following UEQ+ scale shows the standardized format of four items related to a particular UX aspect, perspicuity, followed by a question on the importance of these items for the user:

In my opinion, handling and using the product are:

not understandable	o o o o o o o	understandable
difficult to learn	o o o o o o o	easy to learn
complicated	o o o o o o o	easy
confusing	o o o o o o o	clear

I consider the product property described by these terms as

completely irrelevant	o o o o o o o	very important

As shown above, the items of a UEQ+ scale consist of two terms that represent the opposite ends of a semantic dimension (for example, confusing/clear). Participants can describe their impression on a 7-point answer scale. An introductory sentence is used to set a common context for the four items. Following this, the question at the end is used to calculate an overall UX KPI by weighting the rating of a scale with its importance. The idea behind this KPI is identical to the KPI calculation for the original UEQ, see [14]. Since it is central for this paper, the calculation of the UEQ+ KPI is described in more detail in the following paragraphs.

Assume the scales S_1, ..., S_m have been chosen from the available UEQ+ scales. Thus, the final questionnaire contains m scales. Assume further that data has been collected from n participants in the study. Let s_{ij} be the score (the average of the 4 items in the scale) and w_{ij} the importance rating of participant i concerning scale j. Now, firstly, the relative importance r_{ij} of scale j for participant i is calculated by (1).

$$r_{ij} = \frac{w_{ij}}{\sum_{j=1}^{m} w_{ij}} \tag{1}$$

Thus, r_{ij} is the importance rating of scale j by participant i divided by the sum of all importance ratings of participant i. The KPI is then calculated by (2).

$$KPI = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} s_{ij} \tag{2}$$

Thus, the ratings per scale are first weighted with their relative importance per participant, and then the KPI is calculated as the average over the obtained values of all participants. Semantically, the UX KPI represents the overall UX impression that the participants got from the evaluated product. This KPI can easily be calculated in the data analysis Excel tool available on the UEQ+ homepage https://ueqplus.ueq-research.org.

The UEQ+ was developed as an extension of the UEQ and thus contains the 6 scales from the UEQ plus additional scales. The item format is slightly different to allow a free combination of scales. The main difference is that in the original UEQ the introductory sentence is missing, the items are not grouped by scales but appear in random order and that half of the items show the positive term on the left and the other half on the right. The reasons for the changes compared to the original UEQ item format are described in detail in [9].

As of today, the UEQ+ contains 20 UX scales (more scales may be added in the future) that represent different UX aspects of interactive products. Some examples of other scales are:

- *Efficiency:* Can users solve their tasks without unnecessary effort? Does the product react quickly?
- *Visual Aesthetics:* Does the product look beautiful and appealing?
- *Usefulness:* Does using the product bring advantages to the user?
- *Response quality:* Do the responses of a voice assistant satisfy the user's needs of information?
- *Acoustics:* Impact of sounds or operating noise of the product to the user experience.

Some scales are applicable to a large variety of products, for example efficiency, perspicuity, usefulness, or trust. Others make sense only for specific product types, for example acoustics (sound created by operation of a device – which was constructed to evaluate household appliances), aesthetics (only for products with a graphical user interface), or response quality (only for voice assistants).

The complete list of available scales can be found on the UEQ+ homepage https://ueqplus.ueq-research.org. Translations of the scales to more than 30 different languages are also available on this page. The homepage additionally offers supporting material, for example a data analysis Excel tool and a handbook that describes best practices concerning the usage of the UEQ+. The UEQ+ itself and all the provided material on the homepage are free to use.

Creating a concrete UX questionnaire for a product evaluation based on the UEQ+ is simple. The researcher must decide which scales are relevant for the product that should be evaluated. Then these scales are placed in a sequence to form the concrete questionnaire. The recommendation is to use at most 6 scales in a questionnaire constructed from the UEQ+ framework to keep the time required to fill it out within a reasonable range. If more scales are needed to get a clear picture on the UX of a product, it is recommended to split them into two different questionnaires, i.e., to collect data from two samples with a reduced number of scales.

Scales for the UEQ+ framework can be constructed independently and tested for their reliability and validity. This makes it possible to enhance the framework step by step with additional scales [9].

III. Related Research: Different UX Benchmarks

A UX questionnaire allows us to compare products with respect to the scales found in that questionnaire. If product A obtains a significantly higher score in a scale than product B, then A is better than B concerning the UX quality measured by this scale. However, if we have only a single measurement, it is usually difficult to interpret its value directly. For example, is a mean value *Efficiency* = 1,1 on the UEQ+ scale a good or bad result [15]? Does it indicate that the efficiency of the product is sufficiently high? This question can only be answered by comparing the measured score with scores obtained for other products.

This is the idea behind benchmarks. A benchmark is the result of a collection of measurements obtained from different products with a UX questionnaire. Thus, a benchmark allows us to determine how well the evaluated product has performed compared to the products in the benchmark data set. Several standardized questionnaires, for example the User Experience Questionnaire (UEQ) [16], the System Usability Scale (SUS) [17], the Software Usability Measurement Inventory (SUMI) [18], the Usability Metric for User Experience (UMUX) [19], the shorter version UMUX-Lite [20], the Standardized User Experience Percentile Rank Questionnaire (SUPR-Q) [21] or the Visual Aesthetics of Websites Inventory (VisAWI) [22], include such benchmarks.

How exactly a benchmark is defined depends on the questionnaire. This is illustrated by some examples in what follows.

First, we look at the SUS [23] benchmarks. The SUS questionnaire contains 10 items that can be answered on a 5-point Likert scale, with scale values ranging from 0 to 4. The questionnaire does not provide separate scales to measure sub-aspects of the UX. Rather, it offers a single value that represents the overall usability of the product. To calculate this overall value, all 10 item scores of a participant are summed up, resulting in a value between 0 and 40. This value is then multiplied by 2,5, so that the SUS score is stretched to a range from 0 to 100. According to [24], this rescaling was done primarily because a scale of 0 to 100 is easier to communicate to product managers than a scale of 0 to 40. Therefore, the rescaling is done to improve the communication of the results.

Another benchmark is found in [25] and [26], in which a 7-level rating is derived from a very large collection of SUS data. Here, the SUS values per person were compared to an overall assessment of the product. This overall rating was made possible by seven terms, one of which had to be chosen. This allows a relationship between the overall assessment of a person and the SUS score of the person for the evaluated product. The seven terms are listed below. The value behind the term is the average SUS rating associated with the term (average of the SUS ratings from participants that used this term as an overall rating), whereas the numerical value in parentheses is the standard deviation of this rating.

- *Best imaginable*: 90,9 (13,4)
- *Excellent*: 85,5 (10,4)
- *Good*: 71,4 (11,6)
- *OK*: 50,9 (13,8)
- *Poor*: 35,7 (12,6)
- *Awful*: 20,3 (11,3)
- *Worst Imaginable*: 12,5 (13,1)

Let's assume we measure a product with the SUS and get a score of 25. This would correspond to an overall rating between *Awful* and *Poor* (leaning towards *Awful*). If we get an overall score of 87, this corresponds to an overall rating of *Excellent*. Thus, this simple benchmark helps to interpret a single SUS result by relating the SUS score to a statement about overall UX quality.

Another SUS benchmark with more recent data is presented in [23]. This paper provides 11 categories for the results (see Table I) based on a benchmark set of 241 industrial usability studies. The percentile x-y can be interpreted as follows: x percent of the products from the benchmark showed a result lower than your score, 100-y of the products showed a better result.

Thus, if you obtain a score of 25 then you are in category F and your product belongs to the 14% worst products in the benchmark. If you get a score of 85, then your product would be in category A+ and you are amongst the best 4% of products in the benchmark. Again, this benchmark is very helpful to decide if a single measurement obtained for a product indicates a good, average, or bad UX.

TABLE I. Sus Benchmark as Formulated in [23]

Category	Score Interval		Percentile
A+	84,10	100,00	96–100
A	80,80	84,00	90–95
A-	78,90	80,70	85–89
B+	77,20	78,80	80–84
B	74,10	77,10	70–79
B-	72,60	74,00	65–69
C+	71,10	72,50	60–64
C	65,00	71,00	41–59
C-	62,70	64,90	35–40
D	51,70	62,60	15–34
F	0,00	51,60	0–14

These examples show that the goal of a benchmark is to provide some interpretation concerning how good or bad a measured result is overall. How this is formulated in detail is not standardized, but rather decided by the researchers that set up the benchmark.

The benchmark of the UEQ, as described in [16] and [27], works in a similar way to the benchmark described in [23]. It is based on the data of 21.175 participants from 468 different studies, where one study corresponds to one measurement of one product made with the UEQ. For each scale, the measured value is divided into 5 categories:

- ***Excellent:*** The measured scale value is among the top 10% of the best results.
- ***Good:*** The measured scale value is better than 75% of the measured results and worse than the 10% best results.
- ***Above Average:*** The measured scale value is better than 50% of the measured results and worse than the 25% best results.
- ***Below Average:*** The measured scale value is better than 25% of the measured results and worse than the 50% best results.
- ***Bad:*** The measured scale value is among the 25% worst results.

The UEQ benchmark is available in the data analysis Excel tool for the UEQ that can be downloaded from www.ueq-online.org. The graphical representation in Fig. 1 is automatically calculated from the measured scale values of the 6 UEQ scales for a product. For the example product shown in Fig. 1 (black line) we can see that it is highly rated in the pragmatic quality aspects *Perspicuity*, *Efficiency* and *Dependability*, but shows a low rating for aspects that are related to fun of use, i.e., *Stimulation* and *Novelty*. Thus, in this example the result also provides some clear indication on how to improve the product in the future. One could conclude that investments in the usability of the product are not necessarily important, but improvements concerning fun of use will most likely have a big impact on the overall UX.

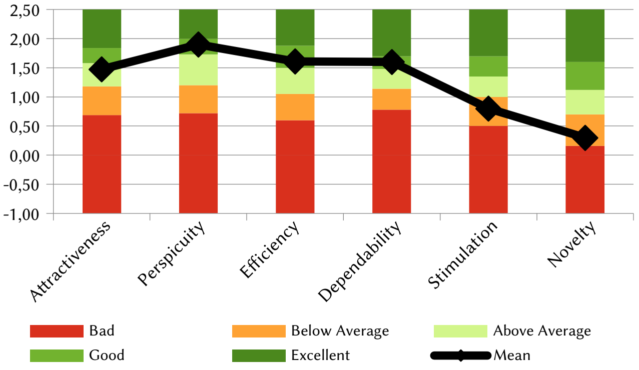


Fig. 1. Comparison of the UEQ measurement of a hypothetical product with the products in the benchmark data set. The UEQ scale ranges from -3 (worst) to 3 (best).

The VisAWI, as presented in [22] and [28], pursues a somewhat different type of benchmarking. A large amount of existing data from more than 160 different web pages was evaluated. These were additionally divided into categories, for example weblogs and social sharing, e-commerce, and information. For each group, the benchmark shows the mean value and standard deviation for the 4 scales of the VisAWI. A convenient feature of this benchmark is that one can always compare a result with a group of very similar web pages. However, the evaluation is somewhat limited, since one can only conclude whether the VisAWI result for a web page is above or below average.

To sum up, there are various ways of creating a benchmark, and existing questionnaires have taken slightly different paths in this respect. However, the aim of a benchmark is always to make the results of a single UX questionnaire easier to interpret, and to do so it is providing a comparison with a large amount of existing data from other product evaluations.

IV. Designing the Study: Preliminary Thoughts on the UEQ+ Benchmark

As was shown in the last section, benchmarks support the interpretation of UX questionnaire results. For the UEQ+, no such benchmark is available so far, and the goal of this paper is to provide at least a first benchmark to fill this gap.

As we have seen, the UEQ+ is a flexible framework that offers a catalog of currently 20 scales to the researcher [9]. The flexibility of such a modular approach has some drawbacks. Providing a benchmark for each scale is extremely difficult since it requires the collection of many product evaluations with the same scale. Since some of the scales are only useful for specific product types, this is very difficult in practice. Thus, providing a benchmark for the single scales, as needed for the UEQ, will require a long time. Since scales will be added from time to time, such a benchmark will also never cover all UEQ+ scales with the same quality.

Therefore, this study follows a different approach: It defines a first simple benchmark based on a limited set of product evaluations, which will provide some quick guidance to UX researchers.

This benchmark is based on the UX KPI, i.e., not on single scales but on the overall user's impression of a product. Of course, this KPI depends on the scales selected for evaluation and this selection varies from product to product. So, it is highly questionable if products from different categories can be reasonably compared by comparing their KPIs. However, from a practical perspective this is not relevant in most cases. Often a comparison to some competitors or products serving the same use cases is sufficient to get a first idea as to whether the own product is "good enough" concerning UX. And since the importance of UX aspects depends on the product type [6]–[8], a similar set of scales will most likely be chosen to measure products of the same type.

Thus, this benchmark was constructed based on one or two representatives for each product type. The KPI of each product was measured via a questionnaire with scales from the UEQ+. For each product, scales were chosen based on how important certain UX aspects are to the users of that product, as reported in previous research [6]–[8].

V. Conducting the Study: Creation of the UEQ+ Benchmark

The data used to create the benchmark was collected in two sets of studies: (1) the data of four studies was taken from our research repository from already existing UEQ+ studies perfectly matching our requirements (inventory data), and (2) the data of 22 additional studies that were collected in two waves, in the fall of 2020 and the fall of 2021 (original data). Details of the data collection are described in the following paragraphs.

Of the inventory data sets, three (regarding Amazon Prime Video, otto.de & zalando.de) were collected as part of validation studies for the UEQ+ [9], and one (regarding Facebook) was part of an interpretation analysis study [29]. The Facebook data was collected via an English social panel. All admissible participants stated that they used Facebook at least once a month.

The questionnaires used for the validation studies were distributed via e-mail and linked on websites. Table II shows participant details for these inventory data sets, specifically their language, mean age and gender attributes.

TABLE II. Participant Details of Inventory Data Sets (Responses, Number of People Stating They Are Male, Female, and Diverse or Not Specified, Mean Age and Language Version of the UEQ+ Used)

UEQ+ Study	N	m	f	d/ not specified	Mean age	Language version
Amazon Prime Video	57	36	21	0	32	German
Facebook	248	112	132	4	30	English
otto.de	42	16	25	1	34	German
zalando.de	46	20	24	2	31	German

Of the original data sets, three were collected via social panels (Alexa, bbc.com & Ebay). The remaining 19 data sets were collected via multi-channel distribution of questionnaires supported by university students at University of Applied Sciences Emden/Leer. Students shared the questionnaires via e-mail, forums, social media, and messengers.

All of these questionnaires were made available in multi-language versions using standardized translations for German and English, so that participants could answer in the language they preferred, and the results could be merged into one database per study. The collected data was then cleaned using the following exclusion criteria:

(1) perfect duplication of entries (oldest entry kept)

(2) time needed to complete questionnaire was below 50 seconds

(3) stated age was over 90 or below 16

(4) less than 80% of UEQ+ items were answered

(5) control questions were answered wrong

Information on the remaining participants in the original studies is found in Table III.

Each questionnaire started with a short introduction followed by demographic questions (gender and age). Then the sequence of UEQ+ items (see Fig. 3 in the Appendix) was displayed. After this block, in most cases, comment fields and questions were added, in order to detect persons that do not answer the questions seriously.

The scales contained in the UEQ+ questionnaires were selected by product. Previous studies [6]–[8] investigated the importance of common UX aspects (corresponding to UEQ+ scales) for typical product types. We used these results for the selection of the scales. Thus, for each investigated product, the corresponding product type was determined and then the most important scales according to the results of these studies were selected. As suggested in the UEQ+ handbook, a maximum of 6 scales was used in a single study. Table V in the Appendix summarizes which scales were used in the different questionnaires.

TABLE III. Participant Details of Original Data Sets (Responses, Number of People Stating they Are Male, Female, and Diverse or Not Specified, Mean Age and Language Version of the UEQ+ Used)

UEQ+ Study	N	m	f	d/ not specified	Mean age	Language version
AirBnB	91	39	49	3	30	89 German, 2 English
Alexa	100	55	43	2	27	English
Amazon	208	110	92	6	38	German
bbc.com	98	31	67	0	37	English
Booking.com	49	20	26	3	36	46 German, 3 English
Ebay	100	49	48	3	30	English
Google Maps	111	63	41	7	31	German
Instagram	97	36	56	5	27	German
Moodle	93	39	49	3	31	German
MS Excel	120	53	61	6	34	89 German, 15 English
MS Teams	130	84	30	16	38	German
MS Word	70	42	22	6	38	German
Netflix	46	27	17	2	30	German
Skype	57	26	24	7	37	German
Spotify	245	116	120	9	28	243 German, 2 English
TikTok	51	21	29	1	26	49 German, 2 English
Trello	28	14	12	2	35	27 German, 1 English
Udemy	41	23	17	1	32	40 German, 1 English
WhatsApp	176	72	92	12	32	174 German, 2 English
Wikipedia	444	104	251	89	28	439 German, 5 English
YouTube	517	409	91	17	25	German
Zoom	25	12	11	2	37	German

TABLE IV: Investigated Product, Mean, Standard Deviation and Confidence Interval of the KPIs and Number of Responses per Study

Product	KPI	Std	95% Conf. Int.		Responses
Google Maps	1,82	0,67	1,78	1,86	111
Wikipedia	1,79	0,63	1,76	1,81	444
zalando.de	1,70	0,69	1,63	1,77	46
Spotify	1,66	0,76	1,63	1,69	245
Udemy	1,66	0,71	1,59	1,74	41
YouTube	1,60	0,67	1,58	1,61	517
BBC.com	1,54	0,89	1,48	1,60	98
Zoom	1,47	0,85	1,35	1,58	25
MS Excel	1,46	0,89	1,41	1,52	120
Alexa	1,46	0,74	1,41	1,51	100
Netflix	1,43	0,86	1,34	1,51	46
Booking.com	1,41	0,83	1,33	1,49	49
WhatsApp	1,39	0,87	1,34	1,43	176
Amazon Prime Video	1,35	0,87	1,27	1,43	57
Trello	1,29	0,67	1,20	1,37	28
otto.de	1,27	0,90	1,18	1,36	42
Ebay	1,25	0,95	1,19	1,32	100
Amazon	1,25	0,82	1,21	1,29	208
MS Teams	1,18	0,92	1,13	1,24	130
MS Word	1,03	0,96	0,95	1,10	70
Instagram	0,97	0,84	0,91	1,03	97
AirBnB	0,96	0,99	0,89	1,03	91
Moodle	0,71	0,88	0,65	0,77	93
Skype	0,60	1,05	0,50	0,69	57
TikTok	0,54	0,95	0,45	0,63	51
Facebook	0,39	1,06	0,34	0,43	248

VI. Results: The UEQ+ Benchmark

Following exclusion, a total of 3.290 responses to our surveys were used for this study. 1.629 participants identified as male, 1.447 as female, and 214 didn't specify their gender. The vast majority of questionnaires was answered in the German language version (2.700 responses); another 590 questionnaires were answered in the English language version. The mean age of participants was 33 years.

From the collected data a single UX KPI was calculated per product. As explained, these KPIs make up the initial simple benchmark we were striving to introduce in this study.

Table IV shows the measured values for the UX KPI per product. The scale for the UEQ+ KPI ranges from -3 (worst possible) to +3 (best possible). The measured KPIs ranged from +0,39 (for the social network Facebook) to +1,82 (for the navigation software Google Maps).

Fig. 2 displays the confidence intervals of the KPIs for the products in our benchmark data set, providing a quick overview of the results.

VII. Conclusions: Interpreting the UEQ+ Benchmark

The benchmark data (see Table IV and Fig. 2) already allows for a first rough interpretation of UEQ+ KPIs. The products investigated are all well-known, established products, and in many cases, they are the market leaders in their segments. Thus, they have a certain level of UX maturity. Overall, a rating between +1 and +2 for the KPI seems to represent a good level of UX.

Some examples can give an idea of how this benchmark can be used and interpreted more in depth. Assume you are a UX researcher and in your recent project you evaluate a web shop. The measured KPI is around 1,5. Is this an indicator of a good or of a poor UX impression? If you look into the list of evaluated products, you find two popular German web shops (otto.de and zalando.de). Their 95% confidence intervals are 1,18 to 1,36 for otto.de and 1,63 to 1,77 for zalando.de. Thus, the UX of your shop is between these two established shops, which clearly indicates that your UX is most likely in an acceptable range.

Assume now that you evaluate a new messenger. The product is new, offers some special services, but must still compete against established products to gain market share. Thus, it makes sense to check if the UX impression of the new messenger is at least close to the UX impression of *WhatsApp* (whose KPI has a confidence interval of 1,38–1,46. Thus, if you obtain a 0,5 score you immediately see that you cannot compete in terms of UX. On the other hand, if you scored 1,4, you would be in a comparable range to the UX of *WhatsApp*.

This benchmark is helpful even in cases where there is no direct match between the products in this benchmark and the evaluated product. Since these are all common products, it is easy to get a personal impression of the UX of these products. So, even a simple first-glance-statement such as "our product generates a UX impression similar to Ebay" can help interpret the results semantically.

In conclusion, with the help of this simple benchmark, UEQ+ results can be compared quickly and easily, which aids practitioners and UX researchers alike by giving an orientation on how to interpret the results of the UEQ+.

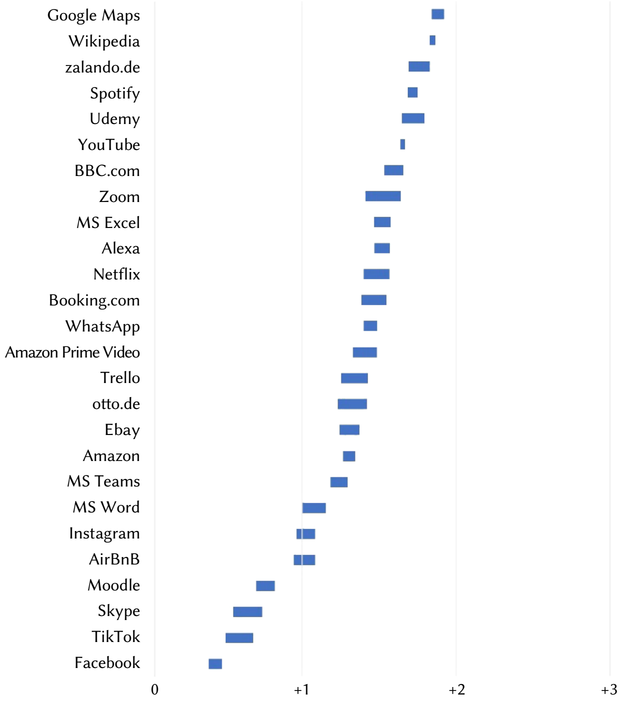


Fig. 2: Confidence intervals for the UEQ+ KPIs of the evaluated products, sorted by KPI.

VIII. Summary and Outlook

A first benchmark for the UEQ+ framework was created. Due to the modular structure of the UEQ+, it is difficult to provide a classical benchmark on the level of single scales. Indeed, since some of the scales are only useful for specific products, it would require a long time to collect enough data for such a benchmark, as with the benchmark of the original UEQ.

Therefore, this first benchmark is based on the UEQ+ KPI and on the evaluation of several well-known products by over 3.200 study participations. A comparison to the KPI values of these products enables UX researchers to develop a first, quick understanding of how good the UX of the evaluated product is compared to user expectations.

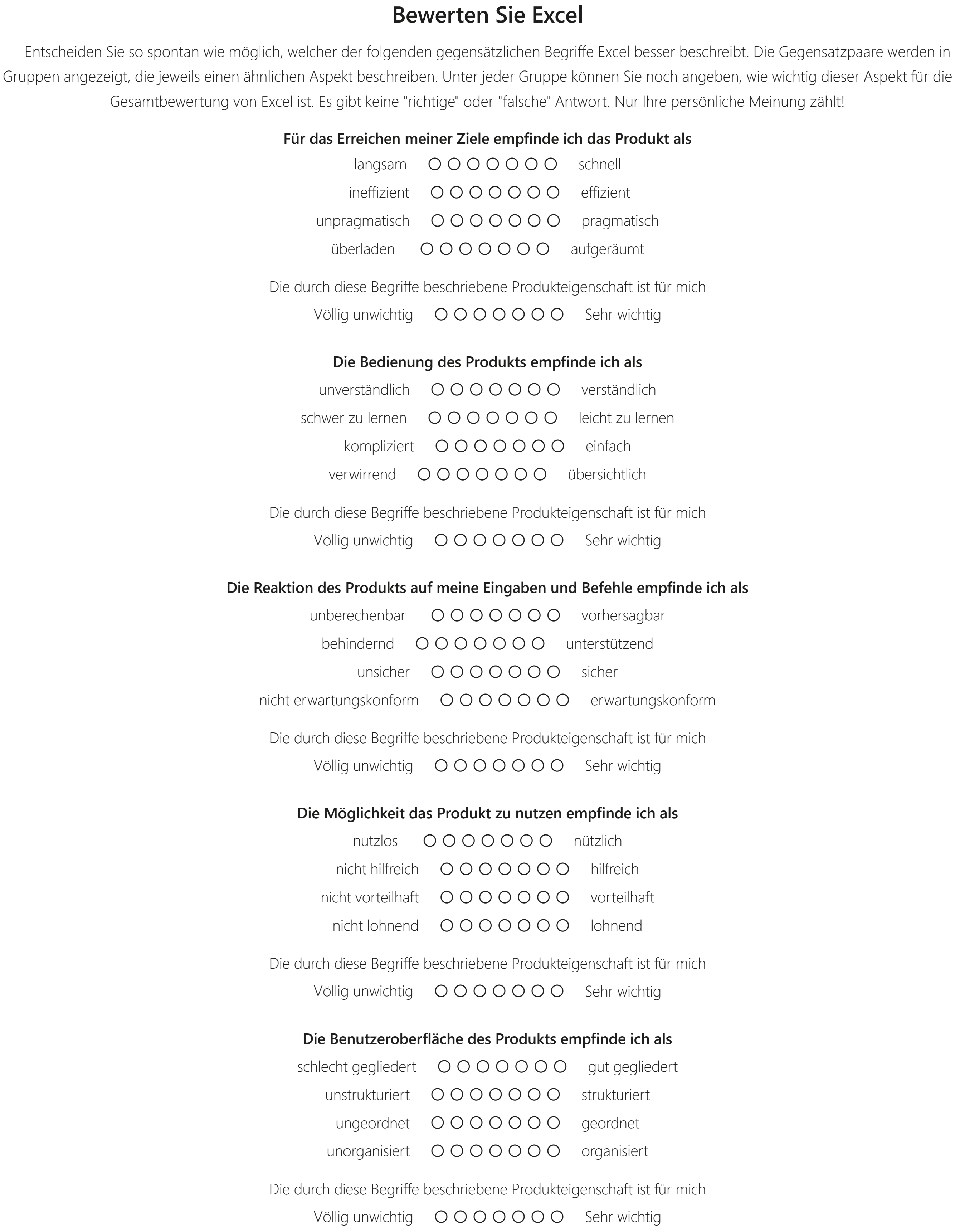
Future lines of work should include studies that give insight into the comparability of products through their UEQ+ KPIs. As mentioned before, it is questionable whether products from different categories are comparable through their UEQ+ KPI. This should be investigated further as it would result in knowledge useful for an easier interpretation of UX data. Further studies should also include more transparent and/or standardized ways of choosing UEQ+ scales when creating the questionnaires as this seems to have happened almost arbitrarily at some points in the past. This would benefit the replicability of the studies and enable expedient comparability of results, which in turn allows for more meaningful and nuanced interpretation of a KPI benchmark. Some research supporting UEQ+ users in this regard has already been conducted [6]–[8], but should be expanded on. It is also possible that product categories such as "banking software" and "shops" don't partition the available data in the most insightful way regarding the comparability of the KPIs, and that instead more abstract categories such as "software for work" or "tools people use rarely" would generate more useful insights. Studies in this regard could begin to create a model of product categories that categorize products from a UX perspective.

Nevertheless, this first benchmark provides a valuable insight for UX practitioners to judge the UX quality of the products they design and evaluate.

Appendix

TABLEV. Investigated Products and Scales Used in the Study

Product	UEQ+ Scales
Facebook	Quality of Content, Trustworthiness of Content, Intuitive Use, Trust, Stimulation
Alexa	Response behavior, Response quality, Comprehensibility, Trust, Usefulness
BBC.com	Perspicuity, Value, Intuitive Use, Quality of Content, Clarity
zalando.de	Attractiveness, Dependability, Intuitive Use, Visual Aesthetic, Quality of Content, Trustworthiness of Content, Trust, Value
Ebay	Trust, Quality of Content, Dependability, Clarity, Intuitive Use
Google Maps	Efficiency, Perspicuity, Usefulness, Intuitive Use, Trustworthiness of Content, Quality of Content
Instagram	Attractiveness, Stimulation, Novelty, Trust, Visual Aesthetic, Intuitive Use, Clarity
Netflix	Attractiveness, Perspicuity, Stimulation, Visual Aesthetics, Intuitive Use, Quality of Content
Teams	Efficiency, Perspicuity, Dependability, Trust, Usefulness, Clarity
Trello	Efficiency, Perspicuity, Trust, Adaptability, Usefulness, Clarity
WhatsApp	Efficiency, Perspicuity, Dependability, Trust, Intuitive Use, Clarity
Word	Efficiency, Perspicuity, Dependability, Usefulness, Intuitive Use, Clarity
YouTube	Attractiveness, Perspicuity, Dependability, Stimulation, Intuitive Use, Clarity
Amazon Prime Video	Attractiveness, Perspicuity, Intuitive Use, Visual Aesthetic, Quality of Content, Trustworthiness of Content, Trust
Moodle	Attractiveness, Perspicuity, Dependability, Adaptability, Usefulness, Clarity
otto.de	Attractiveness, Dependability, Intuitive Use, Visual Aesthetic, Quality of Content, Trustworthiness of Content, Trust, Value
AirBnB	Trust, Quality of Content, Dependability, Efficiency, Clarity
Amazon	Trust, Quality of Content, Dependability, Clarity, Intuitive Use
Booking.com	Trust, Quality of Content, Dependability, Efficiency, Clarity
Excel	Usefulness, Dependability, Efficiency, Perspicuity, Clarity
Skype	Trust, Dependability, Efficiency, Usefulness, Intuitive Use
Spotify	Perspicuity, Dependability, Stimulation, Adaptability, Intuitive Use
Tiktok	Trust, Dependability, Intuitive Use, Quality of Content, Stimulation
Udemy	Quality of Content, Usefulness, Clarity, Perspicuity, Efficiency
Wikipedia	Quality of Content, Clarity, Perspicuity, Visual Aesthetic, Intuitive Use
Zoom	Trust, Dependability, Efficiency, Usefulness, Intuitive Use



Bewerten Sie Excel

Entscheiden Sie so spontan wie möglich, welcher der folgenden gegensätzlichen Begriffe Excel besser beschreibt. Die Gegensatzpaare werden in Gruppen angezeigt, die jeweils einen ähnlichen Aspekt beschreiben. Unter jeder Gruppe können Sie noch angeben, wie wichtig dieser Aspekt für die Gesamtbewertung von Excel ist. Es gibt keine "richtige" oder "falsche" Antwort. Nur Ihre persönliche Meinung zählt!

Für das Erreichen meiner Ziele empfinde ich das Produkt als

langsam	○ ○ ○ ○ ○ ○ ○	schnell
ineffizient	○ ○ ○ ○ ○ ○ ○	effizient
unpragmatisch	○ ○ ○ ○ ○ ○ ○	pragmatisch
überladen	○ ○ ○ ○ ○ ○ ○	aufgeräumt

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich

Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Die Bedienung des Produkts empfinde ich als

unverständlich	○ ○ ○ ○ ○ ○ ○	verständlich
schwer zu lernen	○ ○ ○ ○ ○ ○ ○	leicht zu lernen
kompliziert	○ ○ ○ ○ ○ ○ ○	einfach
verwirrend	○ ○ ○ ○ ○ ○ ○	übersichtlich

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich

Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Die Reaktion des Produkts auf meine Eingaben und Befehle empfinde ich als

unberechenbar	○ ○ ○ ○ ○ ○ ○	vorhersagbar
behindernd	○ ○ ○ ○ ○ ○ ○	unterstützend
unsicher	○ ○ ○ ○ ○ ○ ○	sicher
nicht erwartungskonform	○ ○ ○ ○ ○ ○ ○	erwartungskonform

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich

Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Die Möglichkeit das Produkt zu nutzen empfinde ich als

nutzlos	○ ○ ○ ○ ○ ○ ○	nützlich
nicht hilfreich	○ ○ ○ ○ ○ ○ ○	hilfreich
nicht vorteilhaft	○ ○ ○ ○ ○ ○ ○	vorteilhaft
nicht lohnend	○ ○ ○ ○ ○ ○ ○	lohnend

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich

Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Die Benutzeroberfläche des Produkts empfinde ich als

schlecht gegliedert	○ ○ ○ ○ ○ ○ ○	gut gegliedert
unstrukturiert	○ ○ ○ ○ ○ ○ ○	strukturiert
ungeordnet	○ ○ ○ ○ ○ ○ ○	geordnet
unorganisiert	○ ○ ○ ○ ○ ○ ○	organisiert

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich

Völlig unwichtig ○ ○ ○ ○ ○ ○ ○ Sehr wichtig

Fig. 3: Example of the UEQ+ part of the survey (German).

References

[1] M. Thüring, and S. Mahlke, ”Usability, aesthetics and emotions in human–technology interaction," *International journal of psychology*, vol. 42, no. 4, 2007, pp. 253-264.

[2] J. R. Lewis, and J. Sauro, "Usability and user experience: Design and evaluation," *Handbook of Human Factors and Ergonomics*, 2021, pp. 972-1015.

[3] M. Schrepp, *User Experience Questionnaires: How to use questionnaires to measure the user experience of your products?*, KDP, 2021, ISBN-13: 979-8736459766.

[4] J. Preece, Y. Rogers, and H. Sharpe, "Interaction design: Beyond human-computer interaction," Wiley, New York, 2002.

[5] M. Hassenzahl, "The effect of perceived hedonic quality on product appealingness," *International Journal of Human-Computer Interaction*, vol. 13, no. 4, 2001, pp. 481-499.

[6] A.-L. Meiners, J. Kollmorgen, M. Schrepp, and J. Thomaschewski, "Which UX aspects are important for a software product? Importance ratings of UX aspects for software products for measurement with the UEQ+," in *Proceedings of Mensch und Computer 2021 (MuC '21)*, Association for Computing Machinery, New York, NY, USA, 2021, pp. 136–139.

[7] H. B. Santoso and M. Schrepp, "The impact of culture and product on the subjective importance of user experience aspects," *Heliyon*, vol. 5, no. 9, 2019, doi: 10.1016/j.heliyon.2019.e02434.

[8] M. Schrepp, J. Kollmorgen, A.-L. Meiners, A. Hinderks, D. Winter, H. B. Santoso, J. Thomaschewski. On the Importance of UX Quality Aspects for Different Product Categories, International Journal of Interactive Multimedia and Artificial Intelligence, (2023), http://dx.doi.org/10.9781/ijimai.2023.03.001.

[9] M. Schrepp, and J. Thomaschewski, "Design and Validation of a Framework for the Creation of User Experience Questionnaires," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, 2019, doi: 10.9781/ijimai.2019.06.006.

[10] H. Sandkühler, M. Schrepp, and J. Thomaschewski, "UX Messung mithilfe des UEQ+ Frameworks (Measuring UX with the UEQ+ framework)",, in *Mensch und Computer 2020 - Workshopband*, Gesellschaft für Informatik, Bonn, 2020.

[11] M. Schrepp, H. Sandkühler, and J. Thomaschewski, "How to create short forms of UEQ+ based questionnaires?," in *Mensch und Computer 2021-Workshopband*, 2021.

[12] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *Symposium of the Austrian HCI and usability engineering group*, Springer, Berlin, Heidelberg, 2008, pp. 63-76.

[13] B. Laugwitz, M. Schrepp, and T. Held, "Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten (Construction of a questionnaire to measure UX of software products)," in *Mensch und Computer 2006,* Oldenbourg Wissenschaftsverlag, 2006, pp. 125-134.

[14] A. Hinderks, M. Schrepp, F. J. D. Mayo, M. J. Escalona, and J. Thomaschewski, "Developing a UX KPI based on the user experience questionnaire,“ *Computer Standards & Interfaces*, vol. 65, 2019 pp. 38-44.

[15] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Applying the user experience questionnaire (UEQ) in different evaluation scenarios," in *International Conference of Design, User Experience, and Usability*, Springer, Cham, 2014, pp. 383-392.

[16] M. Schrepp, A. Hinderks, and J. Thomaschewski, ”Construction of a benchmark for the User Experience Questionnaire (UEQ*)*," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, 2017, pp. 40-44.

[17] J. Brooke, "SUS-A quick and dirty usability scale," *Usability Evaluation in Industry*, vol. 189, no. 194, 1996, pp. 4-7.

[18] J. Kirakowski, and M. Corbett, "SUMI: The software usability measurement inventory," *British Journal of Educational Technology*, vol. 24, no. 3, 1993, pp.210-212.

[19] K. Finstad, "The usability metric for user experience," *Interacting with Computers*, vol. 22, no. 5, 2010, pp. 323-327.

[20] J. R. Lewis, B. Utesch, and D. E. Maher, D. E., "UMUX-LITE: when there's no time for the SUS," *in Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 2099-2102.

[21] J. Sauro, "SUPR-Q: A comprehensive measure of the quality of the website user experience," *Journal of usability studies*, vol. 10, no. 2, 2015.

[22] M. Moshagen, and M. Thielsch, "Facets of visual aesthetics," *International Journal of Human-Computer Studies*, vol. 68, no. 10, 2010, pp. 689-709.

[23] J. Sauro, and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*, 2nd ed., Cambridge, MA: Morgan-Kaufmann, 2016.

[24] J. Brooke, "SUS A Retrospective," *Journal of Usability Studies*, vol. 8, no. 2, 2013, pp. 29-40.

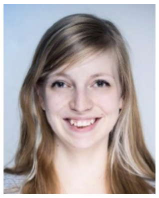
[25] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the System Usability Scale," *International Journal of Human-Computer Interaction*, vol. 24, no. 6, 2008, pp. 574-594.

[26] A. Bangor, P. T. Kortum, and J. T. Miller, J. T., "Determining what individual SUS scores mean: Adding an adjective rating scale," *Journal of Usability Studies,* vol. 4, no. 3, 2009 pp. 114-123.

[27] M. Schrepp, S. Olschner, and U. Schubert, „User Experience Questionnaire (UEQ) Benchmark," in *Tagungsband UP13*, 2013.

[28] M. Moshagen and M. Thielsch, "A short version of the visual aesthetics of websites inventory," *Behaviour & Information Technology*, vol. 32, no. 12, 2013, pp. 1305-1311.

[29] A. Hinderks, F. J. Domínguez-Mayo, A.-L. Meiners, and J. Thomaschewski, "Applying Importance-Performance Analysis (IPA) to interpret the results of the User Experience Questionnaire (UEQ)," *Journal of Web Engineering*, vol. 19, no. 2, 2020, pp. 243-266.



Anna-Lena Meiners

Anna-Lena Meiners received a Bachelor's degree in Theatre Studies, Philosophy and Dutch Language and Literature from Freie Universität Berlin and a Bachelor's as well as a Master's degree in Computer Science and Digital Media from University of Applied Sciences Emden/Leer with a focus on Human-Computer Interaction. Currently, she is a PhD researcher at Karlsruhe Institute of Technology (KIT). Her research focusses on technology design for positive and enriching user experiences.

Martin Schrepp

Martin Schrepp has been working as a user interface designer and researcher for SAP SE since 1994. He finished his diploma in mathematics in 1990 at the University of Heidelberg (Germany). In 1993 he received a PhD in Psychology (also from the University of Heidelberg). His research interests are the application of psychological theories to improve the design of software interfaces, the application of *Design for All* principles to increase accessibility of business software, measurement of usability and user experience, and the development of general data analysis methods. He has published several papers concerning these research fields.

Andreas Hinderks

Andreas Hinderks holds a PhD in Computer Science by University of Sevilla. He has worked in various management roles as a Business Analyst and a programmer from 2001 to 2016. His focus lay on developing user-friendly business software. Currently, he is a freelancing Product Owner, Business Analyst and Senior UX Architect. He is involved in research activities dealing with UX questionnaires, measuring user experience and User Experience Management since 2011.

Jörg Thomaschewski

Jörg Thomaschewski received a PhD in physics from the University of Bremen (Germany) in 1996. He became Full Professor at the University of Applied Sciences Emden/Leer (Germany) in September 2000. His research interests are human-computer interaction, agile software development, and e-learning. Dr. Thomaschewski is the founder of the research group "Agile Software Development and User Experience".