Construction of a Benchmark for the User Experience Questionnaire (UEQ)

Martin Schrepp[1], Andreas Hinderks[2], Jörg Thomaschewski[2]

[1]SAP AG, Germany
[2]University of Applied Sciences Emden/Leer, Germany

***Abstract* — Questionnaires are a cheap and highly efficient tool for achieving a quantitative measure of a product's user experience (UX). However, it is not always easy to decide, if a questionnaire result can really show whether a product satisfies this quality aspect. So a benchmark is useful. It allows comparing the results of one product to a large set of other products. In this paper we describe a benchmark for the User Experience Questionnaire (UEQ), a widely used evaluation tool for interactive products. We also describe how the benchmark can be applied to the quality assurance process for concrete projects.**

***Keywords* — User Experience, UEQ, Questionnaire, Benchmark.**

I. Introduction

In today's competitive market, outstanding user experience (UX) is a must for any product's commercial success. UX is a very subjective impression, so in principle it is difficult to measure. However, given the importance of this characteristic, it is important to measure it accurately. This measure can be used, for example, to check if a new product version offers improved UX, or if a product is better or worse than the competition [1].

There are several methods to quantify UX. One of the most widespread are usability tests [2], where the number of observed problems and the time participants need to solve tasks are quantitative indicators for the UX quality of a product. However, this method requires enormous effort: finding suitable participants, preparing tasks and a test system, and setting up a test site. Therefore typical sample sizes are very small (about 10-15 users).

In addition, it is a purely problem-centered method, i.e. it focuses on detecting usability problems. Usability tests are not able to provide information about users' impression of hedonic quality aspects, such as novelty or stimulation, although such aspects are crucial to a person's overall impression concerning UX [3].

Other well-known methods rely on expert judgment, for example, cognitive walkthrough [4] or usability reviews [5] against established principles, such as Nielsen's usability heuristics [6]. Like usability tests, these methods focus on detecting usability issues or deviations from accepted guidelines and principles. They do not provide a broader view of a product's UX.

A method that is able to measure all types of quality aspects and at the same time collect feedback from larger samples are standardized UX questionnaires. "Standardized" means that these questionnaires are not a more or less random or subjective collection of questions, but result from a careful construction process. This process guarantees accurate measuring of the intended UX qualities.

Such standardized questionnaires try to capture the concept of UX through a set of questions or items. The items are grouped into several dimensions or scales. Each scale represents a distinct UX aspect, for example efficiency, learnability, novelty or stimulation.

A number of such questionnaires exist. Questionnaires related to pure usability aspects are described, for example, in [8], [9]. Questionnaires covering the broader aspect of UX are, for example, described in [10], [11], and [12]. Each questionnaire contains different scales for measuring groups of UX aspects. So the choice of the best questionnaire depends on an evaluation study's research question, i.e. on the quality aspects to measure. For broader evaluations, it may make sense to use more than one questionnaire.

One of the problems in using UX questionnaires is how to interpret results, if no direct comparison is available. Assume that a UX questionnaire is used to evaluate a new program version. If a test result from an older version exists, the interpretation is easy. The numerical scale values of the two versions can be compared by statistical test to show whether the new version is a significant improvement.

However, in many cases the question is not "*Is UX of the evaluated product better than UX of another product or a previous version of the same product?" but "Does the product show sufficient UX?*" So there is no separate result to compare with. This is typically the case when a new product is released for the first time. Here it is often hard to interpret whether a numerical result, for example a value of 1.5 on the *Efficiency* scale, is sufficient. This is the typical situation where a benchmark, i.e. a collection of measurement results from a larger set of other products, is helpful.

In this paper we describe the construction of a benchmark for the User Experience Questionnaire (UEQ) [12], [13]. This benchmark helps interpret measurement results. The benchmark is especially helpful in situations where a product is measured with the UEQ for the first time, i.e. without results from previous evaluations.

II. The User Experience Questionnaire (UEQ)

A. Goal of the UEQ

The main goal of the UEQ is a fast and direct measurement of UX. The questionnaire was designed for use as part of a normal usability test, but also as an online questionnaire. For online use, it must be possible to complete the questionnaire quickly, to avoid participants not finishing it. So a semantic differential was chosen as item format, since this allows a fast and intuitive response.

Each item of the UEQ consists of a pair of terms with opposite meanings.

Examples:

Not understandable o o o o o o o *Understandable*

Efficient o o o o o o o *Inefficient*

Each item can be rated on a 7-point Likert scale. Answers to an item therefore range from -3 (fully agree with negative term) to +3 (fully

DOI: 10.9781/ijimai.2017.445

agree with positive term). Half of the items start with the positive term, the rest with the negative term (in randomized order).

B. Construction process

The original German version of the UEQ uses a data analytics approach to ensure the practical relevance of the constructed scales. Each scale represents a distinct UX quality aspect.

An initial set of more than 200 potential items related to UX was created in two brainstorming sessions with two different groups of usability experts. A number of these experts then reduced the selection to a raw version with 80 items. The raw version was used in several studies on the quality of interactive products, including a statistics software package, cell phone address books, online collaboration software or business software.

In these studies, 153 participants rated the 80 items. Finally, the scales and the items representing each scale were extracted from this data set by principal component analysis [12], [13].

C. Scale structure

This analysis produced the final questionnaire with 26 items grouped into six scales:

- *Attractiveness*: Overall impression of the product. Do users like or dislike it? Is it attractive, enjoyable or pleasing?

  6 items: *annoying / enjoyable, good / bad, unlikable / pleasing, unpleasant / pleasant, attractive / unattractive, friendly / unfriendly*.

- *Perspicuity*: Is it easy to get familiar with the product? Is it easy to learn? Is the product easy to understand and clear?

  4 items: *not understandable / understandable, easy to learn / difficult to learn, complicated / easy, clear / confusing*.

- *Efficiency*: Can users solve their tasks without unnecessary effort? Is the interaction efficient and fast? Does the product react fast to user input?

  4 items: *fast / slow, inefficient / efficient, impractical / practical, organized / cluttered.*

- *Dependability*: Does the user feel in control of the interaction? Can he or she predict the system behavior? Does the user feel safe when working with the product?

  4 items: *unpredictable / predictable, obstructive / supportive, secure / not secure, meets expectations / does not meet expectations*.

- *Stimulation*: Is it exciting and motivating to use the product? Is it fun to use?

  4 items: *valuable / inferior, boring / exciting, not interesting / interesting, motivating / demotivating*.

- *Novelty*: Is the product innovative and creative? Does it capture users' attention?

  4 items: *creative / dull, inventive / conventional, usual / leading-edge, conservative / innovative.*

Scales are not assumed to be independent. In fact, a user's general impression is captured by the *Attractiveness* scale, which should be influenced by the values on the other 5 scales (see Fig. 1).

Attractiveness is a pure valence dimension. *Perspicuity*, *Efficiency* and *Dependability* are pragmatic quality aspects (goal-directed), while *Stimulation* and *Novelty* are hedonic quality aspects (not goal-directed) [14].

Applying the UEQ does not require much effort. Usually 3-5 minutes are sufficient for a participant to read the instructions and complete the questionnaire. The UEQ can either be used in a paper-pencil form as part of a classical usability test (and this still is the most common application), but also as an online questionnaire.

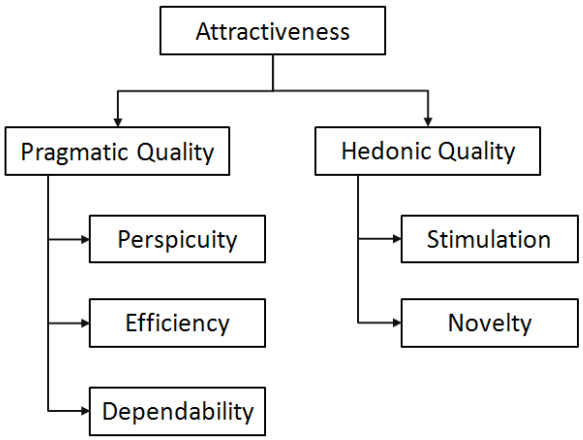

Fig. 1. Assumed scale structure of the User Experience Questionnaire (UEQ).

D. Validation

The reliability (i.e. the consistency of the scales) and validity (i.e. that scales really measure what they intend to measure) of the UEQ scales was investigated in several usability tests with a total of 144 participants and an online survey with 722 participants. These studies showed a sufficient reliability of the scales (measured by Cronbach's Alpha). In addition, several studies have shown a good construct validity of the scales. For details see [12], [13].

E. Availability and language versions

For a semantic differential like the UEQ, it is very important that participants can fill it out in their natural language. Thus, several contributors created a number of translations.

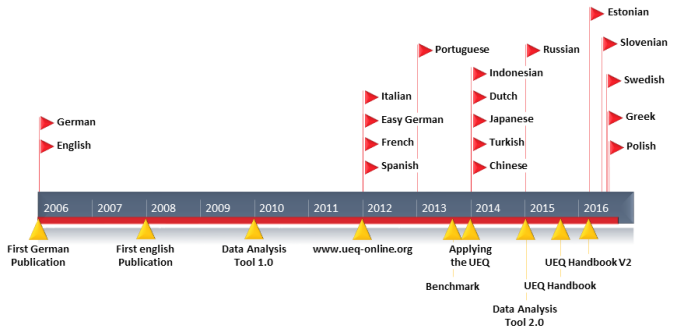

Fig. 2. Timeline of UEQ development.

The UEQ is currently available in 17 languages (German, English, French, Italian, Russian, Spanish, Portuguese, Turkish, Chinese, Japanese, Indonesian, Dutch, Estonian, Slovene, Swedish, Greek and Polish).

The UEQ in all available languages, an Excel sheet to help with evaluation, and the UEQ Handbook are available free of charge at www.ueq-online.org.

Helpful hints on using the UEQ are also available from Rauschenberger et al. [15].

III. Why do we need a benchmark?

The goal of the benchmark is to help UX practitioners interpret scale results from UEQ evaluations.

Where only a single UEQ measurement exists, it is difficult to judge

whether the product fulfills the quality goals. See Fig. 3 as an example of an evaluation result.

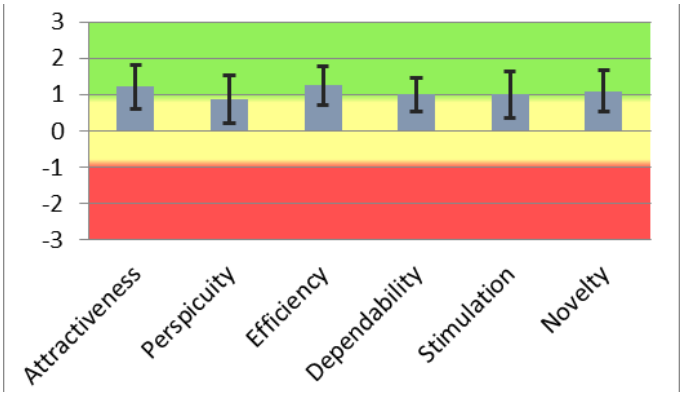

Fig. 3. Example chart from the data analysis Excel sheet showing the observed scale values and error bars for an example product.

Is this a good or bad result? Scale values above 0 represent a positive evaluation of the quality aspect; values below 0 represent a negative evaluation. But what does this actually mean? How do other products score?

If we have, for example, a comparison to a previous version of the same product or to a competitor product, then it is easy to interpret the results.

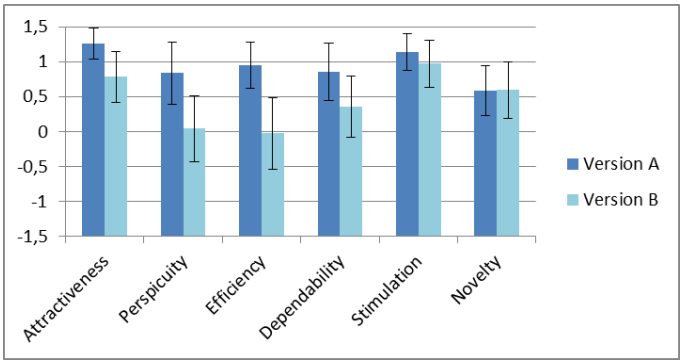

Fig. 4. Comparison between two different products. Here it is much easier to interpret the results, since the mean scale values can be directly compared.

A simple statistical test, for example a t-test, can be used to find out whether version A shows a significantly higher UX than version B.

But when a new product is launched, a typical question is whether the product's UX is sufficient to fulfill users' general expectations. Obviously no comparison to previous versions is possible in this case. It is also typically not possible to get evaluations of competitor products. The same is true for a product that has been on the market for a while, but is being measured for the first time.

Users form expectations of UX during interactions with typical software products. These products need not belong to the same product category. For example, users' everyday experience with modern websites and interactive devices, like tablets or smartphones, has also heavily raised expectations for professional software, such as business applications. So if a user sees a nice interaction concept in a new product, which makes difficult things easier, this will raise his or her expectations for other products. A typical question in such situations is: "*Why can't it be as simple as in the new product?*".

Thus, the question whether a new product's UX is sufficient can be answered by comparing its results to a large sample of other commonly used products, i.e. a benchmark data set. If a product scores high compared to the products in the benchmark, this can indicate that users will generally find the product's UX satisfactory.

IV. Construction of the Benchmark

Over the last couple of years, such a benchmark was created for the UEQ by collecting data from all available UEQ evaluations. The benchmark was only made possible by a huge number of contributors, who shared the results of their UEQ evaluation studies. Some of the data comes from scientific studies using the UEQ, but most of the data comes from industry projects.

The benchmark currently contains data from 246 product evaluations using the UEQ. These evaluated products cover a wide range of applications. The benchmark contains complex business applications (100), development tools (4), web shops or services (64), social networks (3), mobile applications (16), household appliances (20) and a couple of other (39) products.

The benchmark contains a total of 9,905 responses. The number of respondents per evaluated product varied from extremely small samples (3 respondents) to huge samples (1,390 respondents). The mean number of respondents per study was 40.26.

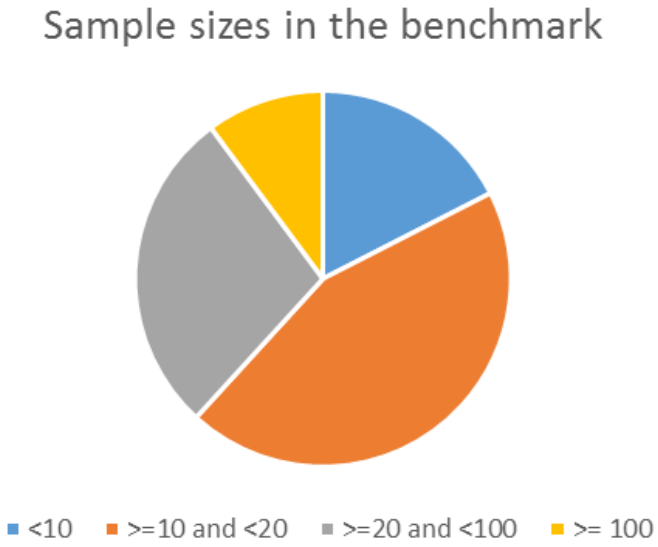

Fig. 5. Distribution of the sample sizes in the benchmark data set.

Many evaluations were part of usability tests, so the majority of the samples had less than 20 respondents (65.45%). The samples with more than 20 respondents were usually collected online.

Of course, the studies based on tiny samples with fewer than 10 respondents (17.07%) do not carry much information. It was therefore verified whether these small samples had an influence on the benchmark data. Since the results do not change much when studies with less than 10 respondents are eliminated, it was decided to keep them in the benchmark data set.

The mean values and standard deviations (in brackets) of the UEQ scales in the benchmark data set are:

- Attractiveness: 1.04 (0.64)
- Efficiency: 0.97 (0.62)
- Perspicuity: 1.06 (0.67)
- Dependability: 1.07 (0.52)
- Stimulation: 0.87 (0.63)
- Originality: 0.61 (0.72)

Nearly all of the data comes from evaluations of mature products, which are commercially developed and designed. Thus, it is no surprise that the mean value is above the neutral value (i.e. 0) of the 7-point Likert scale.

Since the benchmark data set currently contains only a limited number of evaluation results, it was decided to limit the feedback per scale to 5 categories:

- *Excellent*: The evaluated product is among the best 10% of results.
- *Good*: 10% of the results in the benchmark are better than the evaluated product, 75% of the results are worse.
- *Above average*: 25% of the results in the benchmark are better than the evaluated product, 50% of the results are worse.
- *Below average*: 50% of the results in the benchmark are better than the evaluated product, 25% of the results are worse.
- *Bad*: The evaluated product is among the worst 25% of results.

Table 1 shows how the categories relate to observed mean scale values.

TABLE I
BENCHMARK INTERVALS FOR THE UEQ SCALES

	Att.	**Eff.**	**Per.**	**Dep.**	**Sti.**	**Nov.**
Excellent	≥ 1.75	≥ 1.78	≥ 1.9	≥ 1.65	≥ 1.55	≥ 1.4
Good	≥ 1.52 < 1.75	≥ 1.47 < 1.78	≥ 1.56 < 1.9	≥ 1.48 < 1.65	≥ 1.31 < 1.55	≥ 1.05 < 1.4
Above average	≥ 1.17 < 1.52	≥ 0.98 < 1.47	≥ 1.08 < 1.56	≥ 1.14 < 1.48	≥ 0.99 < 1.31	≥ 0.71 < 1.05
Below average	≥ 0.7 < 1.17	≥ 0.54 < 0.98	≥ 0.64 < 1.08	≥ 0.78 < 1.14	≥ 0.5 < 0.99	≥ 0.3 < 0.71
Bad	< 0.7	< 0.54	< 0.64	< 0.78	<0.5	< 0.3

The comparison to the benchmark is a first indicator for whether a new product offers sufficient UX to be successful in the market. It is sufficient to measure UX by a large representative sample of users. Usually 20-30 users already provide a quite stable measurement. Comparing the different scale results to the products in the benchmark allows conclusions regarding the relative strengths and weaknesses of the product.

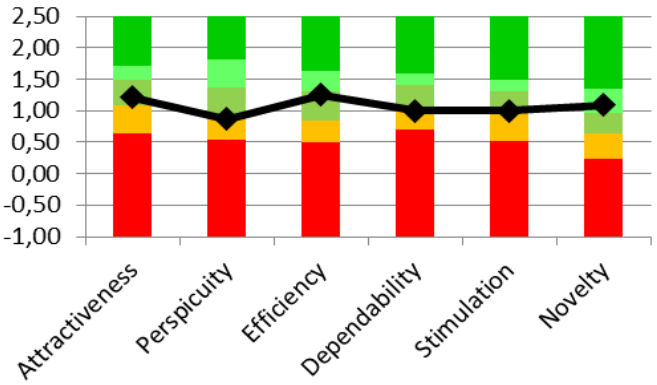

Fig. 6. Visualization of the benchmark in the data analysis Excel sheet of the UEQ. The line represents the results for the evaluated product. The colored bars represent the ranges for the scales' mean values.

It must be noted that the general UX expectations have grown over time. Since the benchmark also contains data from established products, a new product should reach at least the *Good* category on all scales.

V. Benchmark as part of quality assurance

A UX benchmark can be a natural part of the quality assurance process for a new product. Assume that a new product is planned. The crucial quality aspects for a successful launch can easily be identified according to the product type and the intended market positioning. These identified quality aspects should reach a very good value in a later UEQ evaluation.

Let us assume that a new Web application should be developed. Users should be able to handle this application intuitively, without help or reading of documentation, to order services over the Web. The new application's design should be original and unconventional to grab users' attention. In addition, it should not be boring to use, so that users will come back.

In this example it is clear that *Perspicuity*, *Originality* and *Stimulation* are the most important UX aspects. So it would be a natural goal for the application to reach the *Excellent* category on these scales and at least an *Above Average* on the other UEQ scales. A benchmark – together with a clear idea of the importance of the UX quality aspects / UEQ scales – can help define clear and understandable quality goals for product development. These goals can easily be verified by using the UEQ questionnaire later on.

VI. Conclusion

We described the development of a benchmark for the User Experience Questionnaire (UEQ). This benchmark helps interpret UX evaluations of products. It is currently available in 17 languages at www.ueq-online.org inside the "UEQ Data Analysis Tool" Excel file. The benchmark is especially helpful in situations where a product is measured for the first time with the UEQ, i.e. where no results from previous evaluations exist for comparison. In this article we also described how the benchmark can be used to formulate precise and transparent UX quality goals for new products.

A weakness of the current benchmark is that it does not distinguish between different product categories, i.e. there is only one benchmark data set for all types of products. Since most of the data in the benchmark comes from business applications or websites, it may be difficult to use for special applications or products, such as games, social networks or household appliances. The quality expectations for such types of products may simply be quite different from those expressed in the benchmark.

In the future we will try to create different benchmarks for different product categories. However, this requires collecting a larger number of data points per product category in UEQ evaluations and will therefore take some time.

References

[1] Schrepp, M.; Hinderks, A. & Thomaschewski, J. (2014). Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In: Marcus, A. (Ed.): Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience. Lecture Notes in Computer Science, Volume 8517, S. 383-392, Springer International Publishing.

[2] Nielsen, J. (1994). Usability engineering. Elsevier.

[3] Preece, J., Rogers, Y.; Sharpe, H. (2002): Interaction design: Beyond human-computer interaction. New York: Wiley.

[4] Rieman, J., Franzke, M., & Redmiles, D. (1995, May). Usability evaluation with the cognitive walkthrough. In Conference companion on Human factors in computing systems (pp. 387-388). ACM.

[5] Nielsen, J. (1992, June). Finding usability problems through heuristic evaluation. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 373-380). ACM.

[6] Nielsen, J. (1994, April). Enhancing the explanatory power of usability heuristics. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 152-158). ACM.

[7] Nielsen, J. (1994): Heuristic Evaluation. In: J. Nielsen; R.L. Mack (Eds.): Usability Inspection Methods. New York: Wiley. S. 25-62.

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

[9] Kirakowski, J.; Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. British Journal of Educational Technology, Vol. 24, Nr. 3, S. 210–212.

[10] Hassenzahl, M.; Burmester, M.; Koller, F. (2003): AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. [AttrakDiff: A questionnaire to measure perceived hedonic and pragmatic quality] In: J.Ziegler; G. Szwillus (Eds.): Mensch & Computer 2003. Interaktion in Bewegung. Stuttgart: Teubner. S. 187-196.

[11] Visual Aesthetics of Websites Inventory (Moshagen, M. & Thielsch, M. T. (2010). Facets of visual aesthetics. International Journal of Human-Computer Studies, 68 (10), 689-709.).

[12] Laugwitz, B.; Schrepp, M. & Held, T. (2006). Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. [Construction of a questionnaire for the measurement of user experience of software products] In: A.M. Heinecke & H. Paul (Eds.): Mensch & Computer 2006 – Mensch und Computer im Strukturwandel. Oldenbourg Verlag, S. 125 – 134.

[13] Laugwitz, B.; Schrepp; M. & Held, T. (2008). *Construction and evaluation of a user experience questionnaire.* In: Holzinger, A. (Ed.): USAB 2008, LNCS 5298, pp. 63-76.

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

[15] Rauschenberger, M.; Schrepp, M.; Perez-Cota, M.; Olschner, S.; Thomaschewski, J. (2013): Efficient Measurement of the User Experience of Interactive Products. How to use the User Experience Questionnaire (UEQ). Example: Spanish Language Version. In: IJIMAI, 2(1), pp. 39-45.

Martin Schrepp has been working as a user interface designer for SAP AG 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 in these research fields.

Andreas Hinderks holds a diploma in Computer Science and is Master of Science in Media Informatics by University of Applied Science Emden/Leer. He has worked as a Business Analyst and a programmer from 2001 to 2016. His focus then lay on developing user-friendly business software. Currently, he is a freelancing Business Analyst and Senior UX Architect. Also, he is a Ph.D. student at the University of Applied Science Emden/Leer. He is involved in research activities dealing with UX questionnaires, process optimization, information architecture, and user experience since 2011.

Jörg Thomaschewski was born in 1963. He 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 Internet applications for human-computer interaction, e-learning, and software engineering. Dr. Thomaschewski is the author of various online modules, e.g., “Human-Computer Communication,” which are used by the Virtual University (online) at six university sites. He has wide experience in usability training, analysis, and consulting.