Human-Centered Design of Mixed Reality Applications in Medical Education – GreifbAR

Robert Luzsa[1,2], Moritz Queisner[3], Christopher Remde[3], Igor Sauer[3], Nadia Robertini[4] and Susanne Mayr[1]

[1] University of Passau

[2] University of Applied Sciences Landshut

[3] Charité – Universitätsmedizin Berlin

[4] German Research Center for Artificial Intelligence (DFKI)

This is the author's version of the following chapter. Please refer to and cite the published version:

Luzsa, R., Queisner, M., Remde, C., Sauer, I., Robertini, N. & Mayr, S. (2025). Human-Centered Design of Mixed Reality Applications in Medical Education – GreifbAR. In J. Tröge, J. Stepczynski, H. Wiesner & C. Runde (Eds.). *Virtuelle Beteiligung, reale Teilhabe - Transformative Technologien für eine inklusivere Gesellschaft* (pp. 385 – 408). Campus. https://doi.org/10.12907/978-3-593-46260-8

Human-Centered Design of Mixed Reality Applications in Medical Education – GreifbAR

Robert Luzsa, Moritz Queisner, Christopher Remde, Igor Sauer, Nadia Robertini and Susanne Mayr

Project and Chapter Description

The project „Tangible Reality – Interaction of Hands with Physical Tools in Mixed Reality“ investigates how the user's hands can be integrated into virtual reality applications to learn and train work-related tasks with physical objects. As part of the „Interactive systems in virtual and real spaces“ program funded by the German Federal Ministry of Education and Research, the project developed mixed reality (MR) prototypes to improve and facilitate the acquisition of practical skills. Based on a cooperation between surgery (Charité – Universitätsmedizin Berlin), psychology (Universität Passau), artificial intelligence (German Research Center for Artificial Intelligence) and interaction design (NMY Mixed-Reality Communication GmbH) we found that spatial and tangible interaction in mixed reality systems helped trainees to acquire practical knowledge in medical training when learning tasks such as tying knots or using instruments. This chapter first introduces the project. Then, it gives an overview of methods that were used to assess the perspectives of trainees and experts during the project. These are integrated into a generalized model for the human-centered design of MR-based learning tools that may inspire similar projects. Finally, three studies carried out as part of the project are presented and their results and implications discussed.

Abstract

The capabilities of mixed reality (MR) applications to combine visualizations with a user's real environment offer benefits for the training of complex surgical skills like the tying of surgical knots. By wearing a virtual reality headset combined with a passthrough function, users can see additional information that adapts to their body or physical objects, such as visual instructions to perform a work task or they can receive individualized feedback based on their movement. To ensure usability, user acceptance and effective learning outcomes of MR applications, it is essential to assess the knowledge of domain experts as well as to derive requirements from state-of-the-art training concepts. This chapter provides an overview of tools and challenges associated with integrating users into the development of medical MR applications. Drawing on findings and insights gained during the development of an MR-based knot-tying trainer in the Tangible Reality project, the chapter offers recommendations for effectively incorporating users into the design of similar MR applications.

1. Human-Centered Design for Surgical Skills Training

In the development of innovative digital applications, various approaches have emerged that highlight the importance of involving users and other stakeholders in the design and development process. The terms user-centered design (UCD) and human-centered design (HCD) were coined to describe a way of designing systems that are not primarily defined by the capabilities of a technology but instead optimized to fit the needs of its users (Norman, 1986: 32 ff.). A key consequence of this conceptualization of design is that users and stakeholders should not only evaluate a system once it is fully developed. Instead, they should be involved in all phases of the design process – from identifying needs, generating ideas, and developing concepts through to iterative testing of increasingly high-fidelity prototypes (International Organization for Standardization, 2019: 6; Stanton et al., 2013: 2 ff.).

The integration of user perspectives in all phases of development is particularly important in scenarios where specialized expert user groups are involved and where comprehensibility, precision and reliability of the system are critical. A prime example for this is the field of surgical skills training, where systems are developed to teach surgeons skills that are necessary, for example in laparoscopic surgery and knot tying (Hagelsteen et al., 2019: 309 ff.). To design effective surgical training applications, it is essential to consider not only explicit information about the surgical skills to be taught but also implicit standards in surgical practice which are not explicitly outlined in textbooks. These implicit standards encompass, for example, the knowledge and experience that experts utilize and rely on in their daily work. Additionally, it is crucial to understand the specific needs of the user group and the capabilities of the training technology used. Yet, what is the best way to obtain all this information, and what are the best methods for user involvement?

This chapter provides answers to this question by describing the process and methods of user involvement in the development of a mixed reality-based (MR) training application for surgical knot tying.

2. The Tangible Reality Project

The Tangible Reality project aims to bridge the gap between the real and virtual worlds. It focuses on the creation of mixed reality environments that are tangible and graspable, incorporating both virtual and augmented reality. Our major goal in the project is to enable users to interact with real objects, virtually recreated real objects and purely virtual objects using their bare hands seamlessly and naturally, without the need for additional controllers.

This chapter describes a case study that focuses on the development of an MR training application that teaches surgical knot-tying techniques to medical students. The skilled handling of sutures is one of the basics of any surgical procedure. Usually, surgical knotting techniques are learnt through personal instruction by experienced surgeons, diagrams, and video-based tutorials. Learning surgical knot tying techniques through participatory observation and direct instruction by experienced surgeons is costly and difficult to scale. This type of training is increasingly reaching its limits in day-to-day hospital operations, due to the changing economic, social, and regulatory conditions of surgical practice. Students and trainees, as well as professionals in training, are therefore faced with the problem of applying and practicing the theoretical knowledge they have acquired. Sometimes, universities offer workshops in which experienced surgeons demonstrate their knotting techniques. Although such workshops are didactically valuable due to the intensive support in small groups and individual feedback, they are time-consuming and resource-intensive, making them impractical on a larger scale.

Students may rely on textbooks for theoretical knowledge and videos for observational learning. Text- and image-based media allow a scalable acquisition of theoretical knowledge independent of time and place. Yet, this way of learning lacks interactivity, personalization, feedback, and expert guidance, making it insufficient for effective skill acquisition. Knot tying gestures and work steps can only be passively observed and then imitated. Furthermore, learning success cannot be measured and verified. In the project, we concluded that MR-based training could combine the benefits of self-guided, autonomous, and instructor-led learning: Users could wear an MR headset while tying knots on their own, and the headset would display instructions and provide individualized feedback based on hand and finger tracking. This would allow users to train independently, without the need for an instructor, while still receiving individual guidance.

Yet, how should such an MR-based learning system be designed so that learners benefit from it, accept it, and enjoy working with it? And what information should be conveyed, that is, what knotting techniques surgeons use and teach in workshops, and with which means (visual, acoustic etc.) should the system present information and provide feedback?

These questions can only be answered by involving users and experts. For this reason, in the project, existing models and procedures of user involvement were first examined and integrated into a generalized model for the user-centered design of the MR-based learning systems. This model and its scientific foundations are presented in sub-chapter „Stages of Human-Centered MR Development“. Based on this, two studies to assess user and expert perspectives were conducted, which inspired the development of different design variants of the learning system. Finally, students rated and evaluated these variants in an online experiment. The results and the technical details of the developed systems are presented in the final sub-chapter „Case Study: Development and User Testing of a Mixed Reality Knot Tying Trainer“.

3. Stages of Human-Centered MR Development

3.1 Existing Models and Approaches

There are various procedural models for the human-centered development of products and services that can be applied in the context of MR-based surgical learning. However, they all focus only on certain parts of the development process and reflect the research and design tradition from which they originate. To effectively guide the development of surgical training applications, these perspectives must be integrated and combined. To achieve this, the following section introduces three models with different focuses, which are subsequently merged into a comprehensive model. This model encompasses all stages of design and development, from analyzing the current state to evaluating the individual and organizational impacts of application deployment.

3.1.1 ISO 9241-210

The ISO 9241-210 suggests four stages for the human-centered design of digital systems. These should be followed to achieve efficient and effective systems, usability, and well-being of users while minimizing adverse effects on their health, safety, and performance (International Organization for Standardization, 2019: 10 f.). The stages are:

1) *Context analysis:* The usage context of the systems needs to be defined. This includes, for example, characteristics of future users, activities to be performed, workplace conditions and environments, such as available space, lighting, and noise levels. Literature research, e. g., on similar projects and products, should also be conducted at this stage.
2) *Definition of requirements:* Concrete requirements are formulated, from the perspective of future system users, the organization in which the system will be deployed, and other stakeholders (e. g., patients in the domain of medicine).
3) *Design:* Based on contextual knowledge and requirements, increasingly high-fidelity prototypes are created, from pen-and-paper mock-ups and wireframes up to the final working product.
4) *Evaluation:* The evaluation is conducted iteratively, alternating with step three. Users and stakeholders test prototypes and provide feedback, which is integrated into design.

3.1.2 The Double Diamond Model

A similar approach is the double diamond model. It originates from design thinking and aims to involve users and stakeholders to solve problems and develop ideas that meet user needs and market demands (Lindberg et al., 2011: 4 ff.). In the model, divergent and convergent thinking alternate: In the first stage, ideas are generated (divergent *discover* stage), which are then selected (convergent *define* stage), after which solutions are developed (divergent *develop* stage) and finally implemented (convergent *deliver* stage). Hence the term double diamond, as the divergent stages can be visualized as arrows pointing outwards and the convergent stages as arrows pointing inwards, forming two diamonds when put together. The stages are iterative, meaning that stages can be repeated with additional insights until a satisfactory final product or system is reached.

Both models emphasize the importance of stakeholder integration and iterative development. Following these steps is especially important during the development of surgical training tools: Designers and developers are usually unfamiliar with surgical procedures and terms, and learners' surgical knowledge and experience need to be considered. Previous studies have shown that a human-centered design process can improve the usability of training applications in surgery, for example in ventriculostomy (Reyes et al., 2022: 52 ff.) and minimally invasive surgery (Freudenthal et al., 2011: 198 ff.).

3.1.3 The Kirkpatrick Evaluation Model

The models discussed so far provide helpful approaches for application development yet overlook one key aspect: What happens after a system is developed and deployed? How does it impact users and the organization in real-world work or learning contexts? This is especially crucial for medical and surgical learning applications: These should not only offer benefits for users and patients but must also be successfully integrated into the hospital or academic teaching organization. For this reason, a further model was used in the project that focuses on the successful roll-out of developments in real-world environments, namely Kirkpatrick's four-level evaluation model (Kirkpatrick, 1996: 55 ff.; Smidt et al., 2009: 267 ff.). It assesses four levels of effects:

1) *Reaction*: Users' immediate reactions to the training application.
2) *Learning*: The effectiveness of the system in teaching relevant skills.
3) *Behavior*: How users apply learned skills in their work.
4) *Results*: The organizational impact, such as increased productivity or reduced errors.

Combining these levels of effects with the development stages of the models described above, it becomes clear that reactions can primarily be measured in the context of early iterative user tests with mock-ups and prototypes. Learning effectiveness can be measured with summative evaluations with the final prototype or product. However, long-term changes in behavior and organizational outcomes can only be observed after the product has been rolled out in the organization and integrated into learning routines and contexts. For medical and surgical training scenarios, this means assessing the on-the-job performance of nurses or surgeons on tasks they have learned using a new training application. In addition, evaluations from different user groups can be gathered (e. g. patient satisfaction, employee satisfaction, supervisor ratings) and performance measures such as number of errors may be used (Praslova, 2010: 219 ff.). Based on this, recommendations can be made for the use of the system and any necessary steps such as further qualifications of users.

3.2 A Generalized Model of Human-Centered Design and Testing

The three models described above can be combined into a generalized, three-stage, iterative process model that encompasses all phases of the design, development, testing, and implementation of a system (see figure 1).

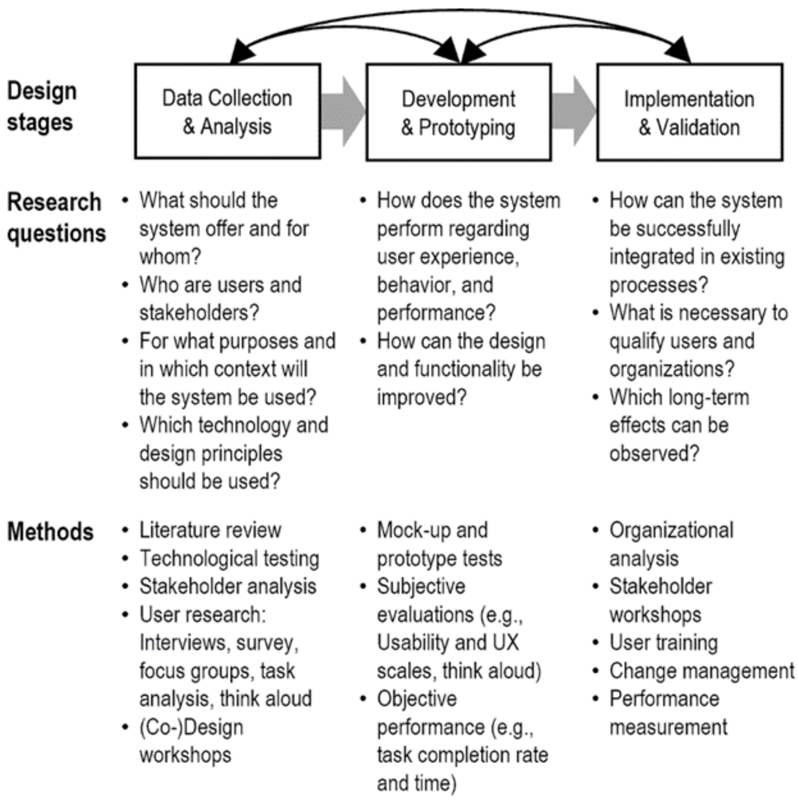

Figure 1: A generalized model of human-centered design and testing (Source: Own figure)

The first stage of the model – data collection and analysis – encompasses all activities that focus on learning about users, their needs, perspectives, workflows, and usage scenarios, as well as exploring the technological possibilities for the development project. It summarizes the first two stages of the double diamond model and the ISO 9421-210. The second phase – development and prototyping – focuses on the actual development, ideally of different variants, which are then subjected to user testing and compared concerning different acceptance and success criteria. It summarizes the third and fourth stages of the above models. In the final stage – implementation and validation – the system is introduced into working and learning environments, to measure its real-world impact on and to derive recommendations for further action. There may be feedback loops between the stages, for example lessons learned during development and testing which may require additional data collection and field research, or the identification of additional needs during implementation inspiring further developments. These feedback loops are shown as arrows in figure 1.

4. Methods for Human-Centered MR Development

Once the relevant stages of development have been identified, suitable methods need to be selected to address the research questions in each stage, for example observational techniques, user testing or ideation methods. In addition, measures of success must be selected that provide valid, objective, and reliable information about the performance and user acceptance of developed mock-ups and prototypes, usually in the form of standardized questionnaire scales (DeVellis & Thorpe, 2021: 6 ff.; Stanton et al., 2013: 11 ff.).

There are several overviews of methods for user-centered product development that researchers and designers of MR-based learning systems can refer to. Stanton et al. (2013: 39 ff.) describe methods from human factors research and engineering psychology that are suitable for analyzing user tasks, interaction patterns with technical systems, and testing prototypes and user interfaces. These methods include, for example, hierarchical task analysis, which observes and breaks down work tasks into smaller steps to optimize technological solutions, the human error template, which identifies possible sources of error in activities, and heuristic analysis, where usability experts evaluate a system using checklists and templates.

Stickdorn et al. (2018: 36 ff.) and Tidwell et al. (2020: 1 ff.) describe methods from design thinking, which primarily incorporate user perspectives into design decisions. Prominent methods here include customer journey maps, which record and visualize user interactions with a system, the development of personas, detailed characterizations of user groups regarding their characteristics (e. g., prior knowledge, skills, socio-demographics, requirements), and co-creation workshops, in which users design drafts for user interfaces.

The following four methods are described in detail as they are general and applicable to the development of almost any product or technological system. They also appear particularly suitable for the design of MR-based surgical training applications because their combined use can provide both extensive qualitative and quantitative insights during all phases of development and because they can be easily adapted towards different target groups such as expert surgeons or students.

4.1 Stakeholder Analysis

The first stage of design is to identify the stakeholders, that is, any person, group, or organization that is directly or indirectly involved or affected by the development project (Bryson, 2004: 22). This is especially important for a surgical training application, as stakeholders include not only trainee surgeons using the system but also experienced surgeons and teachers who may provide learning content, and patients, whose needs should also be reflected in training materials. A common method for identifying stakeholders is stakeholder mapping, where all possible stakeholders are brainstormed and clustered according to different criteria. The involvement of each stakeholder group in the development process is discussed, for example whether they should provide information or evaluate prototypes.

4.2 Think Aloud

Think aloud can be used in the first design stage to understand complex tasks, such as surgical skills, and in the second stage to evaluate users' interactions with prototypes. Participants perform a task or interact with a system while continually verbalizing their thoughts and impressions. The verbalizations are recorded and analyzed (e. g., via qualitative content analysis, see Kuckartz, 2022: 15 ff.), providing insights into detailed task descriptions for integration into a training application or lists of usability problems to be addressed.

Compared to post-hoc interviews, think aloud provides more comprehensive and detailed insights as participants verbalize their thoughts while performing tasks. It is also very well suited to understanding procedural knowledge and actions that experts perform automatically and that they would not normally verbalize because they take them for granted. This is the case in many medical fields, including surgical knot tying, where think aloud can help identify the knot tying steps and understand requirements for correct knotting.

Detailed instructions and probing questions are essential for think aloud, as participants are often not used to verbalizing their thoughts continuously. Researchers might demonstrate the concept by thinking aloud themselves (e. g., counting windows in a building) and prepare probing questions for when participants stop verbalizing (Charters, 2003: 71 ff.).

4.3 The Technology Acceptance Model

The Technology Acceptance Model (TAM) (Davis, 1989: 319 ff.) provides measures of success to determine users' acceptance of prototypes and their perceived strengths and weaknesses. Moreover, it can assess relevant characteristics of the user group.

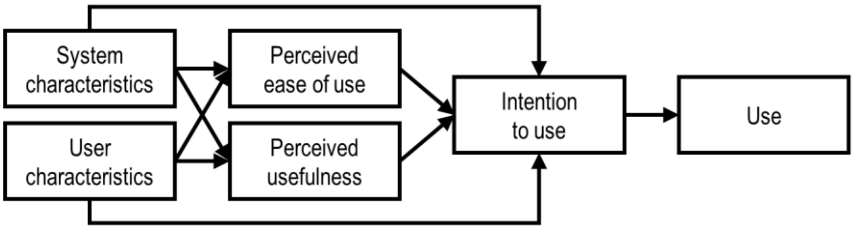

Figure 2: The technology acceptance model including influencing factors (Source: Own figure, adapted from Venkatesh & Bala, 2008: 276)

TAM conceptualizes users' intention to use a system as determined by perceived ease of use (how easy it is to use and interact with the system) and perceived usefulness (how efficiently and effectively the system helps users achieve their goals, see figure 2). Using standardized scales, these facets of technology acceptance can be measured during prototype evaluation to identify weaknesses and areas for improvement, such as whether perceived ease of use is rated high or low.

Additionally, the facets of technology acceptance are influenced by factors that originate from both the users (e. g., how competent they rate themselves when using technology) and the system (e. g., how well it provides instructions and feedback). Measuring these influences can help to identify relevant user needs and user groups. For example, if users' acceptance depends on their technological competence, the system could be optimized to provide different interaction modes and functions tailored to the needs of more or less technologically competent users. A general overview of user- and system-related antecedents of technology acceptance and scales to measure them is given by Venkatesh and Bala (2008: 275 ff.) and Mlekus et al. (2020: 275 ff.). Moreover, for MR-based training applications, Graser and Böhm (2022: 3 ff.) provide a review of 33 studies that identify factors influencing system acceptance, ranging from the system's interface style to learners' perceptions of self-efficacy and teachers' attitudes towards integrating technology into training. Granić (2022: 9733 ff.) also provides a general overview of research on factors influencing acceptance of educational technology.

4.4 The User Experience Questionnaire

Finally, the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008: 63 ff.) is another standardized questionnaire scale, which can be used during prototype testing to get a more detailed picture of the usability and user experience of the system. Usability refers to the degree to which a product can support the user in a specific context to achieve goals in an effective, efficient, and satisfactory manner (International Organization for Standardization, 2018: 6). The concept of user experience is broader and encompasses user perceptions and reactions resulting from the use of a product, system or service (International Organization for Standardization, 2019: 7). It adds emotional and hedonic aspects such as enjoyment to the concept of usability. The standard version of the UEQ uses a semantic differential to measure six facets of usability and user experience: Attractiveness (e. g. annoying vs. enjoyable), perspicuity (e. g. not understandable vs. understandable), dependability (e. g. unpredictable vs. predictable), efficiency (e. g. obstructive vs. supportive), novelty (e. g. dull vs. creative), and stimulation (e. g. boring vs. exciting)

After averaging the items of each scale, a profile can be created and the rating of each aspect can be compared to identify strengths and weaknesses. The UEQ also facilitates the comparison of several variants, e. g. to find the best scoring variant of a user interface. Additionally, there are benchmark values for each scale that are based on 246 product evaluations (Schrepp et al., 2017a: 42 ff.). The results of a user test can be compared with the benchmark values and conclusions can be drawn as to whether the usability and user experience dimensions are significantly below average, and whether the system needs to be revised accordingly.

If only a general assessment of the user experience is required, and the 26 questions of the full UEQ are too comprehensive, the short version of the UEQ-S with only eight questions can be used as provided by Schrepp et al. (2017b: 105). This measures pragmatic and hedonic quality and an overall assessment of the system.

Some scales of UEQ and TAM measure closely related concepts (Mlekus et al., 2020: 276 f.), therefore it is recommendable to decide for each development project which TAM and UEQ scales are most relevant, and to combine them into an appropriate questionnaire.

5. Case Study: Development and User Testing of a Mixed Reality Knot Tying Trainer

This chapter focuses on how the design and evaluation methods were applied in the Tangible Reality project for the development of an MR-based knot tying trainer. It gives an overview of three studies that took place in stage one (data collection and analysis) and stage two (development and prototyping) of development.[1] It also describes the MR-based learning system and how it builds on findings from user research and available technology.

[1] We cannot report on results in implementation and validation as development is still going on at the time of writing.

5.1 Status Quo: Students' and Experts' perspectives on surgical knot tying

5.1.1 User Study: Assessing students' perspectives

The first study investigated how medical students currently learn to tie surgical knots, which knot-tying tasks and techniques they find most important and difficult, and what they expect from an MR-based knot-tying training application.

Method: We conducted an exploratory online questionnaire study with 80 medical students from Charité – Universitätsmedizin Berlin in October and November 2022. The participants were mostly female (56 female, 22 male, 2 diverse), with a mean age of 24.56 years (*SD* = 4.5, *range* = 18–41). Most were in the practice-oriented phase of their studies (44 in the fourth to tenth semesters), with the rest in the theory-oriented phase (18 in the first to fourth semesters) or the final practical year (16).

Participants were informed about the aims and procedures of the study and gave their consent for their responses to be used for scientific purposes. They first answered closed-ended questions on socio-demographics, study phase, knot-tying experience, learning and training methods, preferred types of training, and openness to and perceived advantages and disadvantages of MR-based training. Those with previous knot tying experience also answered open-ended questions about the most challenging and important steps in knot tying.

The open responses were evaluated according to Kuckartz's content-structuring approach (2022: 129 ff.) by first categorizing the statements and then counting the frequencies of the categories.

Results: Most students initially learned knot tying by attending a workshop where instructors give guidance and feedback on knotting (see figure 3A): 59 out of 80 students had already attended such a workshop, while the rest mostly learned from online videos (*n* = 7 mentions) or had no prior knotting experience (*n* = 9). However, many students who had attended a workshop did not continue practicing knot tying afterwards (*n* = 24, see figure 3B). Of those who did continue practicing, most did so on strings attached to objects (*n* = 21) or from re-tying knots shown in tutorial videos (*n* = 17), and only rarely on professional practice benches (*n* = 14) or on patients (*n* = 8).

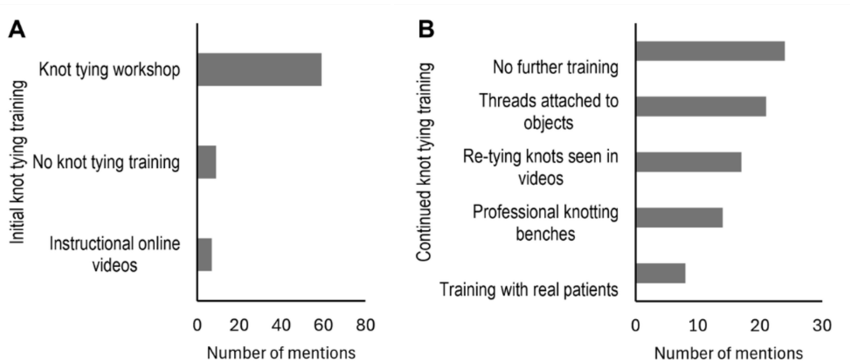

Figure 3: Reported frequency of strategies used in initial knot training (3A) and continued knot training (3B) (Source: Own figure)

Regarding prior experience with simulators in general and MR, 33 reported having interacted with non-MR medical simulators, whereas only 15 had been using MR or VR at all. Despite this lack of first-hand experience, students were quite open to learning and training medical and surgical tasks with simulators (*M* = 4.59, *SD* = 0.57, on a scale of 1 to 5, with 5 indicating high openness). They were also rather interested in learning knot tying with MR (*M* = 4.18, *SD* = 1.00), citing the possibility of realistic yet risk-free training (*n* = 16), the ability to train repeatedly and at any time (*n* = 7) and a general fascination with this new technology (*n* = 18) as main advantages. Still, they considered instructor support essential for successful learning (*M* = 4.06, *SD* = 0.89) and mostly did not agree that simulators could be used to train without human instructors (*M* = 3.10, *SD* = 1.12).

Other responses also underlined the impression that students value the personal guidance and feedback from an instructor: Learning with instructor guidance was rated as the most helpful learning method (*M* = 4.84, *SD* = 0.37), while independent practice on training benches without further aids and independent tying of knots were rated as the least helpful (*M* = 2.84, *SD* = 1.07 and *M* = 3.11, *SD* = 0.88). Students also expressed concerns that an MR-based training application could not fully substitute human instructors and their direct support (*n* = 5 mentions).

The most important and most difficult steps in knot tying, according to the students, were tight and correct thread guidance throughout the entire tying process (*n* = 11), correct hand coordination (*n* = 10), correct tightening of the partial and complete knots (*n* = 9), and correct tying of counter knots, i.e., knots tied in opposite directions to stabilize the overall knot (*n* = 8).

Conclusions: The first survey showed that an MR-based system could support knot-tying training, which is hardly done systematically, especially outside of workshops. Students are open and interested but want the system to provide comprehensive support, guidance, and individualized feedback similar to a human instructor. It should teach, among other things, correct thread handling, hand coordination, knot tightening, and the correct sequence of counter knots. Additionally, the system must be user-friendly, especially for users without MR experience.

5.1.2 Expert Study: Analyzing Surgeons' Knot Tying Techniques

The second study focused on the perspectives of practicing surgeons as knot tying experts. It aimed to identify the knot tying technique used in surgical practice, the individual steps of the knot tying process, and experts' recommendations and requirements for an MR-based knot tying trainer.

Method: Four expert interviews including think aloud were conducted with physicians from Charité – Universitätsmedizin Berlin (three male, one female; two assistant physicians, two surgeons; self-reported number of knots tied between 100 and 1,000,000) in November and December 2022.

Following Wassermann's (2015: 55 ff.) recommendations for expert interviews, a semi-structured format was used and recorded via audio. The interviews covered three topics: the experts' knot tying techniques, quality criteria for surgical knots, and their teaching experiences and recommendations. The interviewees also demonstrated knot tying using a knot bank and braided threads, thinking aloud and explaining their steps while being recorded on video. For the second knot, they wore the Varjo XR-3 XR glasses used in the project to allow them to provide feedback, for example on its weight and pass-through display quality.

Data analysis involved transcribing the interviews and applying qualitative content analysis using a mixed deductive-inductive approach, with three top-level categories (knot technique, quality criteria, teaching recommendations) and additional sub-categories. Two independent coders developed a category system from one interview, which was used to code the remaining interviews. Video recordings were analyzed to identify knot tying steps.

Results: All experts reported using Granny Knots, which is a kind of slip knot formed by first crossing the two threads, and then tying two knots. Both knots should be tied in the same direction, but the initial crossing of the threads results in the first knot being inverted. Opposing knots are necessary to achieve a neat and flatly packed knot, minimizing irritations of the tissue. It can be tightened further after tying, which prevents it from loosening easily and makes it resistant to untight air knots (see figure 4). Finally, it is secured by one or more counter knots each tied in the opposite direction to the previous one. The experts report to use up to four counter knots when using monofilament threads but usually only one when using polyfilament threads.

Three of them preferred to tie the first two knots of the Granny Knot with the middle finger and the opposite knot with the index finger, and one did it the other way round. Two experts emphasized the importance of crossing the threads at the beginning, to avoid twisting the first knot and reduce the force necessary for tightening the knots. Two experts tie knots with their dominant, two with their non-dominant hand, mostly because there is less risk of injury when using a needle.

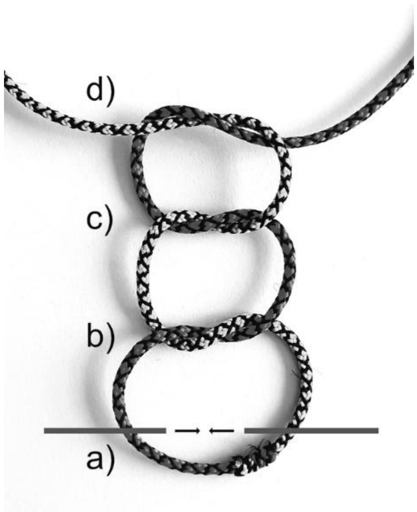

Figure 4: Granny Knot used by experts and taught by the MR training application. a) Subcutaneous thread section. Tissue layer symbolized by red lines. b) First partial knot, tied forwards with middle finger technique. c) Second partial knot, tied backwards using the middle finger technique. d) Third partial knot, tied forwards using the index finger technique. (Source: Own figure)

Regarding quality criteria, the interviewees identified two main categories: Criteria regarding the knot tying process, and criteria regarding the final knot.

During the knot tying process, maintaining thread tension is crucial for keeping the tissue in position (n = 4). Keeping the correct course of the thread is important, as a twisted run increases the force needed to tighten the knot (n = 2). Movements and tensile force should be minimized to avoid tissue damage (n = 3). Additional aspects include high tying speed (n = 1), injury-free needle handling (n = 2), and awareness of the specifics of the used thread material (n = 3).

For the final knot, the strength of the knot is paramount. It should not come undone (n = 4) and should fit tightly (n = 2). The knot should also consist of as few individual knots as possible to reduce foreign material in the body and potential inflammation (n = 1).

Conclusions: Based on these expert insights, it was decided that the MR-based knot tying trainer should teach a knot consisting of two single knots in the same direction, tied with the middle finger and secured by a counter knot with the index finger, using a polyfilament thread. A visual indicator of tensile force was developed to indicate if too much force is being used.

5.2 Implementation: Developing the MR-based Surgical Knot Tying Trainer

Based on the requirements analysis with students and experts, a concept for the digital knot tying trainer was developed and made available to the developers in the form of a design document. The studies' findings suggested the development of a prototype that combines real-world physical elements with virtual instructions and automated feedback to meet students' needs for a realistic and practice-based yet risk-free and interactive system. The derived learning concept is based on three main principles: Firstly, the learning content should be displayed three-dimensionally, as knot tying requires a particularly distinctive presentation of spatial relations. Secondly, the learning instructions should be applied to the learner's own body and movements, which is achieved by spatially mapping the instructions on the learner's hands. And thirdly, the learning content should be automatically adapted to the learner's progress and learning speed. A prototype app, based on this concept was implemented by the developers, as seen in figure 5.

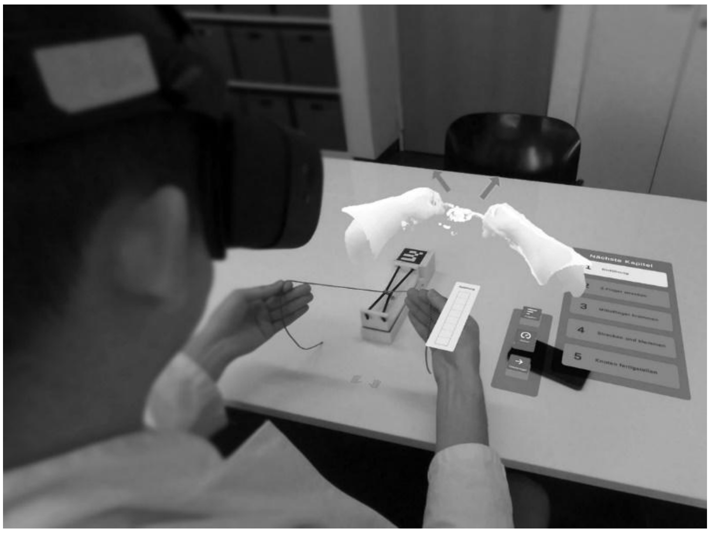

Figure 5: A student uses the knot tying trainer prototype implementation. The volumetric video-based learning instructions are visible in the center. Elements for controlling the application are seen on the right side. (Source: Own figure)

The prototype of the knot tying trainer consists of three components:

a) Knotting bench: Learners use a physical knotting bench with two strings that must be knotted together with a provided thread. The bench can digitally measure and display the thread tension and tensile strength of the knot during the exercise, allowing learners to avoid excessive pull that could damage patients' tissue.

b) VR goggles with hand tracking: We use Varjo XR3 VR goggles, a state-of-the-art VR/MR headset that offers a high resolution and good visual fidelity for the stereoscopic visualization and three-dimensional capture of the training scene. Using the goggles' pass-through function, the learner's real environment can be overlaid with virtual elements. The hand tracking technology UltraLeap accurately registers the hand and finger movements of users in real-time.

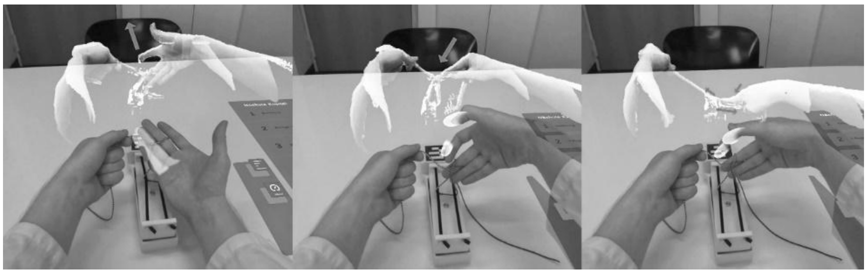

Figure 6: An exemplary sequence of the volumetric video-based learning instructions, in the order from left to right. The arrows highlight the movements which need to be performed. (Source: Own figure)

c) Learning software: The digital learning application provides step-by-step instructions in the form of virtual instructor hands that demonstrate the Granny Knot favored by experts. For this purpose, volumetric videos of experts' knot tying techniques were recorded and integrated into the system, superimposed on the learner's movements (figure 6). The steps can then be imitated by the learner using visual instructions mapped on the hands via superimposed arrows and a voice generated by text-to-speech (figure 7). The system uses hand and finger tracking to recognize individual learning progress based on key gestures and adapts the learning content accordingly. During a knot tying task, the application can overlay visual feedback over learners' fingers, for example by color-highlighting wrongly positioned fingers. The system also provides comprehensive onboarding and step-by-step guidance to assist users without prior MR experience.

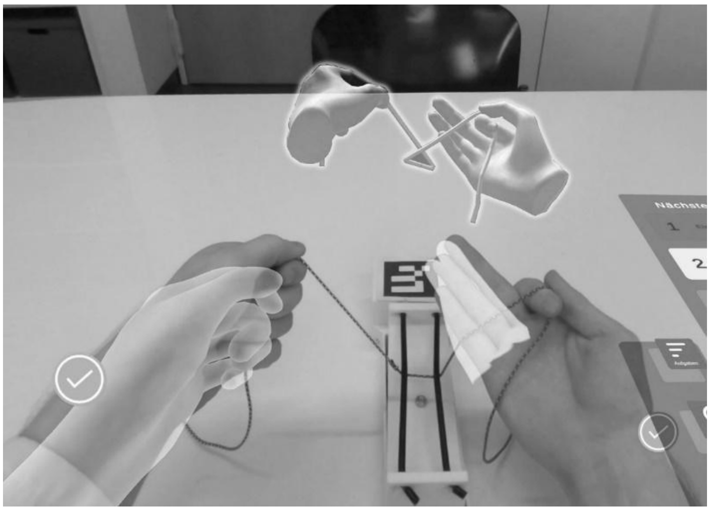

Figure 7: Superimposed instructions and information, which is projected onto the hands of the user. The green augmentation on the left indicates that the user has successfully executed a step. The blue overlay highlights fingers which need to be moved to perform the next step. (Source: Own figure)

During development, questions arose that could only be answered by user testing: It was necessary to know how much visual feedback is useful – do instructor hands, arrows, and colored highlights on the learner hands all support learning, or should individual elements be removed to avoid distraction and visual clutter? Several design variants were developed for this purpose, which included all or only individual elements and were to be evaluated by users. Additionally, we wanted to receive general feedback on the system and improvement potential from the user's perspective.

5.3 Testing: Online Evaluation of the Proof-of-Concept Prototype

Therefore, in a third, experimental study, three design variants of the MR-based learning system were evaluated by medical students at Charité – Universitätsmedizin Berlin. The goal was to identify the best design for further development and to get feedback on the system's usability, user experience, and learnability.

Method: In November 2023, fifty-five medical students participated in the online experiment, of whom 46 remained after exclusions (30 women, 16 men, M_{age} = 23.5 years, *SD* = 4.6, *range* = 19–42). Participants watched videos of the app from the viewpoint of someone wearing mixed reality glasses. A between-subject experimental design with three conditions was used. Participants were randomly assigned to one of three app variants:

- *Complete:* The app included virtual instructor hands, verbal instructions, visual feedback on tensile force, arrows indicating direction of movement, and hand highlighting (*n* = 11).
- *Without arrows:* The app included all elements except arrows (*n* = 16).
- *Basic:* The app included only basic elements, no arrows, and no highlighting (*n* = 19).

Participants first answered questions on socio-demographics and knot tying experience. Then, they watched the assigned video, split into three parts: An interactive tutorial (2 min), knot bank preparation (2 min), and the first steps of knot tying (1 min). After each part, they evaluated the functions and elements of the application using rating scales and open questions.

Dependent variables included user experience (short User Experience Questionnaire [UEQ-S]) and willingness to use (taken from the TAM), using 5- and 7-point Likert scales. Open questions addressed strengths, weaknesses, and improvements for the system.

For analysis, mean values of dependent variables were calculated and compared descriptively. Due to the exploratory nature of the analysis and the small sample size, only descriptive statistics are reported. Open-ended responses were categorized inductively.

Results: Regarding user experience, the complete design variant was overall rated best (M_{UEQS} = 2.20, SD_{UEQS} = 0.27, with values ranging between -3 and 3) and better than the variant without arrows (M_{UEQS} = 1.90, SD_{UEQS} = 0.73) and the basic variant (M_{UEQS} = 1.58, SD_{UEQS} = 0.89). A similar pattern of differences emerged for pragmatic and hedonic quality (see figure 8). When compared to the UEQ-S benchmark values, all variants are ranked at least above average, with the complete variant and the one without arrows ranked as excellent.

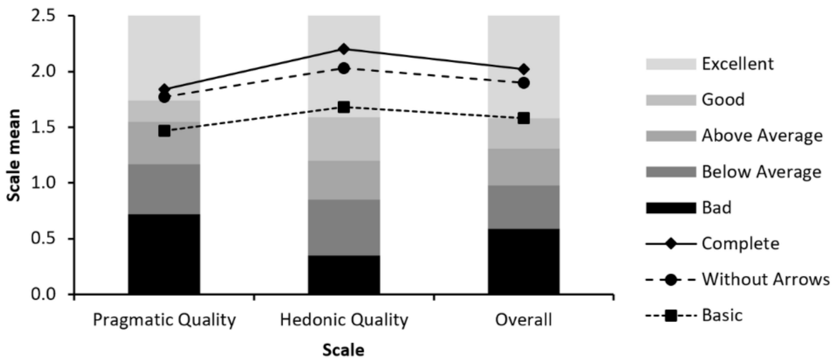

Figure 8: Benchmark of UEQ-S mean values of the three design variants (Source: Own figure)
Note. The colored bars indicate the 25th, 50th, 75th and 90th percentile of values in the benchmark sample of studies to which the values of this study are compared.

The most helpful system element, according to participants' ratings, was the visual feedback on tensile force, that is, whether sufficient or too much tension is applied to the thread (n = 19). Other positive features mentioned were the virtual instructor hands that demonstrated the knot (n = 7) and the colored highlighting of the users' hands when mistakes are made (n = 6).

The main point of criticism was the poor visual representation of the virtual instructor hands and the virtual thread. Respondents wished for higher contrast, more defined, sharper, less shaky visualizations, which also take color blindness into account (n = 22). Participants criticized the text-to-speech-based computer voice as too monotonous, slow and detrimental for concentration and wished for a more human sounding voice (n = 14). Concerning the user interface, participants wished for more self-explanatory or better labelled buttons (n = 7), a menu to select the knot to be trained (n = 3) and overall, for a more visually appealing and modern design (n = 3).

Participants reported the greatest intention to use for the variant without arrows ($M_{intention}$ = 4.62, $SD_{intention}$ = 0.62, with values ranging between 1 and 5) and the complete variant ($M_{intention}$ = 4.36, $SD_{intention}$ = 0.50), and a much lower for the basic variant ($M_{intention}$ = 3.53, $SD_{intention}$ = 1.39).

Conclusions: The system variant with all elements – instructor hands, arrows and highlighting of user hands – was rated best and was therefore developed further in the project. To improve the display quality of the hands, new volumetric videos with additional cameras were recorded. Additionally, high-resolution still images of 3D modelled hands and threads were added to the instruction videos, illustrating the final state of a knot step in greater detail than still images from volumetric video.

6. Discussion and Conclusion

This chapter explored the human-centered design of a mixed reality (MR) application for surgical knot tying, integrating different design models and methods, and reporting on three studies conducted in the Tangible Reality project.

A generalized iterative process model for the design, development, and evaluation of medical and surgical training applications was proposed, which integrates the ISO 9241-210, the double diamond model, and the Kirkpatrick evaluation model. This model emphasizes understanding user contexts, defining requirements, iterative prototyping, and evaluating both short- and long-term effects of the system.

Based on this model, an overview of methods for integrating users and stakeholders in the design and development process was given, highlighting four methods that appear to be particularly suitable for development in medical training: Stakeholder analysis, think aloud, and questionnaires measuring technology acceptance and user experience.

Three studies illustrated the use of these methods in the context of the development of an MR-based surgical knot tying trainer: Inspired by an initial stakeholder analysis, study one measured medical students’ requirements and perspectives on MR-based knot tying training, while study two assessed experienced physicians’ knot tying strategies and common quality criteria. These were integrated in different design variants of the training application. In study three, first-person video mock-ups of three design variants were rated by medical students in terms of user experience and acceptance to identify the best variant and additional requirement for further design.

In conclusion, the findings of the studies and their interplay illustrate that a human-centered, iterative design approach is essential for developing effective MR-based medical training tools. By taking into account stakeholder perspectives and integrating user feedback and expert insights throughout the design process, the development of training solutions that enhance the skills and confidence of medical professionals can be ensured.

Funding information:

The Project GreifbAR is funded by the German Federal Ministry of Education and Research (BMBF) within the funding line „Interactive systems in virtual and real spaces – Innovative technologies for the digital society“.

Literature

Bryson, John M. (2004). What to do when Stakeholders matter: Stakeholder Identification and Analysis Techniques. In: *Public Management Review,* vol. 6(1), p. 21–53. https://doi.org/10.1080/14719030410001675722

Charters, Elizabeth (2003). The Use of Think-aloud Methods in Qualitative Research. An Introduction to Think-aloud Methods. In: *Brock Education Journal,* vol. 12(2), p. 68-82. https://doi.org/10.26522/brocked.v12i2.38

Davis, Fred D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. In: *MIS Quarterly,* vol. 13(3), p. 319–340. https://doi.org/10.2307/249008

DeVellis, Robert F. & Thorpe, Carolyn T. (2021). *Scale Development: Theory and Applications,* SAGE Publications, Thousand Oaks.

Freudenthal, Adinda et al. (2011). Collaborative co-design of emerging multi-technologies for surgery. In: *Journal of Biomedical Informatics,* vol. 44(2), p. 198–215. https://doi.org/10.1016/j.jbi.2010.11.006

Granić, A. (2022). Educational Technology Adoption: A systematic review. In: *Education and Information Technologies,* vol. 27(7), p. 9725–9744. https://doi.org/10.1007/s10639-022-10951-7

Graser, Stefan & Böhm, Stephan (2022). A Systematic Literature Review on Technology Acceptance Research on Augmented Reality in the Field of Training and Education. In: *CENTRIC 2022: The Fifteenth International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies, and Services, 2022*, vol. 12, IARIA, p. 20–28. https://doi.org/10.48550/arXiv.2411.13946

Hagelsteen, Kristine et al. (2019). Performance and perception of haptic feedback in a laparoscopic 3D virtual reality simulator. In: *Minimally Invasive Therapy & Allied Technologies,* vol. 28(5), p. 309–316. https://doi.org/10.1080/13645706.2018.1539012

International Organization for Standardization (2018). *Ergonomics of human-system interaction—Part 11: Usability: Definitions and concepts. (ISO Standard No. 9241-11:2018).* https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en [Zugriff 20.03.2025]

International Organization for Standardization (2019). *Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems (ISO Standard No. 9241-210:2019).* https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en [Zugriff 20.03.2025]

Kirkpatrick, Donald (1996). Great ideas revisited. Techniques for Evaluationg Training Prgrams. Revisiting Kirkpatrick's Four-Level-Model. In: *Training & Development,* vol. 50(1), p. 54–57.

Kuckartz, Udo (2022). Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung, Beltz Juventa, Weinheim, Basel.

Laugwitz, Bettina, Held, Theo & Schrepp, Martin (2008). Construction and Evaluation of a User Experience Questionnaire. In: *HCI and Usability for Education and Work*, Springer, Berlin, Heidelberg, p. 63–76. https://doi.org/10.1007/978-3-540-89350-9_6

Lindberg, Tillmann, Meinel, Christoph & Wagner, Ralf (2011). Design Thinking: A Fruitful Concept for IT Development? In: *Design Thinking,* Springer, Berlin, Heidelberg, p. 3-18. https://doi.org/10.1007/978-3-642-13757-0_1

Mlekus, Lisa et al. (2020). How to raise technology acceptance: User experience characteristics as technology-inherent determinants. In: *Gruppe. Interaktion. Organisation. Zeitschrift Für Angewandte Organisationspsychologie (GIO),* vol. 51(3*)*, p. 273–283. https://doi.org/10.1007/s11612-020-00529-7

Norman, Donald Arthur (1986). Cognitive Engineering. In: *User Centered System Design,* CRC Press, Boca Raton, p. 31–62.

Praslova, Ludmila (2010). Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in Higher Education. In: *Educational Assessment, Evaluation and Accountability,* vol. 22(3), p. 215–225. https://doi.org/10.1007/s11092-010-9098-7

Reyes, Jonatan et al. (2022). User-Centered Design for Surgical Innovations: A Ventriculostomy Case Study. In: Ethical and Philosophical Issues in Medical Imaging, Multimodal Learning and Fusion Across Scales for Clinical Decision Support, and Topological Data Analysis for Biomedical Imaging, Springer, Cham, p. 51–62. https://doi.org/10.1007/978-3-031-23223-7_5

Schrepp, Martin, Hinderks, Andreas & Thomaschewski, Jörg (2017a). Construction of a Benchmark for the User Experience Questionnaire (UEQ). In: *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 4(4), p. 40–44. https://doi.org/10.9781/ijimai.2017.445

Schrepp, Martin, Hinderks, Andreas & Thomaschewski, Jörg (2017b). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). In: *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4(4), p. 103–108. https://doi.org/10.9781/ijimai.2017.09.001

Smidt, Andy et al. (2009). The Kirkpatrick model: A useful tool for evaluating training outcomes. In: *Journal of Intellectual & Developmental Disability,* vol. 34(3), p. 266–274. https://doi.org/10.1080/13668250903093125

Stanton, Neville A. et al. (2013). *Human Factors Methods: A Practical Guide for Engineering and Design*. Ashgate Publishing Ltd., Farnham.

Stickdorn, Marc et al. (2018). This Is Service Design Doing: Applying Service Design Thinking in the Real World, O'Reilly Media Inc., Sebastopol.

Tidwell, Jennifer, Brewer, Charles & Valencia-Brooks, Anne (2020). *Designing Interfaces: Patterns for Effective Interaction Design (3rd Edition)*, O'Reilly Media, Sebastopol.

Venkatesh, Viswanath & Bala, Hillol (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. In: *Decision Sciences,* vol. 39(2), p. 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x