POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Usability assessment a mixed reality platform for pre-operative planning in cardiac interventions.

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria Biomedica

Author: **Clelia Galluzzi**

Student ID: 970820
Advisor: Prof. Emiliano Votta
Co-advisors: Ing. Omar Pappalardo, PhD
Academic Year: 2021-22

Abstract

Usability refers to the capability of a product to be correctly used by specific users in a specific context of use, and to the extent of effectiveness, efficiency, and satisfaction with which they can achieve specified goals through the product. Referring to medical devices, the standards IEC 62366-1 and IEC 62366-2 describe the *usability engineering process*, i.e., the process the manufacturer should follow to design and develop a usable device.

In this comparative study, the usability of a mixed reality (MR) platform for pre-operative planning of cardiac interventions was investigated and compared with the usability of a DICOM viewer software commonly used for the same purpose. MR refers to the technologies that allow the user to visualize virtual 3D objects superimposed on the real world, and to interact with them while maintaining the physical connection with the surrounding environment. The 3D rendering is realized through a powerful workstation which processes digital images and the generated content is visible to the user by wearing a semi-transparent head-mounted display. Given the novelty of the technology, usability evaluation is pivotal to test if it could be positively accepted by potential users. Moreover, the new Medical Device Regulation (MDR) 2017/745 expects the mitigation of risks derived from human errors, thus implicitly requiring usability evaluation.

To this goal, 16 physicians with no experience in MR technologies were enrolled and clustered into two groups: experts (n=7) and newbies (n=9) in using traditional DICOM viewer software. Participants were asked to perform 3 specific tasks both with the MR platform and with a traditional DICOM viewer, and the time required to perform the tasks with the two technologies was measured. Subsequently, for both technologies, they filled in three validated questionnaires, namely User Experience Questionnaire (UEQ), System Usability Scale (SUS), and Surgery Task Load Index (S-TLX), and an ad hoc questionnaire related to the MR, which was designed in the present study. Eventually, users were asked to rate the relative relevance of different features of the MR platform, with the aim of identifying the most important features that could boost MR added value and which could be exploited as evaluation criteria for future assessment of similar technologies.

Users who were newbies to the use of both technologies completed the tasks faster with the MR platform and preferred it in terms of perceived workload, usability, and general user experience. Opposite results were obtained from users experienced in the use of traditional DICOM viewers, even if they assigned high usability scores and low workload scores also to the MR platform. In general, all participants perceived the added value brought by the MR technology in the pre-operatory phase. Deeper studies will be however needed to derive more reliable conclusions.

Keywords: usability, IEC 62366, mixed reality, pre-operative planning.

Abstract in lingua italiana

Il termine *usabilità* si riferisce alla capacità di un prodotto di essere usato correttamente da specifici utenti in uno specifico contesto di utilizzo, e al livello di efficacia, efficienza, e soddisfazione con cui questi possono raggiungere determinati obiettivi sfruttando il prodotto. Riferendosi ai dispositivi medici, gli standard IEC 62366-1 and IEC 62366-2 descrivono il processo di *usability engineering*, cioè quel processo che il fabbricante deve seguire per sviluppare un dispositivo *usabile*.

In questo studio comparativo è stata analizzata l'usabilità di una piattaforma di realtà mista (MR) da usare nella pianificazione pre-operatoria in ambito cardiaco e comparata a quella di un software per la visualizzazione di immagini DICOM comunemente utilizzato per lo stesso scopo. La MR consiste in tecnologie che permettono di visualizzare oggetti virtuali 3D sovrapposti al mondo reale e di interagire con questi mantenendo la connessione fisica con l'ambiente circostante. Il rendering 3D è realizzato attraverso una potente workstation che elabora immagini digitali ed il contenuto generato è visibile all'utente indossando un visore con lenti semitrasparenti. Data la novità della tecnologia, la valutazione dell'usabilità è fondamentale per capire se questa sarà accettata positivamente dai potenziali utilizzatori. Inoltre, il nuovo Regolamento per i Dispositivi Medici (MDR) 2017/745 prevede la mitigazione dei rischi derivanti da errori umani, richiedendo quindi implicitamente la valutazione dell'usabilità.

A questo scopo, 16 medici senza esperienza con tecnologie di MR sono stati arruolati e suddivisi in due gruppi: esperti (n=7) e neofiti (n=9) nell'uso di tradizionali software per la visualizzazione di immagini DICOM. Ai partecipanti è stato chiesto di svolgere 3 task specifici sia con la piattaforma di MR che con un visualizzatore DICOM tradizionale, ed è stato misurato il tempo necessario per eseguire i compiti con le due tecnologie. Successivamente, con riferimento ad entrambe le tecnologie, i partecipanti hanno compilato tre questionari validati (User Experience Questionnaire (UEQ), System Usability Scale (SUS), Surgery Task Load Index (S-TLX)), e un questionario ad hoc relativo alla MR, che è stato sviluppato direttamente in questo studio. Alla fine, agli utenti è stato chiesto di valutare la rilevanza relativa di diverse caratteristiche della piattaforma di MR, con l'obiettivo di identificare le funzionalità più importanti che potrebbero aumentare il valore aggiunto della tecnologia ed essere sfruttate come criteri su cui basare una sua futura valutazione.

Gli utenti inesperti nell'uso di entrambe le tecnologie hanno completato i task più velocemente con la piattaforma di MR e hanno preferito questa in termini di carico di lavoro percepito, usabilità ed esperienza utente in generale. Risultati opposti sono stati ottenuti per gli utenti esperti nell'uso di visualizzatori DICOM tradizionali, anche se questi ultimi hanno attribuito alti punteggi di usabilità e bassi punteggi di carico di lavoro anche alla piattaforma di MR. In generale, tutti i partecipanti hanno percepito il valore aggiunto portato dall'uso della tecnologia di MR nella fase di pianificazione pre-operatoria. Saranno tuttavia necessari studi più approfonditi per trarre conclusioni più attendibili.

Parole chiave: usabilità, IEC 62366, realtà mista, pianificazione pre-operatoria.

Contents

Introduction

Extended Reality (XR) encompasses those environments that allow for human-machine interaction through some combination of the physical world with virtual objects. It consists of a spectrum of technologies referred to as “reality-virtuality continuum”, which allows for generating virtual 3D interfaces, visualizing them through wearable devices, and interacting with them in different ways. XR spectrum ranges from the totally immersive Virtual Reality (VR), where the user finds himself in a fully digital world where he/she can move around and interact, yet losing the connection with the real environment, to Augmented Reality (AR), where the user visualizes his physical environment augmented by virtual objects (annotations, plots, numerical data, . . .) without having the possibility to interact with them, passing through Mixed Reality (MR), where the user, even maintaining the connection and visualization of the real world, can see digital objects, manipulate them and interact with them.

On the top of a workstation dedicated to data processing, the use of these technologies requires head-mounted displays (HMDs), i.e., headsets equipped with ad hoc technologies including screens, sensors (e.g., infrared cameras and gyroscopes), and a central processing unit (CPU).

The recent improvement of HMDs, together with the increased computational power and the improved software capabilities, has brought an increasing number of applications of this technology in various sectors, medicine and health care being key ones. The numerous medical applications share one unmet need that can be fulfilled by XR: the need for a user-friendly and ergonomic interface to visualize patient-specific medical imaging and 3D anatomical reconstructions, as well as to easily navigate them and to fully capture 3D anatomical complexities. Current interfaces rely on the use of physical 2D monitors, which do not allow for a fully 3D rendering and, because of their physical encumbrance, cannot be positioned in the operatory room (OR) in a way that optimizes the ergonomics and comfort of the operators. Moreover, the classical interaction paradigm through mouse, keyboard, or touchscreen monitor can pose problems in the context of ORs or other hospital environments: the operator has to move his hands away from the patient or from the surgical field to interact with the device and the sterile conditions required in some hospital settings can be threatened. XR may overcome these limitations and allow for a broader range of interactions with the represented data through very user-friendly modalities. Thus, medical applications of XR can range from pre-operative planning, intra-operative guidance, doctor training, and patient management.

In the particular context of surgery, the most useful technology of the XR spectrum is MR since it allows the operator to remain conscious of the events happening in the OR and to interact with the rest of the staff while having additional information in the form of holograms that can be easily manipulated, zoomed, explored inside and moved to the most comfortable position.

The fast technological development, in terms of hardware and software, is allowing for the development of increasingly detailed anatomical models, the visualization of progressively neater and better resolved images, and the improvement of the sensors' capability to respond to human commands. However, beyond these technical aspects, other relevant considerations concern the effectiveness of human-machine interactions and the usability and learnability of the system, which are fundamental prerogatives for the technology to become consumer items in healthcare facilities.

Focusing on the use of MR in interventional-structural cardiology and in cardiac surgery, this work aims at analyzing the usability of a MR platform and at comparing it with a traditional DICOM viewer software, the gold standard technology used in pre-operative planning.

Accordingly, the study of the usability of the system was implemented as a summative evaluation, to validate the design of an already existing MR-based solution. To this aim, a series of tasks to be performed by clinical end-users through the MR-based system was defined; these tasks were all related to the analysis and navigation of medical imaging and of 3D anatomical reconstructions. End-users with different levels of medical expertise were enrolled and administrated with validated questionnaires as well as with ad hoc developed questionnaires upon completing the tasks. From the analysis of their answers, general information concerning their perception of the usability of the system was derived together with more specific information relating to the consistency or variability of the evaluations among the different users. Also, the completion time was measured to evaluate whether the performance of end-users depended on their type and degree of medical specialization, and on the level of experience with XR technologies.

Moreover, the enrolled end-users were asked to fulfill the same tasks through a standard DICOM viewer software package, to perform a comparative evaluation vs. a gold standard in terms of general experience, usability, task complexity, time required, and workload perceived.

Eventually, the end-users were asked to grade the relative importance of different features of the MR-based technology before and after having used it, in order to set priorities among the different functionalities to be guaranteed to boost its efficiency and added value. This research question was aimed at three specific goals:

1. allowing the manufacturer to answer efficiently and exhaustively to the user needs, by adopting a developing approach that takes into consideration user's needs, perceptions, and opinions by design;
2. allowing those in charge of technology assessment, namely clinical engineers, to engineer the process of evaluation of subjective aspects like ergonomics and usability, which are commonly left to the physician's discretion;
3. evaluating whether end-users' initial opinion concerning the relevance of the functionalities of the technology has changed after its use.

The logical workflow followed in this work, and the organization of the sections, is clarified in the following lines:

1. **Background.** It contains information concerning the technology covered by this study, i.e., MR platforms, together with notions concerning the research problem, i.e., usability evaluation.
   At first, a general definition of extended reality and related technologies is provided together with some examples of their medical applications derived from the literature. Then, the focus is shifted to mixed reality with the presentation of the principle of stereoscopy and of the hardware required to deliver the holographic content. In the end, the exact MR platform under evaluation in the course of this study is presented.
   For what concerns the research problem, the definition of usability and usability engineering process is provided. Subsequently, the reference regulatory context which regulates usability evaluation is presented together with some validated tools to implement it.
2. **State of the art.** The section contains examples of usability studies concerning XR technologies derived from the literature, which describe some inspiring works performed in the context of interest and the application of tools and methodologies similar to the ones exploited in this study.
3. **Materials and methods.** It describes in detail the implementation of the usability study object of this work: the study design, participants' enrollment, the definition of the test protocol, and the evaluation methods. Also, the statistical tests and analyses performed on the collected data are presented.

4. **Results and discussion.** In this section, the results from the testing phase and from the data processing are reported and discussed in order to derive relevant implications, conclusions and insights concerning the research questions object of this study.

5. **Conclusions and limitations.** A summary of the results obtained from the conduct of this study is reported in this section. Also, as the title suggests, the weaknesses and limitations of the current study are herein presented together with some hints about possible developments and improvements applicable in future studies.

1 | Background

1.1. Extended Reality

Extended reality (XR) is an advanced technology that allows for a human-machine interface in an environment that combines physical and virtual objects; the latter are visualized through wearable devices and, in some cases, the user can interact with them. XR encompasses different sub-categories of technologies that constitute the "virtuality-reality continuum" [16], which can be imagined as a line having the real and the virtual world at the extremes and the technologies of virtual (VR), mixed (MR), and augmented reality (AR) in between, as shown in Figure 1.1

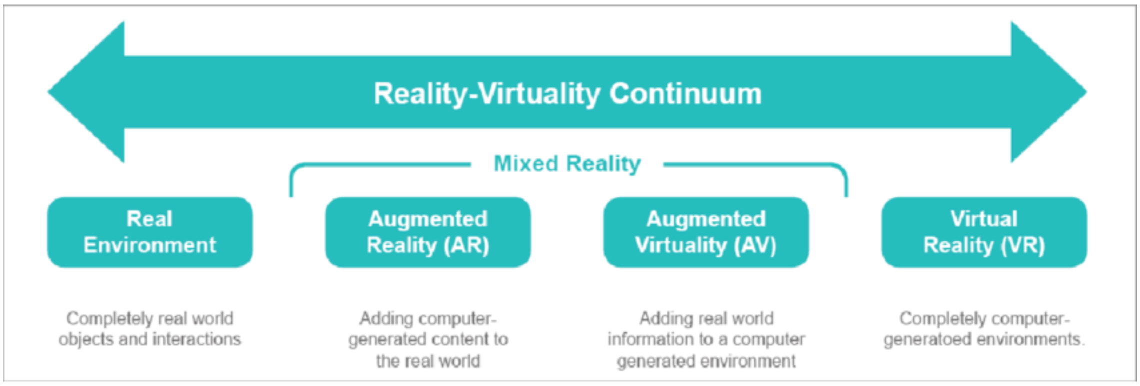

Reality-Virtuality Continuum
Mixed Reality
Real Environment
Augmented Reality (AR)
Augmented Virtuality (AV)
Virtual Reality (VR)
Completely real world objects and interactions
Adding computer-generated content to the real world
Adding real world information to a computer generated environment
Completely computer-generatoed environments.

Figure 1.1: *Virtuality-reality continuum: the set of technologies that allow for reality extension.*

Actually, the schematization by Drascic and Milgram [16] has now been overcome since the meaning of mixed reality has been modified and augmented virtuality has been incorporated into it. Thus, the most recent meaning of the different technologies that form the XR spectrum is presented in the following lines and schematized in Figure 1.2.

Virtual reality (VR) refers to fully immersive digital experiences, in which the user, wearing an immersive visor, founds himself in a completely digital world that substitutes the real one. Because of the occlusive characteristic of the headset, the user is no more able to visualize the real world and interact with the people around him. For what concerns

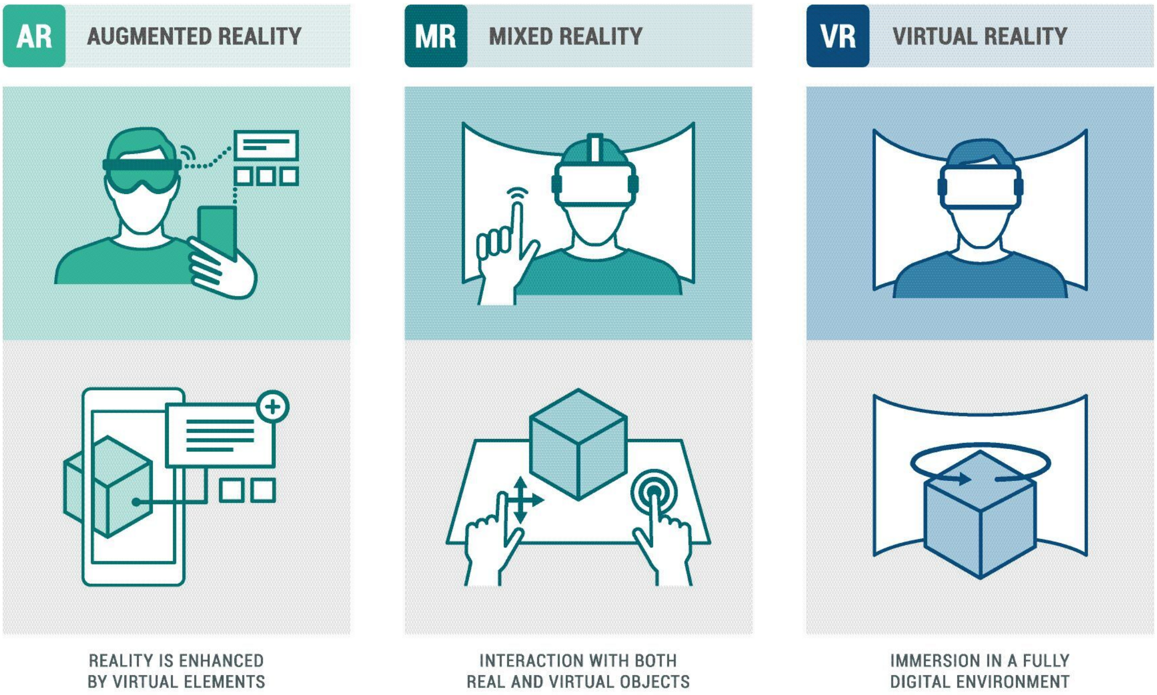

AR AUGMENTED REALITY
MR MIXED REALITY
VR VIRTUAL REALITY
REALITY IS ENHANCED BY VIRTUAL ELEMENTS
INTERACTION WITH BOTH REAL AND VIRTUAL OBJECTS
IMMERSION IN A FULLY DIGITAL ENVIRONMENT

Figure 1.2: *Most recent schematization of the three technologies forming the XR spectrum.*

the hardware, the visor is completed with hand controllers such as motion capture data gloves, that allow the user to interact with the virtual objects and the digital world.

Augmented reality (AR) is the technology that remains most faithful and similar to reality: it consists of an “augmentation” of the real world through digital 2D or 3D objects superimposed on it. In most cases, they take the form of written annotations, numerical data, plots, and lists of instructions, aimed at helping the user to better understand the world and the objects around him. The most important feature of AR is that the user can only visualize this information without having the possibility of interacting with it. The dedicated hardware can range from simple smartphones or tablets to more complex smart glasses and headsets. In the first case, the user has an indirect vision of the augmented world on the screen of the device while in the other cases he can directly look at the environment around him and visualize the added information.

In between VR and AR there is the technology of mixed reality (MR): in this case, the user remains connected to the real world, he can interact with it and with the people around him while having the possibility to visualize and interact with computer-generated contents, for example by moving, rotating, zooming and exploring them. The headset is in most of cases semi-transparent, to always allow the user to have a direct vision of reality, and it is provided with microphones, cameras, and light sensors to scan the surrounding environment and respond to the user's gestures or voice. Since augmented virtuality is now included in MR, also immersive devices can be used to deliver MR technologies. In this case, they need to be equipped with cameras to monitor the external world and to reproduce the real objects as virtual ones so that the user can freely navigate the fictitious environment without the risk of hitting them. In this discussion, with the term MR I will refer only to non-immersive technologies that allow the direct visualisation of the real world, leaving aside the exception of augmented virtuality.

1.1.1. Medical applications of XR

The improvement of both hardware and software performance has led to a rapid spread of XR in different application fields, such as industry and health care. Focusing on the last one, the technology can be exploited to support activities such as pre-operative planning, intra-operative guidance, doctor training, and patient management. In this work, the focus has been on the medical applications of XR in the context of interventional cardiology and cardiac surgery, but the many considerations are generalizable to other types of surgery and medical disciplines. In the following paragraphs, a more precise description of the different applications of XR in medicine will be presented, together with some practical examples of their implementation in interventional cardiology and cardiac surgery derived from a literature review.

In the pre-operative phase, the possibility of visualizing a more accurate and realistic reproduction of the patient's anatomy and the direct depth perception is fundamental when planning surgical procedures or choosing a device to be implanted, particularly in case of complex and abnormal anatomical conformations. XR allows for obtaining 3D holographic models of the patient's anatomy starting from data acquired with traditional imaging techniques (CT, MRI, echography). The holographic model can be used to plan surgical interventions, simulate device implantation, and identify the relationships between different anatomical structures and devices. This can offer significant advantages in terms of planning accuracy and time.

As reported in a study by Ender et al. [17] for example, during heart valve repair the measurement of the size of valvular annulus under direct vision is often challenging and it is not possible in percutaneous interventions. The research group investigated the feasibility and reliability of reality-enhanced 3D echocardiographic ring sizing by exploiting modified computerized ring models that could be superimposed on 3D reconstructions of the mitral annulus. 50 patients undergoing minimally invasive mitral valve repair were involved in this study: a 3D reconstruction of their mitral valve was performed preoperatively through transoesophageal echocardiography (TEE) and 3D reconstruction software. CAD models of different Carpentier-Edwards Physio rings (Edwards Lifescience, Irvine, CA) were created and superimposed on the 3D reconstruction of the valve to define the size of the ring to be implanted (Figure 1.3). The same evaluation was performed with conventional surgical sizing (i.e., through coronary angiography, transthoracic and transesophageal echocardiography, and cardiac computed tomography). Good correlation was observed between preoperative virtual annular sizing and the size of the actual implanted annuloplasty ring (r = 0.83). The study demonstrated that the superimposition of virtual models to 3D images of the mitral valve acquired through TEE allows the measurement of the size of the ring to be substituted in an AR environment. Moreover, this measure correlates well with conventional surgical sizing and may facilitate future percutaneous mitral valve repair techniques.

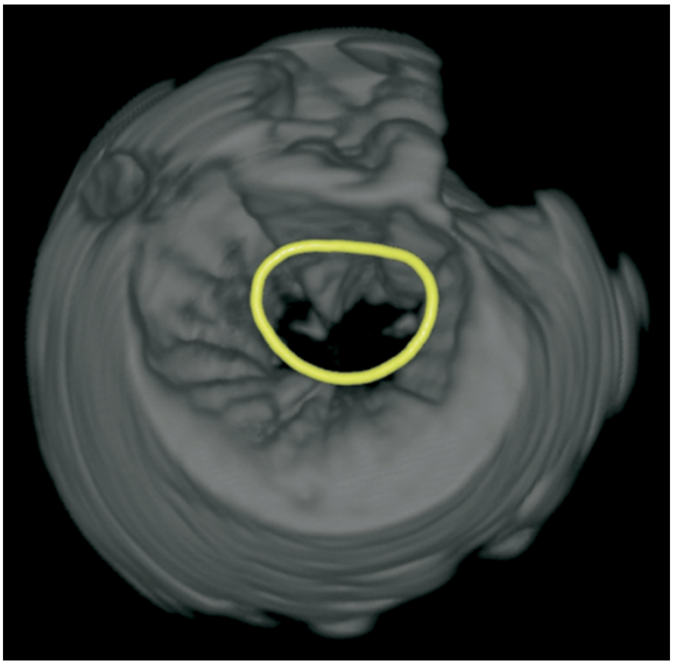


Figure 1.3: *Virtual model of the Carpentier-Edwards Physio rings (Edwards Lifescience, Irvine, CA) superimposed to the 3D reconstruction of the mitral valve obtained through TEE.*

Chan et al. [14] investigated the advantages of the commercial True 3D system developed by EchoPixel. This software platform allows for the visualization of 3D holographic objects through a 3D display and polarized glasses. In this study, the system was exploited to visualize the pulmonary arteries in newborn patients affected by pulmonary atresia [1] with major aortopulmonary collateral arteries, a congenital cardiovascular anomaly. To map the native vessels and plan the surgical correction of the anatomy, catheter angiography is currently performed and combined with 3D information from CT angiography. The study compared the accuracy and time required to interpret the localization of the collateral arteries through traditional tomographic readout and through True 3D. The interpretation time of the image visualized in 3D resulted in being significantly lower as compared to the time required to interpret images on traditional 2D screens, maintaining comparable diagnostic accuracy. In fact, the sensitivity, specificity, and accuracy of tomographic readout were 81%, 93%, and 91% respectively. For True 3D, they resulted 90%, 91%, and 91% respectively. The average time for interpretation was significantly shorter with True 3D (13 +/- 4 min) than with tomographic readout (22 +/- 7 min) ($p = 0.0004$), probably because of the enhanced visual cognition. The results of the study confirmed that advanced digital stereoscopy is recommended for the evaluation of congenital anomalies of the pulmonary vasculature.

A similar study concerning the True 3D system was conducted by Lu et al. [22] to evaluate the impact of stereoscopic visualization on preoperative planning for congenital heart surgery. In this case, preoperative planning for patients with congenital heart disease was performed by 4 cardiac surgeons with 3 different software platforms, exploiting respectively echocardiography, CT/MRI, and VR. At first, 2D and 3D images from echocardiograms and CT/MR images were reviewed on a 2D screen, and surgeons were asked to fill in questionnaires concerning this standard procedure. In a second moment, the same images were reviewed on the stereoscopic 3D platform, and again a satisfaction questionnaire was completed. In all cases, surgeons reported adequate information to operate before adding stereoscopic data but, in most cases, the preferred choice was the 3D VR platform. Stereoscopic data was reported to provide additional information that sometimes resulted in a modification of the surgical plan. In particular, even if the time spent to review data on True 3D was longer (8 minutes vs 3 minutes, $p < 0,0001$), in 96% of cases surgeons reported that this review was very useful, and in 84%, that it improved the understanding of the anatomy. In two cases (8%), the review of cardiac MRI on True 3D altered the surgical plan.

Similar results have been obtained in other studies, such as the one conducted by Haw et al. [20], in which the use of 3D VR models during the pre-operatory planning of congenital cardiac surgery has been demonstrated to result in additional information that

[1] Pulmonary atresia is a congenital malformation of the pulmonary valve in which the valve orifice fails to develop. The valve is completely closed thereby obstructing blood outflow from the right ventricle to the pulmonary artery.

might change the decision-making process, sometimes leading to a revision of the surgical plan and to a better outcome.

In the end, Ong et al. [26] investigated the use of 3D VR models to plan surgery for congenital heart diseases (CHDs) in paediatric patients. Cardiac CT images from two paediatric patients with CHD were segmented and used to reconstruct the 3D virtual model shown in figure 1.4. The study highlighted the advantages of 3D VR visualization compared to 2D visualization of 3D heart models obtained from echocardiography, CT, and MRI. The surgeon's understanding was increased by the improved realism and depth perception and by the possibility of dynamic inspection of the model, both by moving around it or by moving the model itself. The possibility of manipulating it in real-time and sharing its visualization also facilitated communications between the operators. In addition, the study compared 3D printing with VR, highlighting some of the advantages of the latter. Among these, the fact that the resolution of the VR model does not depend on the capabilities of a printer, the possibility to zoom over a specific structure, the absence of the need for purchasing printing materials, the possibility to fuse VR models with intraoperative images and to store the virtual model with the other patient's medical records in a digital form. Limitations of both approaches are the dependence on the quality and resolution of the initial imaging and the need for advanced 3D segmentation tools and in particular, for software that allows automatic and accurate segmentation without human effort, which is not available at the moment.

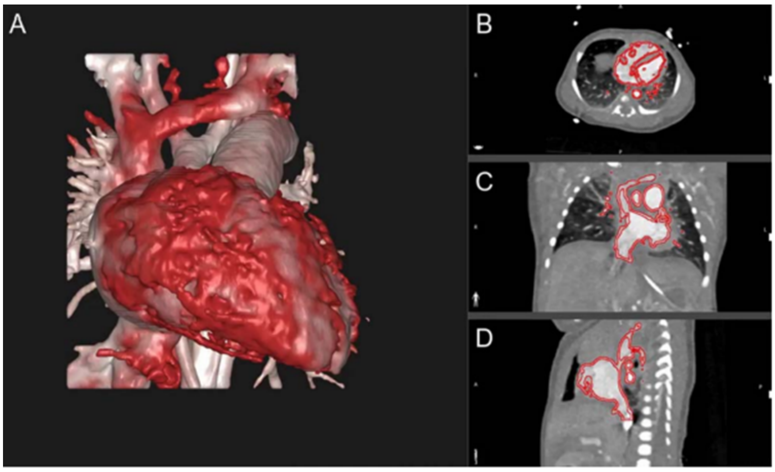


Figure 1.4: *3D segmentation of cardiac CT images. (A) 3D heart model postsegmentation; (B) Axial view; (C) Coronal view; (D) Sagittal view.*

Another relevant application of XR, namely MR, consists of intra-procedural support (Figure 1.5).

During traditional (i.e., open) surgery, MR technology allows for superimposing virtual models onto the surgical field and interacting with them while maintaining the sterile conditions of the OR. These models are typically 3D reconstructions of the patient's anatomy, which allow the surgeon for visualizing the whole anatomical structure and not only the part that is physically visible.

In mini-invasive surgery, MR has great potential in effectively guiding transcatheter interventions such as heart valve implantations, heart valve repair, atrial septal defect closure, and radiofrequency ablation procedures. The augmented vision of the patient's anatomy is very helpful for supporting the operator and reducing the procedure time, and it may also partially compensate for the lack of haptic feedback during the procedure.

Moreover, both in open and mini-invasive surgery, virtual contents can also consist of 2D screens that improve the ergonomics of the surgeon, who is provided with the vision of the relevant data in the most comfortable position, without the need for looking away from the surgical field and to assume uncomfortable positions to look at physical monitors.

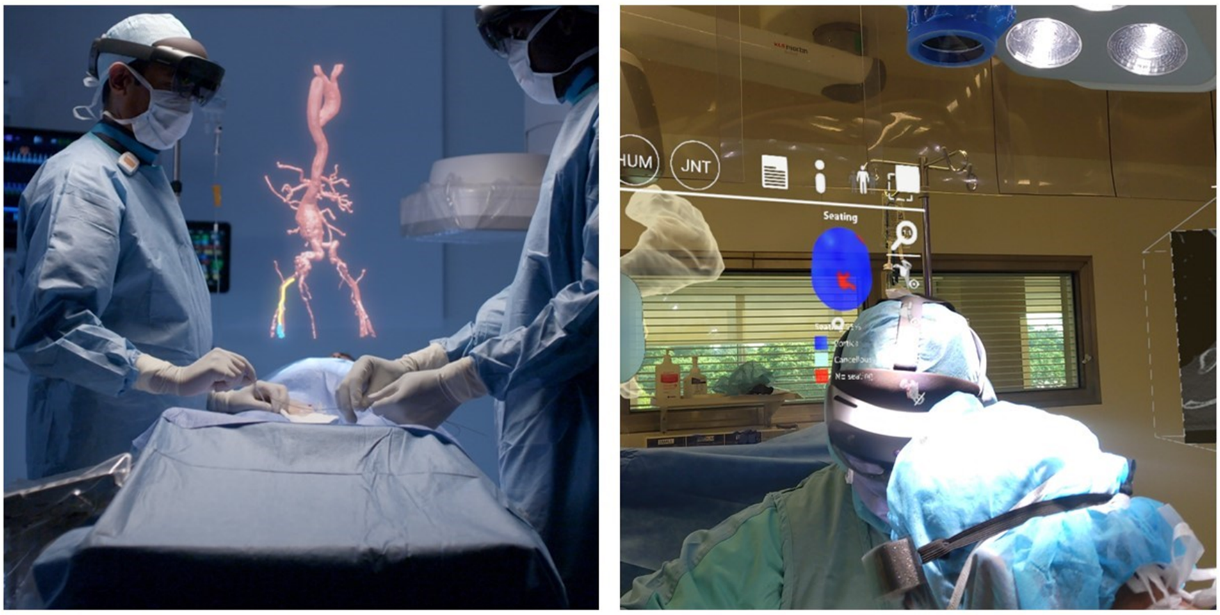


Figure 1.5: *Applications of MR in OR: holografic model of patient's anatomy (left) and holografic monitor (right).*

A significant example in the context of mini-invasive surgery is ELVIS (Enhanced Electrophysiology Visualization and Interaction System). It consists of a MR platform, developed with Windows Mixed Reality platform and loaded on Microsoft HoloLens, that provides electrophysiologists with the possibility of visualizing electroanatomic maps in

3D through a head-mounted display. Typically, the procedure of ablation in patients with heart rhythm abnormalities is performed through catheters introduced percutaneously, that are equipped with electrodes on the tip to record the electrical activity of the heart. The procedure is guided by commercially available mapping systems which exploit magnetic fields to reconstruct the anatomical maps of the heart chambers. Anatomical and electrophysiological data are overlapped to form electroanatomic maps. These allow physicians to precisely identify the zones to be treated with radiofrequency ablation and to visualize the catheters in real time. Electroanatomic maps are typically integrated with intra-procedural fluoroscopy and with pre-procedural CT images. During the ablation procedure, ELVIS allows for the visualization of data exported from electroanatomic mapping systems and pre-procedural CT/MRI. In a single hologram, it displays real time cardiac geometry, local activation time map, catheter localization, and lesion data (Figure 1.6). Interventional cardiologists are therefore provided with the possibility of visualizing patient-specific 3D cardiac models with real-time catheter location, interacting with the display while maintaining the sterility of the environment, and sharing the holographic model with other users. [33]

The same system has been analyzed by Southworth et al. [35]. Image quality, hardware, and software performance analysis resulted in the confirmation of acceptable frame rate, latency, battery runtime, dynamic range, and depth distortion. A subsequent clinical feasibility user validation was performed on 10 patients to evaluate the qualitative accuracy of the maps created with a traditional mapping system and with ELVIS. This study demonstrated that the accuracy and performance of ELVIS allow its use during clinical ablation procedures.

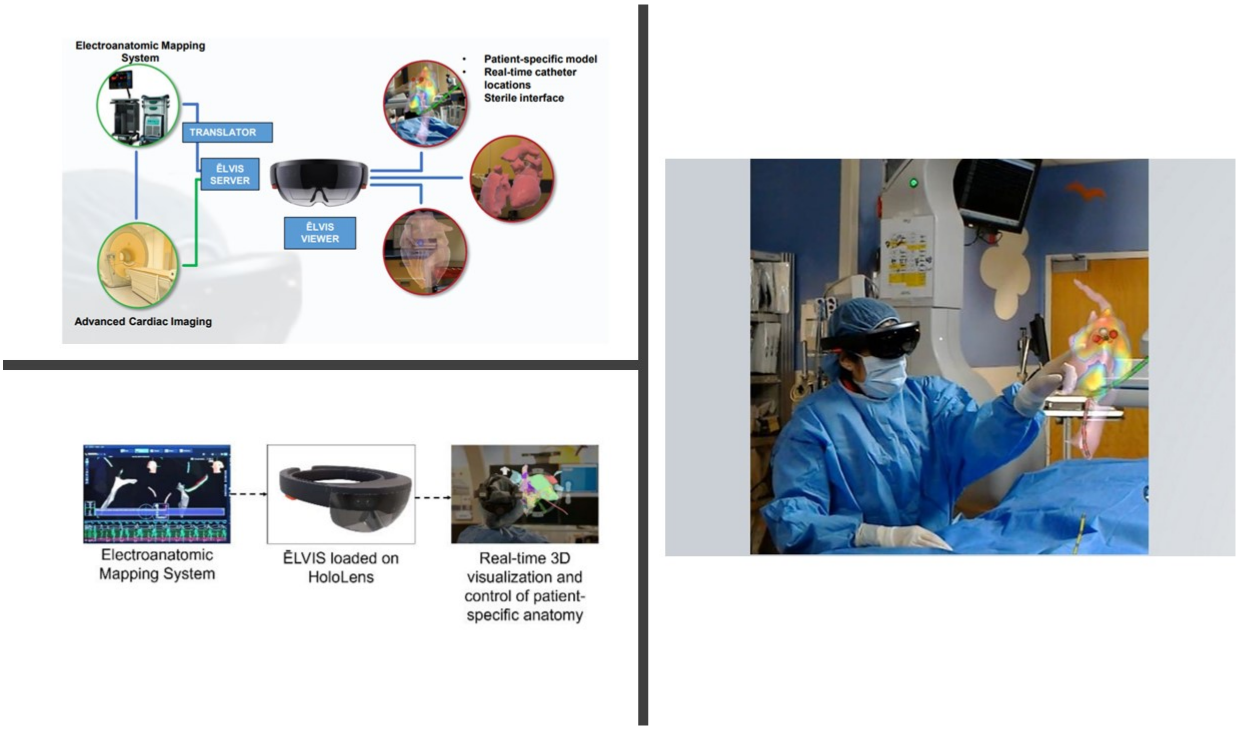

Electroanatomic Mapping System
TRANSLATOR
ĒLVIS SERVER
ĒLVIS VIEWER
Advanced Cardiac Imaging
Patient-specific model
Real-time catheter locations
Sterile interface
Electroanatomic Mapping System
ĒLVIS loaded on HoloLens
Real-time 3D visualization and control of patient-specific anatomy

Figure 1.6: *Left: Workflow for ĒLVIS. From the electroanatomic mapping system, data flows to the ĒLVIS application which runs on Microsoft HoloLens. Through the headset, the real-time patient-specific data is displayed in 3D. Right: holographic information provided by ELVIS.*

Along with procedural time, also the need for intra-operatory imaging performed through ionizing radiations may be significantly reduced by exploiting XR. This would lead to a consequent reduction in the dose for patient and operator and in the use of contrast medium. The final goal of using XR in the context of surgical guidance is therefore to develop a platform able to update the hologram in real-time, exploiting intra-procedural echographic images and pre-procedural CT scans. This would be the disruptive innovation which would allow the elimination of the use of intra- procedural X-ray imaging. A first attempt in this direction was described by Kesprzak et al. [21], whose study focused on the development of a MR platform to support percutaneous mitral valve structural intervention through the rendering of real-time 3DTEE images. The system allows for real-time streaming of 3DTEE images to a 3D DICOM viewer workstation for real-time rendering. The 3D hologram is then transferred wirelessly to an HMD, which allows for is visualization as a semitransparent model superimposed on the surgical field (Figure 1.7). The hologram is shared between echocardiographers and interventional cardiologists and the interaction is voice- and gesture-guided. By testing the system in a real procedure consisting of percutaneous mitral balloon commissurotomy [2], the study demonstrated the feasibility of real-time intraprocedural use of holographic MR models displaying 3DTEE data stream to guide the percutaneous procedure.

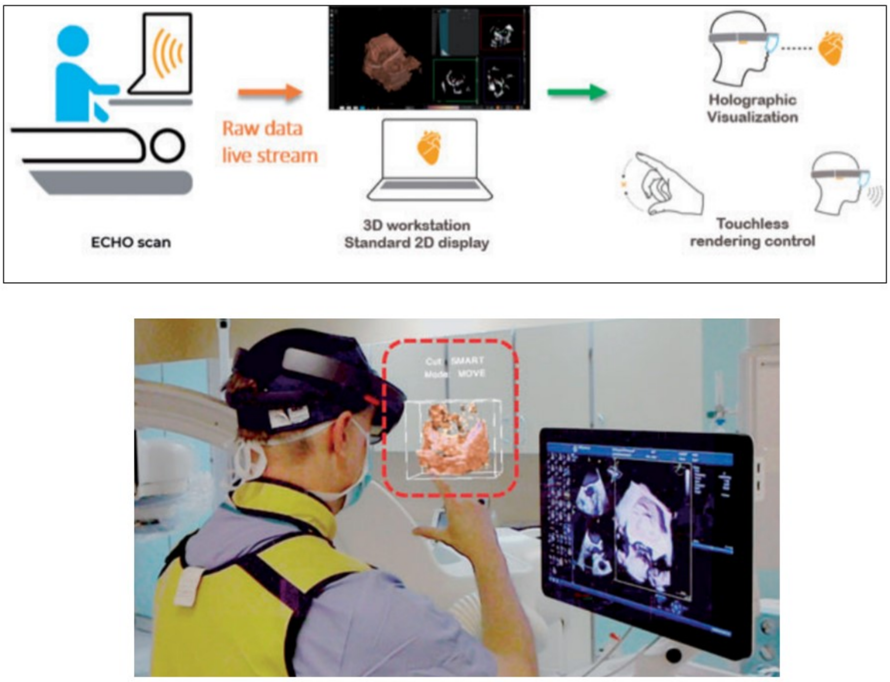


Figure 1.7: *Top: schematic representation of the workflow followed to have the real-time holographic streaming of 3DTEE images during the procedure. Bottom: result of the process.*

As reported by Arian Arjomandi Rad et al. [28], the hope is also that these innovative rendering techniques make it possible to treat more and more cases with minimally invasive procedures rather than via sternotomy, reducing intra-operative trauma and accelerating post-operative recovery.

Referring again to intra-operatory support, another promising application of MR is remote proctoring. The term proctoring refers to the support provided by medical device companies or more expert operators to doctors in the OR when these must learn or perform new surgical practices, or they need supervision during the implantation of new devices. This is normally done in person by expert staff, who supports operators directly inside the OR. Remote proctoring instead allows fulfilling the same tasks without the need for the physical presence of the specialist in the OR: the operator and the specialist have simply to wear an HMD that allows both of them to visualize a shared MR environment

[2] in cardiac surgery, the treatment of mitral valve stenosis through the expansion of a fluid-filled balloon inserted percutaneously, i.e., without thoracotomy.

containing the anatomical model of the patient together with the relevant clinical data acquired before or during the surgery.

Even the post-surgery phase could considerably benefit from the application of XR, namely VR, allowing for easier and more efficient management of the discharged patient. For example, it is possible to deliver rehabilitative treatments to a patient wearing an immersive headset so that he is more involved in the task.

Stress and intra- or post-surgery pain can be effectively managed through VR: by making the patient immersed in a totally virtual environment, he is alienated from the actual condition and this distraction has a positive effect on reducing his pain perception. As demonstrated by Mosso et al. [25], the application of VR distraction therapy can effectively reduce post-operative pain and stress, promoting the overall well-being of the patient. Out of 67 patients tested in this study, 88% experienced a significant reduction in pain and physiologic changes connected to relaxation, such as decreased heart rate, respiratory rate, and arterial blood pressure, after only 30 minutes of VR therapy.

For what concerns remote rehabilitation, a large randomized study by Cacau et al. [15], compared a VR-based platform with traditional rehabilitation protocols in post-cardiac surgery patients. 60 patients were randomized to receive either rehabilitation through conventional physical therapy or through VR. Both groups were treated twice a day (in the morning and in the afternoon) with physical therapy treatment including breathing exercises, airway clearance techniques, metabolic exercises, and motor exercises. Patients allocated in the control group performed the treatment in a conventional way while patients allocated in the VR group performed the motor exercises using virtual reality. The results showed better outcomes for the VR group with respect to the control group in terms of reduced postoperative pain, improved functional performance and walking capacity, higher energy levels, faster recovery, and shorter hospital stay. Thus, VR-assisted rehabilitation proved to be effective for the recovery of patients after cardiac surgery.

In the end, another relevant application of XR concerns communication and education. The use of 3D holographic models facilitates communication between physicians and patients or relatives. This allows for an improved patients understanding of their condition, which is followed by increased situational awareness, self-management skills, and patient empowerment.

Again, the use of MR models but also of immersive VR applications gives medical students and trainees the possibility to easily learn anatomical concepts and train their surgical skills in a realistic environment. They are provided with unlimited cases to be analysed, including complex and rare anatomies, and they are not impacted by the limitations of animal models, such as natural anatomical differences, limited availability, and high costs. Moreover, the possibility of repeating the same learning experience improves the efficiency of the learning process, leading to a more educated generation of doctors, as reported by Battaluga et al. [9]. In fact, he demonstrated better study grades in a group of 100 students instructed through 3D anatomic models versus conventional teaching material.

Beyond the advantages for the patients, in terms of clinical benefits, and for the operators, in terms of enhanced performance possibilities, the technology can also have an economic impact. Referring to consumers like hospitals or medical centers, the costs of acquisition of the technology from an external developing company can range from 10'000 to 100'000 €/year. This figure corresponds to the cost of a single license of the segmentation and rendering software. Depending on the type of contract, it can include the loan/rent/leasing of XR hardware such as headsets, workstations, and monitors. In other cases instead, the acquisition of the hardware is at the expense of the customer. Conversely, services such as operators' training, hardware maintenance, and software updates are often included in the license cost. Despite the high initial costs, in the long term the technology might be cost-effective since it brings several economic advantages, both direct and indirect. The direct savings are related to the fact that it does not require the purchasing of material and the disposal of waste products, it eliminates the costs of cadaveric and animal models for surgical training, and it reduces significantly the cost of patient management, personnel, and travels thanks to the possibility of remote assistance. In an indirect way, the faster surgical planning, shorter duration of the procedure, improved surgical outcome, and shift from open to mini-invasive surgery improve the general efficiency of the system thus reducing the costs for the health care facility.

1.2. Mixed Reality

As previously stated, this work focuses on the pre-procedural and intra-procedural application of XR in cardiac surgery and interventional cardiology. Hence, the most relevant technology in this context is MR, since it allows the operator to visualize holographic 2D or 3D models and to interact with them without losing the view and the possibility to interact with the real environment and with the other people physically present around him.

1.2.1. Stereoscopy

Most MR technologies are based on stereoscopy. Stereoscopy is an image visualization technique that mimics the binocular vision of the human visual system with the goal of showing a tridimensional image. In natural eyesight, the eyes see the same image from two different positions, the brain receives these two pieces of information and, by overlapping them, elaborates the depth perception and computes distances. When the observed object is closer to the subject, the deviation of the object's position in the images which arrive at each eye is greater and a smaller depth is perceived by the brain while, when the object is far from the subject, this deviation is smaller and greater depth is perceived by the brain. The same effect can be reproduced starting from 2D images, by showing each eye a slightly different image, thus obtaining depth perception. For example, it is possible to exploit two different colour filters, as in the case of anaglyph 3D glasses, or polarization filters, as in the case of polarized 3D glasses (Figure 1.8). In the first case, two monochromatic images are generated and received each by one eye. In the second case instead, a projector with two orthogonally polarized lenses generates two differently polarized images which are received by the eye covered by the respectively polarized lens. This principle is exploited in several applications, among which there are cinema and video gaming.

In MR, the head-mounted display projects two slightly different images to the two eyes so that they converge on an artificial depth plane to generate 3D perception. In this work, we will refer to the digital content generated through stereoscopy on MR platforms as holograms even if the term is not appropriate. The principle of hologram generation, in fact, is different from the one of stereoscopy: it is obtained through the interaction of two light beams (the beam reflected by the element to be rendered and the reference laser beam) on a sheet of sensitive material, called holographic film[3].

[3] The process of generation of the hologram starts from image registration: this phase consists in dividing a laser beam through a beam splitter and sending one of the two obtained rays toward the object to be rendered and the other one toward a sheet of sensitive material, called holographic film. On this sheet, the light beam that directly hits it interacts with the one reflected by the object, forming the so-called interference fringes. Laser beams are used instead of white light since all the components go in the same direction, have the same wavelength, and are coherent, i.e., in phase. Two coherent beams create regular interference patterns when they overlap. In the case of hologram generation, the reference beam is coherent while the object beam is not. The interference pattern created when the beams interfere contains all the information on the object to be reproduced in the form of intensity and phase of the reflected light. The following step consists in sending just the reference laser beam toward the interference fringes so that it can pass through their crevices, which work as a diffraction grating. As a result, the laser beam is diffracted and it generates light waves that, overlapping each other, rebuild the original object beam and hence the previously registered image in 3D, which can be visualized by the user as if it was physically present. Simply speaking, if we refer to the reference beam with the letter A and to the image-reflected beam with the letter B, their interference can be viewed as an algebraic summation A+B=C, where C represents the interference fringes. In the second step, the passage of the reference beam through the interference fringes can be simplified as the subtraction C-A=B, which in turn returns B, i.e., the image beam.

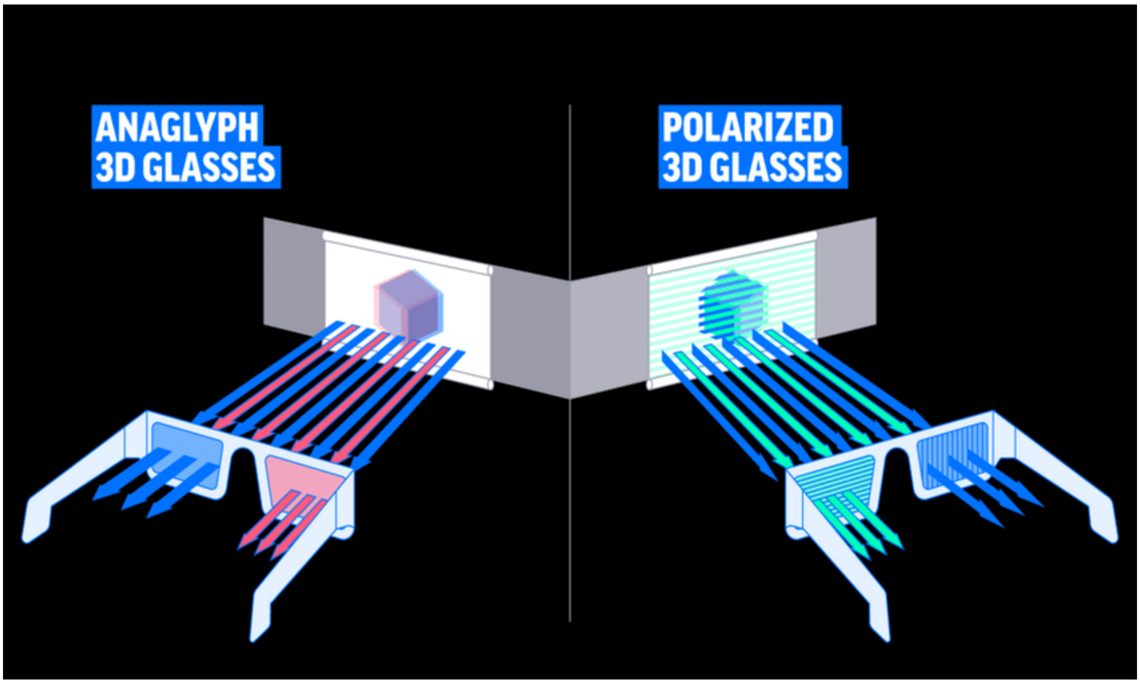


Figure 1.8: *Examples of two types of glasses which allow stereoscopic vision: anaglyph 3D glasses (left) exploit two different colour filters, polarized 3D glasses (right) exploit two differently polarized lenses. In both cases, each eye receives a slightly different parallel image which are then overlapped by the brain to obtain the 3D rendering.*

1.2.2. MR hardware

Beyond the workstation dedicated to data processing, the hardware required to visualize holograms consists of holographic headsets: these are head-mounted see-through displays with semi-transparent lenses mounted on a support to be placed on the head of the user. One of the most promising models, which was used in the work herein presented, is HoloLens 2 (Figure 1.9). It is an untethered holographic computer that runs on the Windows Holographic Operating System, launched in November 2019 as an improved version of the first-generation HoloLens.
The workstation is basically a PC on which, through specific software, digital images are segmented and elaborated to create the 3D virtual model. The model is transferred from the workstation to HoloLens 2 through Wi-Fi connection and image rendering is implemented on the HMD through the on-board processor, also referred to as Holographic Processing Unit (HPU).

The main features of the Hololens 2 device are here summarized, while more details are provided in Table 1.1. The headset weighs 566 grams, it is single size but provided with an adjustable band and can also be worn over eyeglasses. The images to be shown to the

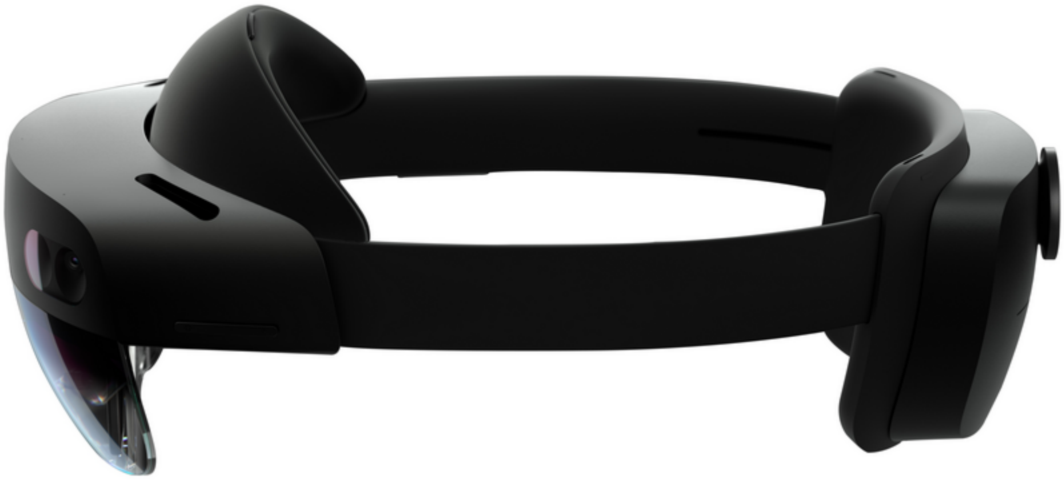


Figure 1.9: *Microsoft HoloLens 2 headset (side view).*

user are projected on semi-transparent lenses, which, as already explained, allow the user to keep the direct vision of the native environment while visualizing the holograms. To project the light information in the user's eyes HoloLens uses a flat waveguide, a thin sheet of transparent material with an entry area. Once inside the guide, the light is maintained there through total internal reflection until it reaches the exit area and is released directly into the user's eye. The resolution of the monitor is 1440x936 pixels and the field of view (FOV) is 43°x29°.

Table 1.1: *Details of the different features of the HoloLens 2 head-mounted display. From https://www.microsoft.com/en-us/hololens/hardware [23]*

COMPONENT	**DETAILS**
Display	**Optics**: See-through holographic lenses (waveguides) **Resolution**: 2k 3:2 light engines **Holographic density**: >2.5k radiants (light points per radian) **Eye-based rendering**: Display optimization for 3D eye position
Sensors	**Head tracking**: 4 visible light cameras **Eye tracking**: 2 IR cameras

(Continue to next page)

(Continued from the previous page)

	Depth: 1-MP time-of-flight (ToF) depth sensor **IMU**: Accelerometer, gyroscope, magnetometer **Camera**: 8-MP stills, 1080p30 video
Audio and Speech	**Microphone array**: 5 channels **Speakers**: Built-in spatial sound
Human Understanding	**Hand tracking**: Two-handed fully articulated model, direct manipulation **Eye tracking**: Real-time tracking **Voice**: Command and control on-device; natural language with internet connectivity **Windows Hello**: Enterprise-grade security with iris recognition
Computer and Connectivity	**SoC**: Qualcomm Snapdragon 850 Compute Platform **HPU**: Second-generation custom-built holographic processing unit **Memory**: 4-GB LPDDR4x system DRAM **Storage**: 64-GB UFS 2.1 **Wi-Fi**: Wi-Fi 5 (802.11ac 2x2) **Bluetooth**: 5 **USB**: USB Type-C
Fit	**Single size**: Yes **Fits over glasses**: Yes

(Continue to next page)

(Continued from the previous page)

	Weight: 566g
Software	**Windows Holographic Operating System** **Microsoft Edge** **Dynamics 365 Remote Assist** **Dynamics 365 Guides** **3D Viewer**
Power	**Lithium batteries** **Battery life**: 2–3 hours of active use **Charging**: USB-PD for fast charging **Cooling**: Passive (no fans)

Table 1.1: *Details of the different features of the HoloLens 2 head mounted display. From https://www.microsoft.com/en-us/hololens/hardware [23]*

The interaction with the hologram takes place in different hands-free modalities, without the need for additional hardware such as mouse, keyboard, or controllers. Different sensors track the position of the user with respect to the environment and enable him to interact with, zoom, or move the virtual objects. Defined gestures are recognized by the cameras of the headset when they are performed inside their field of view; voice control can be configured in order to associate specific words recognized by the microphones to the execution of specific actions; eye-tracking works by monitoring the direction of the gaze of the subject, represented by a ray emitted from the center of the point of view.

Head tracking is performed through 4 visible light cameras while eye- and hand-tracking is performed through 2 infrared (IR) cameras. Other embedded sensors are a depth sensor and an inertial measurement unit (IMU). The latter is composed of an accelerometer (used by the system to determine linear acceleration along the X, Y, and Z axes and gravity), a gyroscope (used by the system to determine rotations), and a magnetometer (used by the system to estimate absolute orientation with respect to the magnetic heart field). Through the IMU, the headset continually tracks the position and orientation of the user's head relative to the surroundings. The depth sensor is based on IR illumination and works

through phase-based time-of-flight (ToF) for spatial mapping and hand tracking. ToF camera functioning consists in illuminating the scene with a modulated light source and observing the reflected light: the phase shift between the illumination and the reflection is measured and translated into a distance.
The focal distance of HoloLens 2 is close to 2 meters, and this is a piece of important information to be considered when planning the hologram placement in order to avoid the vergence-accommodation conflict. In natural viewing, vergence refers to the simultaneous movement of the eyes which allows for the binocular vision of an object without seeing double images. The eyes converge when an object is close to the observer while they diverge as the object moves away. Accommodation instead, refers to the contraction (when the object is far away) and distention (when the object is close to the observer) of the crystalline lens of the eye to focus on an object. In natural eyesight, these two phenomena are connected and they work in agreement. In XR headsets instead, they may not happen at the same distance, meaning that our brain receives different depth cues that generate a conflict situation. Vergence changes based on the distance of the object to get a single image while accommodation is fixed at the focal distance of the display (often close to 2 m) to get a sharp image. This situation is referred to as vergence-accommodation conflict and it is characterized by eye fatigue, visual discomfort, and nuisance. In particular, this problem is more perceptible when the distance of the virtual object is closer than 2 meters and when its position changes in time, i.e., when it moves in space. Therefore, the best choice is to maintain the hologram still at a distance of about 2 m from the user so that the vergence happens on the same plane of the accommodation. When it is not possible, it is better to avoid the user's gaze having to move back and forth between different distances, otherwise, user comfort can be compromised. Furthermore, again due to the vergence-accommodation conflict, the perception of depth and the focus of the hologram can conflict with that of real objects, with consequent alteration of depth perception. As a result, the user may have difficulty interacting with real and virtual objects at the same time. Therefore, it is essential to try to reduce the conflict as much as possible. Since eye vergence is influenced by the interpupillary distance (IPD), it is advisable to perform calibration of the device prior to the use of an XR headset: in this way, the IPD distance is computed and set up in the device. For HoloLens 2 it is possible to skip the calibration process without very negative consequences, especially if the use is short-term.

A USB port allows the recharge of the device and a battery its wireless functioning. The lithium battery is estimated to last 2-3 hours in active use and up to two weeks in standby mode. Moreover, the device is completely functional while charging. In order to ensure the mobility of the user, HoloLens 2 is also equipped with wireless Wi-Fi connectivity and

Bluetooth. The system is provided with 64 GB of storage memory and 4 GB of RAM. The sensor streams can either be processed or stored on the device or transferred wirelessly to another PC or to the cloud for more computationally demanding tasks. This opens a wide range of new computer vision applications for HoloLens 2.

A comparative study led by S. Moosburner et al. [24] compared the performance of HoloLens vs. Meta 2 (Figure 1.10), an alternative MR headset, in visceral surgery applications. The study consisted of a usability analysis run through a modified version of the User Experience Questionnaire (UEQ), a validated tool for measuring usability. It highlighted some of the major differences between the devices. At first, Meta 2 is tethered to a PC while HoloLens is a standalone mobile computer. This difference implies a limitation in the movement of the end-user in the OR when using Meta 2, but it also allows Meta 2 for better image quality, with improved resolution (2560x1440 pixels vs. 1280x720 pixels) and larger FOV (82.2°x52° vs. 35°x17°). The two devices also exploit different techniques to project light information into the user's eyes: while HoloLens exploits a flat waveguide as previously explained, Meta 2 is based on a downwards-facing display reflected by a curved transparent metal-coated surface into the user's eyes. The curved surface causes distortion in the image and this makes the calibration process strictly necessary. The result of this study showed that the broader FOV and the higher resolution display of Meta 2 did not lead to superior evaluations of the performance of the device. The major reasons were its wired design and the necessity for a powerful computer to allow its functioning, whereas HoloLens works in a standalone and wireless modality. Moreover, the calibration of Meta2 was reported by the user to be too long and complex and the image often resulted in being still out of focus. The larger FOV of Meta 2 allowed the user to see easily the whole model, even if it generated eye strain when the model was more than 1 meter away. Both HoloLens and Meta 2 use the same systems for object tracking and gesture recognition but on the first device the tracking was reported to be more spatially stable and the gestures to be easily recognized by the system. To sum up, the study demonstrated the superiority of HoloLens in supporting providers in a surgical setting. Since HoloLens 2 is characterized by improved resolution and larger FOV and by no limitations as compared to HoloLens, the conclusions of the study can be extended also to HoloLens 2.

1.3. The technology under evaluation

This study focuses on the MR platform developed by Artiness, a startup founded in 2018 by researchers and professors from the Bioengineering department of Politecnico di Milano.

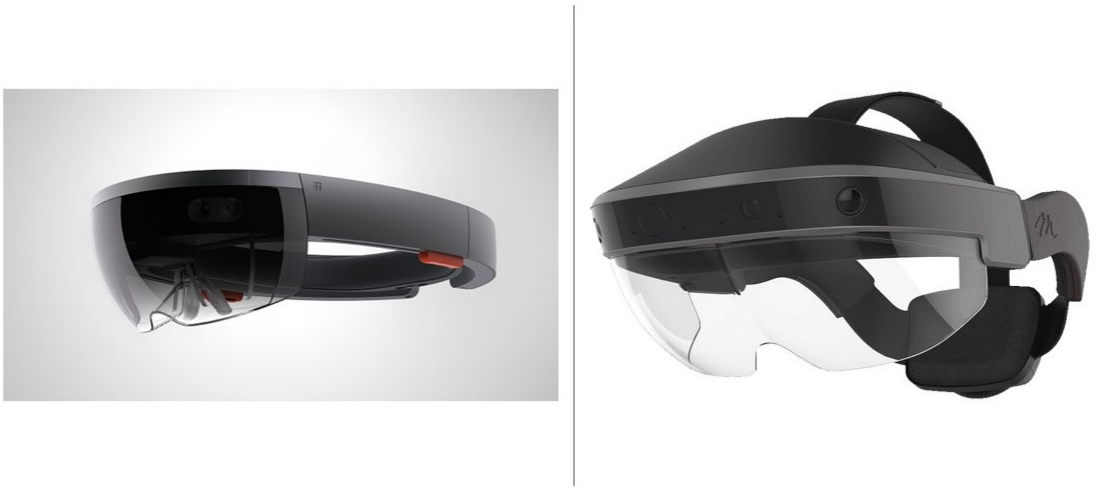


Figure 1.10: *The two HMDs compared in the study by S. Moosburner et al. [24]. Left: Microsoft HoloLens headset (1° generation). Right: Meta2 headset.*

They develop MR platforms for pre-operatory planning and intra-operatory support in the context of interventional, structural, and vascular cardiology, with a particular focus on mini-invasive procedures for valves repair or replacement. The goal is to provide clinicians with a reliable and intuitive tool that can simplify their work and improve the clinical outcomes for the patients. The developed solutions are based on the holographic visualization of patient-specific medical data derived from TC images. The applications are currently developed in Unity and run on HoloLens 2, the head-mounted display created by Microsoft and described in the previous subsection 1.2.2.

In the pre-procedural context, 3D holographic models are meant to allow for the realistic navigation and analysis of the patient's anatomy, thus simplifying procedural planning and making it easier to account for any particular characteristic of the patient to be treated. In this context, the start-up team has developed ARTICOR platform, which has already been certified as class I Medical Device Software (MDSW) in 2021, according to the Medical Device Directive (MDD) 93/42 [2]. This software platform manages the whole process from image segmentation to holographic rendering. Medical images in DICOM format are retrieved by the operator from the hospital PACS and uploaded on a workstation dedicated to data processing: images are segmented through proprietary algorithms and a 3D model of the relevant anatomy is generated. The models created in this way are then stored on secure cloud platforms and sent, in wireless modality, to a standalone head-mounted display (HoloLens 2), through which it is possible to interact with both the hologram and the original images without the need for additional hardware. The platform allows the user to navigate the reconstructed anatomical model and to virtually position implantable devices in the respective implantation sites, in order to evaluate their sizing with respect to the surrounding anatomy on patient-specific mixed reality 3D models[4]. Additionally, ARTICOR offers the possibility to share the hologram among multiple users and to stream the holographic content on a monitor so that the discussion of the clinical case can be supported by the heart team.

In the context of intra-operative applications, Artiness is developing a MR platform allowing for remote proctoring, which is being currently evaluated in an ad hoc clinical trial. Through two head-mounted displays, the operator who's physically present inside the OR and the remote proctor can remotely share the MR visualization of heterogeneous contents, including medical data, signals from the OR, images streamed from monitors of the OR, pre-op or intra-op medical imaging with the corresponding 3D anatomical models. Both end-users can interact with the visualized contents, having the possibility to navigate imaging datasets and 3D models, as well as to annotate any content. The effects of any manipulation by one end-user are visible by the other with minimal latency thanks to 5G technology.

1.4. Usability: definition and relevance

Usability is formally defined in ISO 9241-11 (*Ergonomics of human-system interaction - Usability: Definitions and concepts*) [6] as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" and as "a multidimensional quality that refers to the ability of a human to interact easily and relatively error-free with a system or product" by the British Standards Institution (BSI) Group. Unlike more technical specifications, which concern the device per se, the concept of usability pertains to the interaction between the user and the system in relation to specific use environments, user goals, and user expectations about the way the system should work. Therefore, it is not an intrinsic characteristic of the device, but it strictly depends on the user, the task, and the use environment.

Usability engineering refers to the design and development process of the user interface (UI) of a product and it is intended to identify and minimize use errors and the associated risks and to determine whether the device will meet the intended users' needs and expectations. In order for the product to respond to usability standards, it is clear that its project and design should be "user-centered", i.e., guided by the knowledge of end-user needs, features, expectations, and context of use. This approach is aimed at guarantee-

[4] In ARTICOR, the thickness of the myocardium is not reconstructed but the rendered surface corresponds to the endocardium since the device will come into contact with it. Typically, the devices are 10% oversized relative to the anatomical measurement of the implant site.

ing the device's safety by reducing the probability of use errors and use-associated risks occurring, and at improving its effectiveness by creating an intuitive, easy-to-learn, and easy-to-use device.

The concept of usability is applicable to the development of any kind of product but, in the context of medical devices, it has an even greater relevance since incorrect or time-inefficient use of products due to usability issues can lead to late intervention, wrong diagnosis, and serious injuries for the patient. For these reasons, usability assessment is becoming a widespread activity during medical device design, and it will soon become a common practice since in the new Medical Device Regulation (MDR) 2017/745 [5] there are many references to the concept of usability and human factors. Annex I: *General safety and performance requirements*, refers to the evaluation of situations that can bring to use errors thus implicitly to usability evaluation. In particular, Requirement 1 states that medical devices must be suitable for their intended use under normal use conditions, being safe and effective, without compromising the safety and health of any user. Possible associated risks must be acceptable considering the benefits brought to the patient by the device and the state of the art. Requirements 2, 3, 4, and 8 concern risk reduction and control stating that: manufacturers must implement, document, maintain, and update a risk management system in order to *a) establish and document a risk management plan for each device; b) identify and analyze known and foreseeable hazards associated with each device; (c) estimate and evaluate the risks associated with and occurring during intended use and during reasonably foreseeable misuse; d) eliminate or control these risks; e) evaluate the impact of the information coming from the production phase and, in particular, from the system post-market surveillance.* Risks associated with the device must be reduced as much as possible, in order to be acceptable with respect to the benefits, without affecting the risk-benefit ratio, and risk control measures must be applied to comply with safety principles, taking into account the generally recognized state of the art. Manufacturers must manage risks so that the residual risk associated with each hazard, as well as the overall residual risk, is considered acceptable. To this goal, manufacturers must, in order of priority: *(a) eliminate or reduce risks as far as possible through safety in design and manufacture; b) where appropriate, take protective measures, including warning signals if necessary, in relation to risks that cannot be eliminated; and c) provide safety information (warnings/precautions/contraindications) and, where appropriate, user training.* Manufacturers must always inform users about residual risks. Requirement 5 focuses on the elimination or reduction of risks connected to use errors and specifies in particular that manufacturers must: *a) reduce, as far as possible, the risks associated with the ergonomic characteristics of the device and the environment in which it is intended to be used (design for patient safety); and b) consider the level of technical knowledge, experience, education,* *training and environment of use, and, where possible, the medical and physical conditions of the intended users (design for lay, professional, disabled or other users).* This requirement contains the most explicit reference to usability. As a consequence, when applicable, risk evaluation and management should include the assessment of risks and hazards related to human factors.

The set of records and documents produced during the *usability engineering process* is referred to as *Usability Engineering File* and, when this evaluation is performed, it should be included in the technical documentation to be presented when applying for the CE mark. If the usability assessment is not performed, it is however necessary to justify the reasons why it has been declared non-applicable.

1.4.1. Reference regulatory context

In the context of medical devices, the regulatory framework for usability is based on the Italian standard CEI EN 62366-1: *Application of usability engineering to medical devices*, which derives from the transposition of the respective international standards IEC 62366-1 [3], and on the Technical Report IEC TR 62366-2: *Guidance on the application of usability engineering to medical devices* [4]. IEC 62366-1 was harmonized with MDD 93/42 [2] but not with MDR 2017/745 [5] since the regulation has been published later. Anyway, it is the reference standard for usability both in Europe and in the USA thus its application is very useful to medical device manufacturers who want to demonstrate the usability of their products in the certification process. The standard strictly focuses on usability as it relates to safety, defined in the standard itself as “freedom from the unacceptable risk that can arise from use error and lead to exposure to direct physical hazards or loss or degradation of clinical functionality”. It defines a process for the manufacturer to assess and mitigate risks related to normal use, i.e., correct use and use errors, as well as to identify risks related to abnormal use [5], with the goal of specifying, developing, and evaluating the aspects of usability related to safety. Consequently, the usability engineering process has a strict interrelationship with the risk management process of ISO 14971 (*Application of risk management to medical devices*) [7], as shown in Figure 1.11.

[5] Normal use: operation, including routine inspection and adjustment by any user, and stand-by, according to the instructions for use or in accordance with generally accepted practice for those medical device provided without instruction for use. Use errors can occur in normal use.
Abnormal use: conscious, intentional act or intentional omission of an act that is counter to or violates normal use and is also beyond any further reasonable means of user interface-related risk control by the manufacturer.

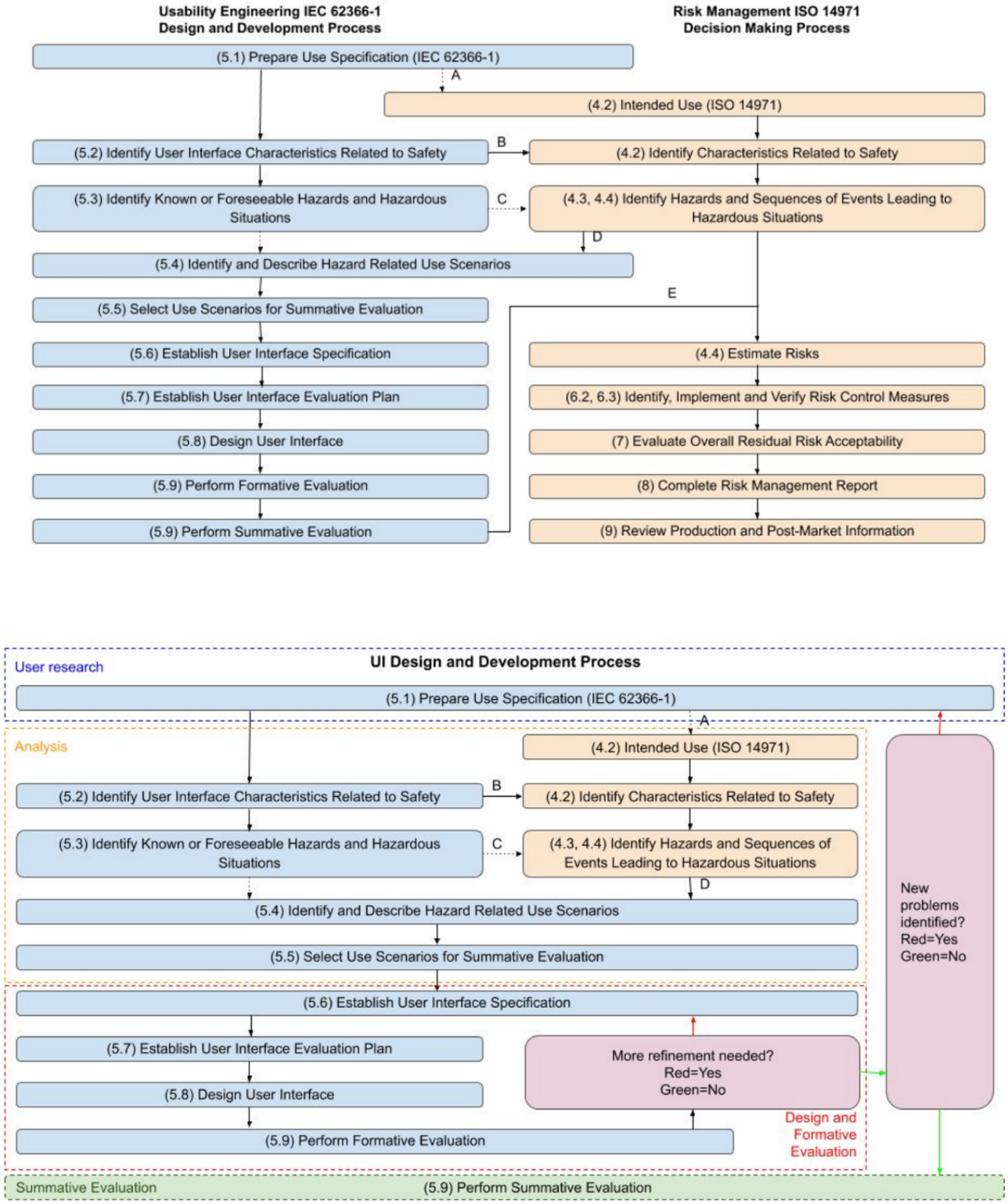

Usability Engineering IEC 62366-1
Design and Development Process
Risk Management ISO 14971
Decision Making Process
(5.1) Prepare Use Specification (IEC 62366-1)
A
(4.2) Intended Use (ISO 14971)
(5.2) Identify User Interface Characteristics Related to Safety
B
(4.2) Identify Characteristics Related to Safety
(5.3) Identify Known or Foreseeable Hazards and Hazardous Situations
C
(4.3, 4.4) Identify Hazards and Sequences of Events Leading to Hazardous Situations
D
(5.4) Identify and Describe Hazard Related Use Scenarios
E
(5.5) Select Use Scenarios for Summative Evaluation
(5.6) Establish User Interface Specification
(4.4) Estimate Risks
(5.7) Establish User Interface Evaluation Plan
(6.2, 6.3) Identify, Implement and Verify Risk Control Measures
(5.8) Design User Interface
(7) Evaluate Overall Residual Risk Acceptability
(5.9) Perform Formative Evaluation
(8) Complete Risk Management Report
(5.9) Perform Summative Evaluation
(9) Review Production and Post-Market Information
User research
UI Design and Development Process
(5.1) Prepare Use Specification (IEC 62366-1)
A
Analysis
(4.2) Intended Use (ISO 14971)
(5.2) Identify User Interface Characteristics Related to Safety
B
(4.2) Identify Characteristics Related to Safety
(5.3) Identify Known or Foreseeable Hazards and Hazardous Situations
C
(4.3, 4.4) Identify Hazards and Sequences of Events Leading to Hazardous Situations
D
New problems identified? Red=Yes Green=No
(5.4) Identify and Describe Hazard Related Use Scenarios
(5.5) Select Use Scenarios for Summative Evaluation
(5.6) Establish User Interface Specification
(5.7) Establish User Interface Evaluation Plan
More refinement needed? Red=Yes Green=No
(5.8) Design User Interface
Design and Formative Evaluation
(5.9) Perform Formative Evaluation
Summative Evaluation
(5.9) Perform Summative Evaluation

Figure 1.11: *Top: interrelations between the usability engineering process described in IEC 62366:1 and the risk management process described in ISO 14971. Bottom: medical device user interface (UI) design and development process according to standards IEC 62366-1 and ISO 14971.*

The technical report [4] provides information and guidance to efficiently implement the requirements of IEC 62366-1 [3] but it has a broader focus. It does not only consider usability aspects related to safety, but it also takes into consideration the relationship between usability and task accuracy, completeness, efficiency, and user satisfaction.

These international standards define the workflow to be followed during the project, design, and implementation of a device in order to ensure that the final product is usable and safe from the point of view of user interface and user interaction. The workflow can be coarsely divided into two main phases, named formative and summative evaluation.

Formative evaluation is intended to be performed early and iteratively throughout the product development cycle, with the goal of identifying product strengths and shortcomings, user needs, and opportunities for improvement. These tests are called "formative" since they aim at shaping the product by highlighting the highest possible number of potential problems that can lead to use errors, thus guiding the design of the device. Formative evaluations are typically conducted through a "quick and dirty" approach that allows for the identification of macroscopic deficiencies early in the developing phase so that it is possible to address them when redesign cost and time are still contained. Usually, formative evaluations focus on hazard-related use scenarios or tasks in which use errors can occur. In this way, they allow to determine if the risk controls implemented can effectively prevent harm to the user.

Summative evaluation instead is carried out at the end of the design and development phase, with the goal of validating the safe use of the user interface. It is intended to confirm that representative users can interact with the given device safely and effectively, and that the device does not induce dangerous use errors. Summative evaluation generally is conducted under conditions of simulated use by participants who represent the different user groups of that device. Data are collected in the form of task completion, time to complete the task, feedback by participants, and descriptions of the observed use errors or difficulties. Since the purpose of this phase is to simulate the realistic use of the device, the moderator should not influence participants' behaviour: as much information as possible should be recorded without commenting and avoiding using the "think aloud" method. "Think aloud" consists in asking participants to express their thoughts while they are performing defined tasks. This methodology is useful to collect information about the strategies applied by users, about their difficulties, expectations, and impressions but, at the same time, it can influence user performance and prevent a natural interaction. Therefore, this practice is considered appropriate in formative usability tests, but not in summative usability tests.

Typically, a single summative evaluation is conducted after many formative ones, before applying for regulatory clearance. Data from this last evaluation should allow the manufacturer to conclude that no further user interface improvement is needed or applicable. If the summative evaluation reveals some residual vulnerability to potentially harmful use errors whose residual risk is deemed unacceptable in relation to the benefit, device re-design and re-testing are required, and a further summative evaluation will be needed to validate the modifications implemented.

Going more into detail, the *usability engineering process* is articulated into several steps, as presented in IEC 62366-1 [3] and carefully described in IEC TR 62366-2 [4], where practical examples of their implementation are also provided to the manufacturer. A schematization of the process is reported in Figure 1.12.

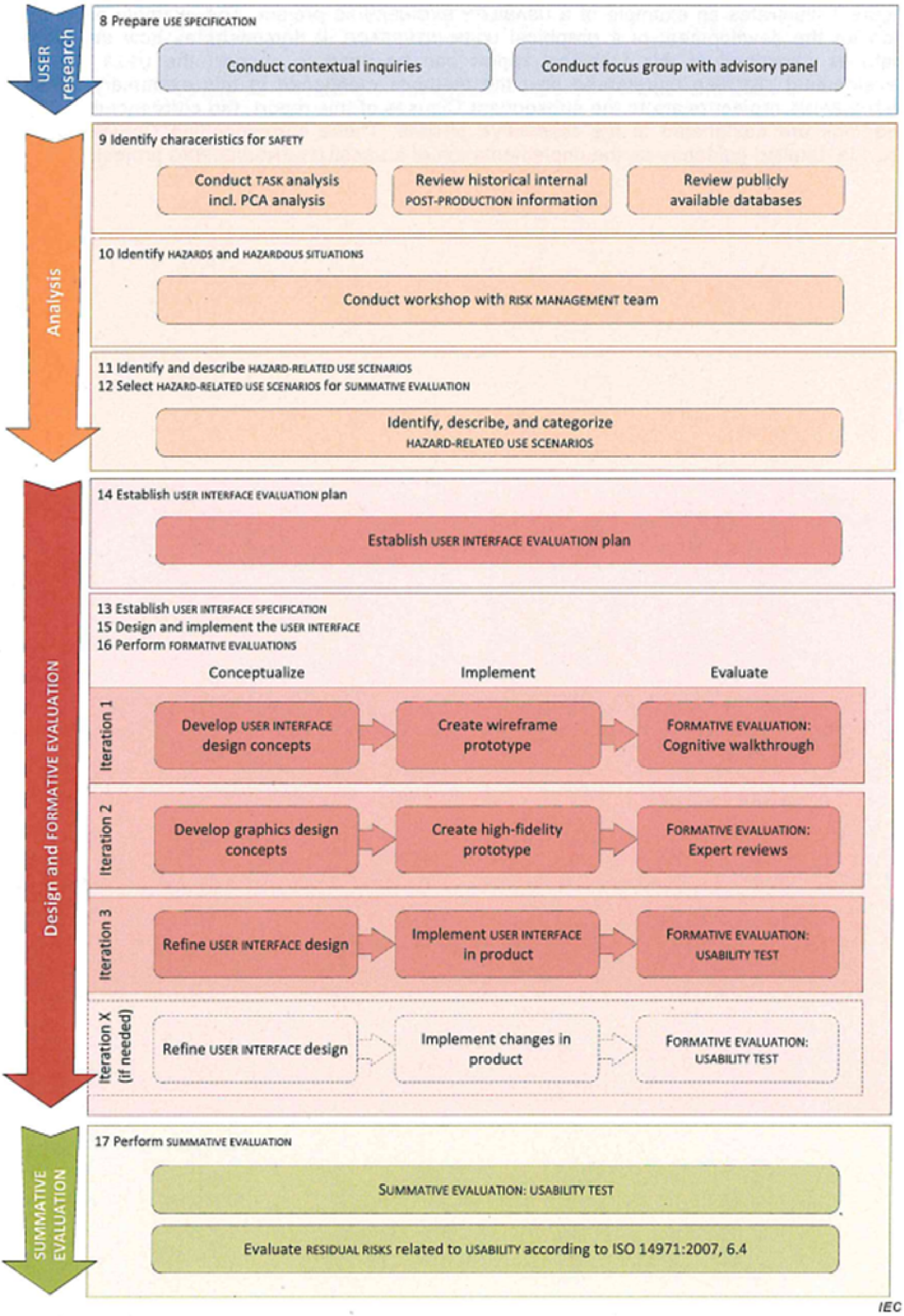

USER research
8 Prepare USE SPECIFICATION
Conduct contextual inquiries
Conduct focus group with advisory panel
Analysis
9 Identify characeristics for SAFETY
Conduct TASK analysis incl. PCA analysis
Review historical internal POST-PRODUCTION information
Review publicly available databases
10 Identify HAZARDS and HAZARDOUS SITUATIONS
Conduct workshop with RISK MANAGEMENT team
11 Identify and describe HAZARD-RELATED USE SCENARIOS
12 Select HAZARD-RELATED USE SCENARIOS for SUMMATIVE EVALUATION
Identify, describe, and categorize HAZARD-RELATED USE SCENARIOS
Design and FORMATIVE EVALUATION
14 Establish USER INTERFACE EVALUATION plan
Establish USER INTERFACE EVALUATION plan
13 Establish USER INTERFACE SPECIFICATION
15 Design and implement the USER INTERFACE
16 Perform FORMATIVE EVALUATIONS
Conceptualize
Implement
Evaluate
Iteration 1
Develop USER INTERFACE design concepts
Create wireframe prototype
FORMATIVE EVALUATION: Cognitive walkthrough
Iteration 2
Develop graphics design concepts
Create high-fidelity prototype
FORMATIVE EVALUATION: Expert reviews
Iteration 3
Refine USER INTERFACE design
Implement USER INTERFACE in product
FORMATIVE EVALUATION: USABILITY TEST
Iteration X (if needed)
Refine USER INTERFACE design
Implement changes in product
FORMATIVE EVALUATION: USABILITY TEST
SUMMATIVE EVALUATION
17 Perform SUMMATIVE EVALUATION
SUMMATIVE EVALUATION: USABILITY TEST
Evaluate RESIDUAL RISKS related to USABILITY according to ISO 14971:2007, 6.4
IEC

Figure 1.12: *Scheme of the usability engineering process (IEC TR 62366-2).*

In the following lines, a description of the single steps composing the *usability engineering process* is reported:

1. **Prepare use specifications**: this step is a fundamental prerequisite to designing the medical device and its user interface. Specifications shall include intended medical indication, intended patient population, intended part of the body or type of tissue involved, intended user profile, anticipated user tasks, intended use environment, and operating principle. Use specifications are refined over time as more knowledge is obtained through user research methodologies such as contextual inquiries, i.e., observation of users interacting with the device while performing realistic tasks, and focus groups, i.e., group interviews led by a moderator to investigate perceptions, opinions, and attitudes of the users. This information also constitutes input for the risk management process described in ISO 14971 [7];

2. **Identify user interface characteristics related to safety and potential use errors**, and update this information throughout the whole design and development process. The identification starts from the definition of the primary operating functions, i.e., those functions of the device involving user interaction that can influence safety. The primary operating functions can be analyzed through task or function analysis. Task analysis consists in identifying the sequence of tasks necessary to perform a function, subdividing each task into small steps, and describing each of them in detail to identify the ones potentially subject to use errors. Function analysis instead identifies which functions of a medical device should be automatic, which ones should be performed only by the user, and which ones should be shared between the user and the device. In order to identify characteristics related to safety, also post-production information on similar products can be exploited;

3. **Identify known or foreseeable hazards and hazardous situations**, the latter occurring when a person is exposed to a hazard. In this phase, it is necessary to account for use specifications, identified potential use errors, and information on hazards related to user interfaces of similar medical devices already existing. Since the aim of this phase is to investigate the potential effect that a use error may have and how it can contribute to harm, communication with the risk management team is fundamental;

4. **Identify and describe hazard-related use scenarios**: use scenario refers to a description of the interaction between a user from a specific user profile and the medical device in order to reach a specific result in a certain use environment. The term can refer both to a positive (i.e., correct use) and to a negative (i.e., use error)

situation. In this phase in particular the focus is on use scenarios that can lead to the hazardous situations identified in the previous phase, which are referred to as hazard-related use scenarios. The identified scenarios shall be described through the sequence of tasks composing them, the probability of occurrence, and the severity of the associated harm. Their investigation starts from the identification of hazards and hazardous situations, and it is essential to define user interface requirements;

5. **Select the hazard-related use scenarios for summative evaluation**, so that summative evaluation might be able to demonstrate safety for what concerns the user interface of the device. Depending on the number of hazard-related use scenarios identified, it is possible to include all of them or to perform a selection. This one can be based on the consequence (i.e., severity) of the derived harm or on the risk (i.e., the combination of the probability of occurrence and severity of the harm), if the probability can be estimated. Several formative evaluations should be performed to examine a complete set of user-medical device possible interactions: in this way, there is a high probability that all the relevant hazard-related use scenarios will be identified and included in the summative evaluation, independently of their probability of occurrence;

6. **Establish user interface specifications** as highlighted by the previous steps and develop a document containing all the user interface testable technical requirements. These must be updated when new insights about user needs, preferences, and risks are identified. User interface specifications must be developed also for instruction for use and other accompanying documentation since these are part of the medical device itself;

7. **Establish a user interface evaluation plan**: basically, this phase consists of the planification of the next two stages (formative and summative evaluation). They must be planned to synchronize these activities with the development project and to allocate the necessary resources in advance. The plan should include the goal and the methods that will be used, the participants involved, the test environment, the accompanying documentation, and the training to be provided prior to or during the test;

8. **Perform user interface design, implementation, and formative evaluations**: this is an iterative process leading to the development of a user interface that meets user needs and prevents use errors. In order to prevent use errors, the manufacturer, at first, must try to eliminate hazards and hazardous situations through the “safety by design” approach. If this is not possible, the design should

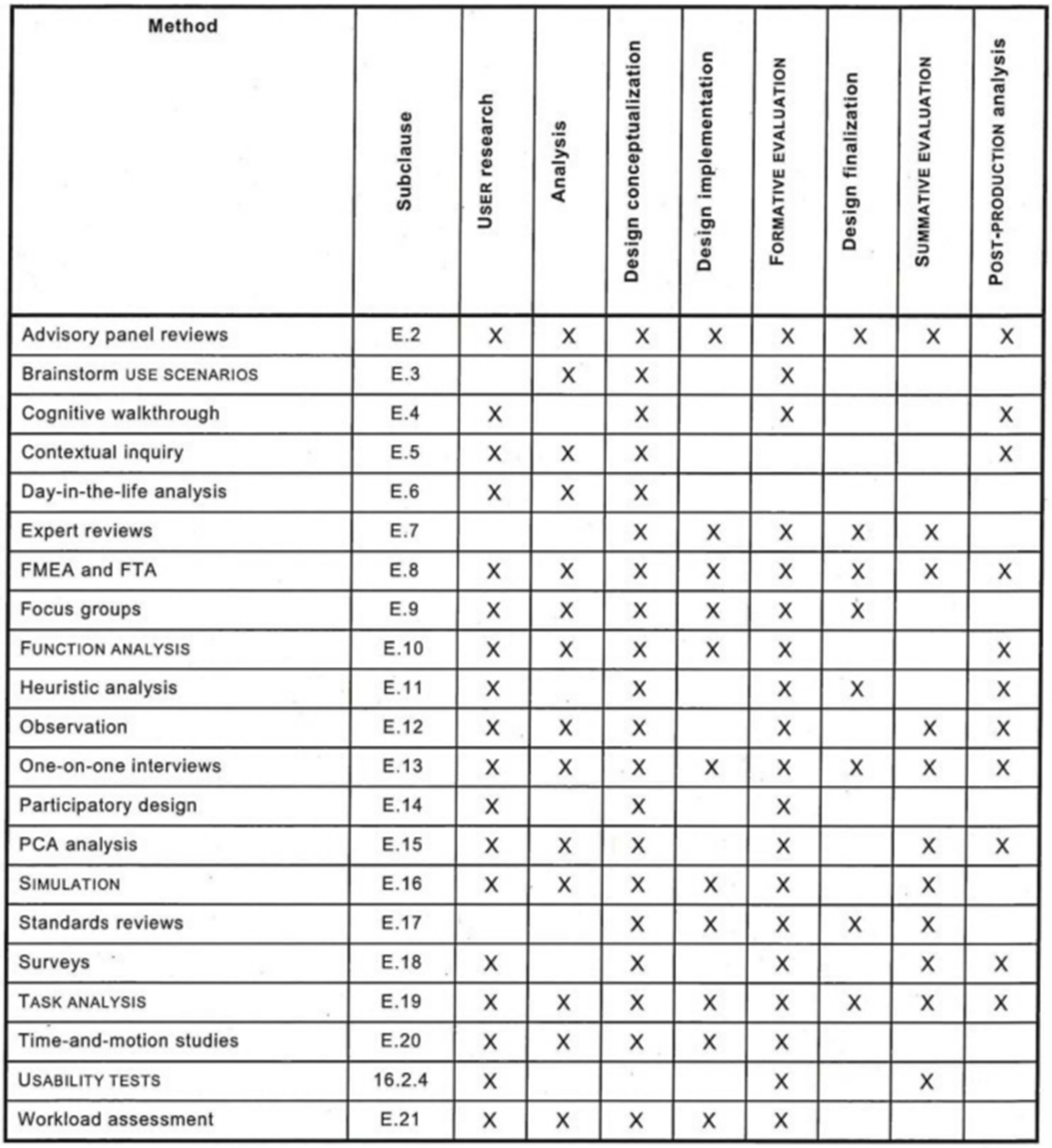
include protections against use errors and, when these might occur despite risk controls, alarm signals and information for safety are necessary to prevent harms. The development process should include also the design of accompanying documentation and the definition of the training required. Along this process, the manufacturer must account for the user interface requirements defined taking into consideration user needs and preferences, use scenarios, and use environments. Feedback from the different steps is collected through formative evaluations: this information is then used to update requirements, leading the design and implementation process from preliminary concepts and prototypes to the final product. Formative evaluations are needed to determine if risk controls can successfully prevent use errors that may lead to harm;

9. **Perform summative evaluation of the usability of the user interfaces**: this last phase should be able to demonstrate that users are able to accomplish the intended purpose of the medical device as described in the use specifications. It should confirm that usability is acceptable, in the sense that use-related risks have been eliminated or reduced to acceptable levels and no further improvements to the user interface are needed or feasible. Again, the evaluation of the residual risk related to usability must be performed considering ISO 14971 [7];

10. **Document the *usability engineering process*** through a report, addressed to internal and external stakeholders. It should include a description of use specifications and user interface and a summary of the results of the previous steps. Differently from the risk management process described in ISO 14971 [7], post-production surveillance is not required by IEC 62366-1 [3] since all use errors should be identified during product development. Anyway, the manufacturer should always be aware that data from post-production surveillance, such as customer complaints, can provide relevant data to support usability engineering activities.

1.4.2. Evaluation methods

TR 62366-2 [4] also carefully describes the different usability engineering methods to be applied in the different phases of the above-mentioned process (Figure 1.13). Focusing on the phase of interest for this work, i.e., summative evaluation, the applicable methods are presented in the following lines:



Method	Subclause	USER research	Analysis	Design conceptualization	Design implementation	FORMATIVE EVALUATION	Design finalization	SUMMATIVE EVALUATION	POST-PRODUCTION analysis
Advisory panel reviews	E.2	X	X	X	X	X	X	X	X
Brainstorm USE SCENARIOS	E.3		X	X		X			
Cognitive walkthrough	E.4	X		X		X			X
Contextual inquiry	E.5	X	X	X					X
Day-in-the-life analysis	E.6	X	X	X					
Expert reviews	E.7			X	X	X	X	X	
FMEA and FTA	E.8	X	X	X	X	X	X	X	X
Focus groups	E.9	X	X	X	X	X	X		
FUNCTION ANALYSIS	E.10	X	X	X	X	X			X
Heuristic analysis	E.11	X		X		X	X		X
Observation	E.12	X	X	X		X		X	X
One-on-one interviews	E.13	X	X	X	X	X	X	X	X
Participatory design	E.14	X		X		X			
PCA analysis	E.15	X	X	X		X		X	X
SIMULATION	E.16	X	X	X	X	X		X	
Standards reviews	E.17			X	X	X	X	X	
Surveys	E.18	X		X		X		X	X
TASK ANALYSIS	E.19	X	X	X	X	X	X	X	X
Time-and-motion studies	E.20	X	X	X	X	X			
USABILITY TESTS	16.2.4	X				X		X	
Workload assessment	E.21	X	X	X	X	X			

Figure 1.13: *Methods to be applied in the usability engineering process (IEC TR 62366-2).*

Advisory panel reviews: typically, they involve 5 to 10 people with different perspectives about the medical device under investigation, therefore not only key opinion leaders or experts but any type of representative users. These reviews should be conducted often and continuously during the development process so that involved people can develop a growing level of knowledge of the product and of its pros and cons. However, this technique does not substitute input from other prospective users who are new to the device.

Expert reviews: these reviews involve instead usability specialists which should identify design strengths and weaknesses and hence propose opportunities for design improvement.

FMEA and FTA: failure modes and effects analysis (FMEA) and fault tree analysis (FTA) are methodologies used to define, identify, and reduce the probability of medical device failures due to inadequate usability. Each failure mode is defined by its frequency of occurrence, the severity of the harm that can result from it, and the effectiveness of the risk control measures implemented. FTA is a "top-down" approach that focuses on one failure event and tries to determine its causes (i.e., failure modes) in successive levels of detail. FMEA is instead a "bottom-up" approach that starts from the identification of potential failure modes and investigates their effect. Both methodologies, however, have the goal of identifying actions to mitigate the failures.

Observation: it allows for the identification of hazard-related use scenarios by observing the use of the medical device in the real environment. In this way, it is possible to gain knowledge about aspects difficult to be analyzed in an interview, such as natural use and unconscious behaviours.

One-on-one interviews: they consist of interviews conducted by a researcher with a user of the medical device. They can be more or less structured but typically a question-answer model is followed in a conversational way, in order to give the respondent the opportunity to provide also open feedback. They are used to identify typical use scenarios, issues, opinions, and attitudes, and to answer specific design questions. Through this technique, it is also possible to identify features that distinguish different user groups.

PCA analysis: for this analysis, tasks are decomposed into different user interactions, and for each user interaction, the relative user perceptions (P), cognitive steps (C), and actions (A) are investigated. In this way, the manufacturer can derive the requirements of each task in terms of perception (e.g., hearing an alarm), cognitive load (e.g., recalling information), and physical load (e.g., pressing a button). A use error is likely to occur if the user is not able to perceive a signal, interpret information, or perform an action. Therefore, for every step composing a task, three "what if" questions are asked: "what if the user is unable to perceive x?", "what if the user is unable to interpret y?", "what if the user is unable to perform the action z?". The answer to these questions allows to draw up a list of potential use problems.

Simulations: they consist of a more or less faithful reproduction of the use environment, in order to support realistic testing of the use of the device. It is required that the experimental setting allows users to naturally interact with the product as well as with human actors representing the other operators usually present in the use environment. Conducting simulations is always very expensive.

Standard reviews: they are conducted by usability specialists to assess a user interface according to established usability engineering practices. Some topics of investigation can be the physical aspect of the product, the control requirements, the display characteristics, and the alarm signals.

Surveys: administering questionnaires is a quick and cost-effective way to collect users' opinions. They can be more or less structured in order to give or not the respondent the possibility to expansively respond to open-ended questions.

Task analysis: it is used to study the interaction of the user with the device, in order to understand which factors could facilitate or hinder user performance. Each task is decomposed into several steps, called operations, that are analyzed individually to know how well the users are able to perform them. The operations that can lead to use errors are further analyzed. This approach allows the manufacturer to develop an appropriate user interface, it provides information for the analysis of use-related risks, and can also contribute to the definition of use scenarios.

Usability test: usability tests involve users from a specific group who are asked to complete a set of tasks, typically involving fundamental medical device functions. They are conducted by one or more moderators, i.e., people experts in the system who have to manage the execution of the test and take note of users' behaviours, actions, difficulties, and performances. The sessions can also be recorded for further subsequent analysis. Often, usability tests are carried out in controlled conditions and environments that affect users interaction and can be conducted on prototypes with different degrees of fidelity with respect to the final product. It is also possible to carry out usability tests on similar devices already on the market, to investigate their strengths and weaknesses. During the test, the moderator has to observe and analyze users' behaviour and report where some difficulties have been encountered. "Think aloud" can be an important source of information for medical device design but it should be used at the correct moment. When the goal is to get insights about design strengths and opportunities for improvement, i.e., during formative evaluations, it is worth to be used, while, when the goal is to have participants interact as naturally as possible with the device, i.e., during summative evaluation, it should not be used. Thinking aloud can in fact alter participants' performance by requiring them to behave differently than they would when using the device in the real world.

Focusing on the tool constituted by the surveys, some validated and open-sourced instruments for usability evaluation are herein presented:

Surgery Task Load Index:

The Surgery Task Load Index (S-TLX) is a validated tool to assess the perceived cognitive load specifically inside the surgical setting. It has been developed starting from a more general index, the NASA TLX, which has been demonstrated to be reliable in measuring individual workload among general populations. The goal of S-TLX is to evaluate the impact of different sources of stress on the cognitive workload of healthcare operators, in order to understand the underlying mechanism and to provide targeted interventions to tackle these issues.

Several tools have been used in the past to evaluate surgery-related stress but all of them provided a unidimensional measure, limiting the possibility to take effective actions to solve the problem. Workload is instead a multidimensional concept, influenced by different factors such as task demand, circumstances under which the task is performed, skills, behaviours, and perceptions of the individual performing it. Multidimensional tools offer stronger diagnosticity since they are capable of discriminating between different causes of workload but, at the same time, they lack generalizability to different application environments due to the fact that they may not reflect the particular sources of stress characterizing different contexts. Therefore, when the tool is applied to a context different from the one for which it has been specifically created, it results in the computation of an aggregated workload measure, losing the advantage of using a multidimensional scale. From these considerations, it is clear the need for deriving context-specific versions of the NASA-TLX in order to provide diagnostic information on the impact of various sources of stress on the demands perceived in specific frameworks.

The different dimensions considered in the Surgery-TLX were derived partially from the general NASA-TLX and partially from another TLX variant designed for car driving (Driving Activity Load Index - DALI). From the first one, the dimensions referred to the task demand were retained (mental, physical, and temporal demand) while from the DALI the dimensions concerning the environmental demand (distractions and situational stress). A further dimension related to task complexity was added to substitute the one indicated as "frustration" in the original version of the questionnaire. Thus, the 6 dimensions of workload considered in the S-TLX questionnaire resulted in:

- mental demand (how mentally fatiguing was the procedure);
- physical demand (how physically fatiguing was the procedure);
- temporal demand (how hurried or rushed was the procedure);
- task complexity (how complex was the procedure);

- situational stress (what level of anxiety was perceived while performing the procedure);
- distractions (how distracting was the operating environment).

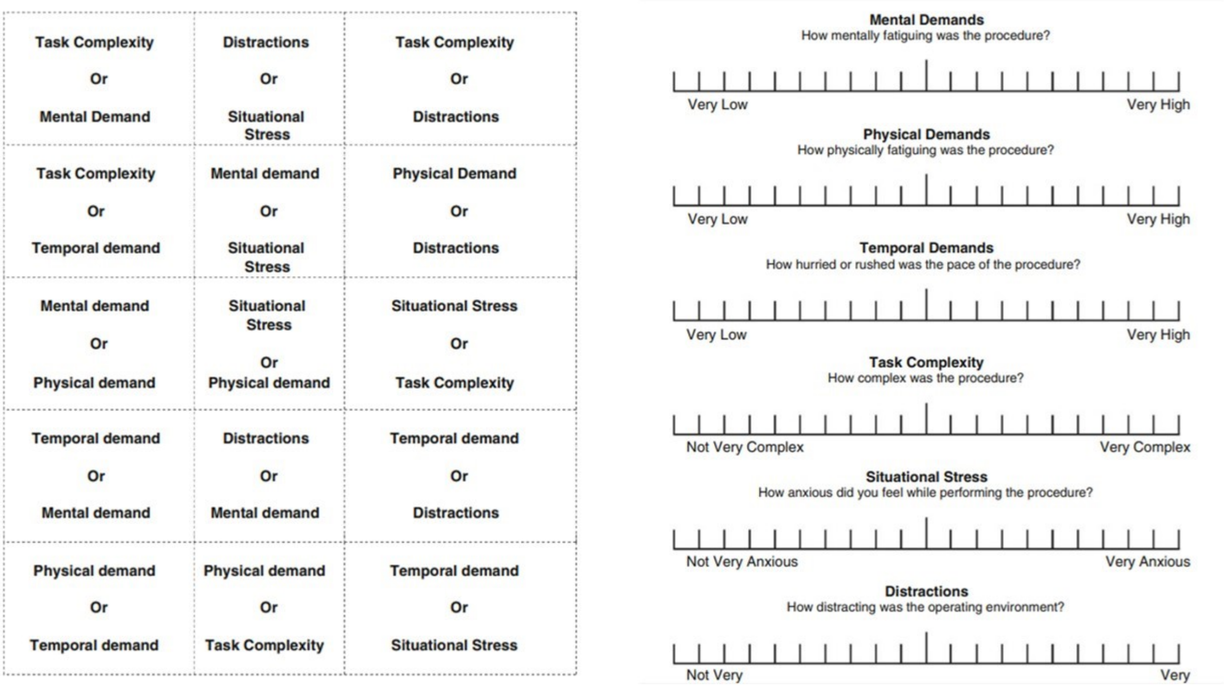
The S-TLX questionnaire is composed of two separate parts (Figure 1.14): the first is dedicated to the computation of the sources of load (i.e., the weights) and the second to the computation of the magnitudes of load (i.e., the ratings).

The weights are computed through a technique referred to as pairwise comparison. It consists of the combination of the 6 workload sources two by two, which results in a total of 15 couples. For each of them, the user has to circle the member of each pair that contributed more to the workload he perceived during that task. Through the tally of the number of times each factor is selected, it is possible to compute its weight, which represents, for each rater, the contribution of each factor to the workload experienced during a specific task. The weight of each factor can range from 0 to 5. In this way, instead of an a priori definition of workload, it is possible to take into account the definition each subject has in relation to a specific task and to weigh the different contributions accordingly.

The rating instead, expresses the magnitude of the different sources of workload in a given task. They are expressed through evaluation scales, represented by lines delimited by two bipolar descriptors, and divided into 20 equal intervals (0 = very low magnitude, 20 = very high magnitude). The user has to put a cross on the level of the scale indicating the level of perception of that source of workload. From the multiplication of the weight times the rating related to each workload source, it is possible to obtain the weighted rating of that stressor.

This tool does not only investigate the extent of the difficulties perceived during a procedure but also contributes to the assessment of the reason why they have been encountered and to the definition of appropriate training and stress management interventions, taking into consideration the subjective definition of workload in a specific task.

Even if the goal of this multidimensional tool is to provide an evaluation of the impact of the different factors on the demands perceived during the task, an overall workload score can be computed as well by summing up all the weighted ratings and dividing the result by 15. Of course, the higher the total S-TLX score, the higher the workload and stress perceived by the individual during the performance of the task. [1] [38]



Task Complexity Or Mental Demand	Distractions Or Situational Stress	Task Complexity Or Distractions
Task Complexity Or Temporal demand	Mental demand Or Situational Stress	Physical Demand Or Distractions
Mental demand Or Physical demand	Situational Stress Or Physical demand	Situational Stress Or Task Complexity
Temporal demand Or Mental demand	Distractions Or Mental demand	Temporal demand Or Distractions
Physical demand Or Temporal demand	Physical demand Or Task Complexity	Temporal demand Or Situational Stress

Mental Demands
How mentally fatiguing was the procedure?
Very Low — Very High

Physical Demands
How physically fatiguing was the procedure?
Very Low — Very High

Temporal Demands
How hurried or rushed was the pace of the procedure?
Very Low — Very High

Task Complexity
How complex was the procedure?
Not Very Complex — Very Complex

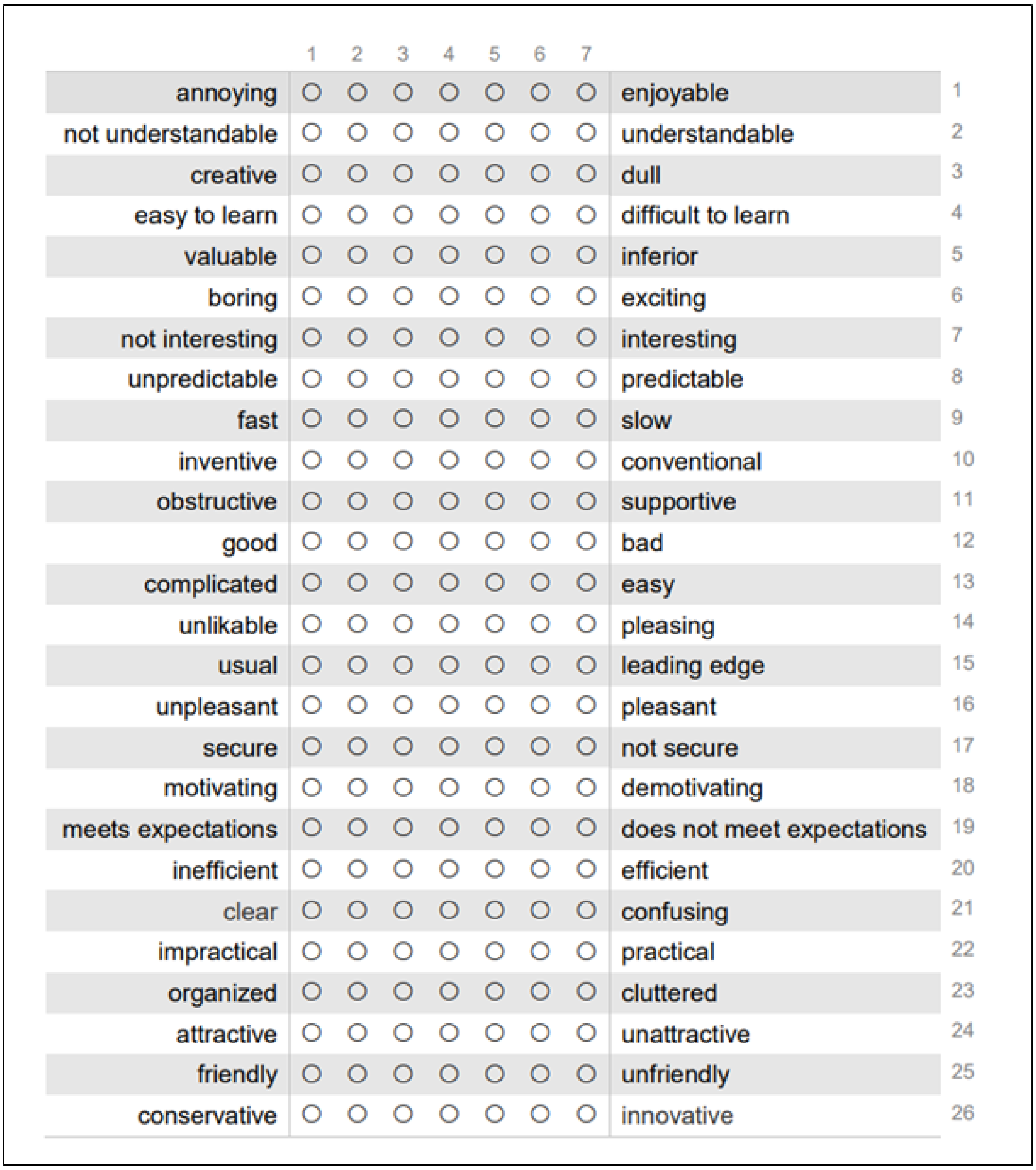
Situational Stress
How anxious did you feel while performing the procedure?
Not Very Anxious — Very Anxious

Distractions
How distracting was the operating environment?
Not Very — Very

Figure 1.14: *Surgery Task Load Index (S-TLX) questionnaire template: pairwise comparison for the computation of the weights (left), ratings for the computation of the magnitude of the sources of load (right).*

User Experience Questionnaire:

The User Experience Questionnaire (UEQ) is a validated tool designed to provide a quantitative measure of user experience and usability of interactive products. The items which form the questionnaire have been defined starting from a pool of 229 elements, among which only 80 were retained after expert evaluation. From these, the final 26 items were extracted by principal component analysis. Each of them has the form of a semantic differential, meaning that it is represented by two opposite terms divided by a 7-point scale (Figure 1.15).



	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

Figure 1.15: *User Experience Questionnaire (UEQ) template.*

Participants have to indicate the term that better describes the product under evaluation by placing an "X" on the point scale. The closer the symbol is to one of the two words, the better that term applies to the product in the respondent's opinion. In order to avoid bias, items starting with the positive term are alternated with items starting with the negative one. The 26 items can be grouped into 6 scales:

- Attractiveness: Overall impression of the product. Do users like or dislike the product?

- Perspicuity: Is it easy to get familiar with the product? Is it easy to learn how to use the product?
- Efficiency: Can users solve their tasks without unnecessary effort?
- Dependability: Does the user feel in control of the interaction?
- Stimulation: Is it exciting and motivating to use the product?
- Novelty: Is the product innovative and creative? Does the product catch the interest of users?

While Attractiveness is a pure valence dimension, Perspicuity, Efficiency and Dependability are considered goal-directed dimensions ("pragmatic quality aspects"), and Stimulation and Novelty are non-goal-directed dimensions ("hedonic quality aspects").

The items composing the different scales are shown in Figure 1.16. [32]

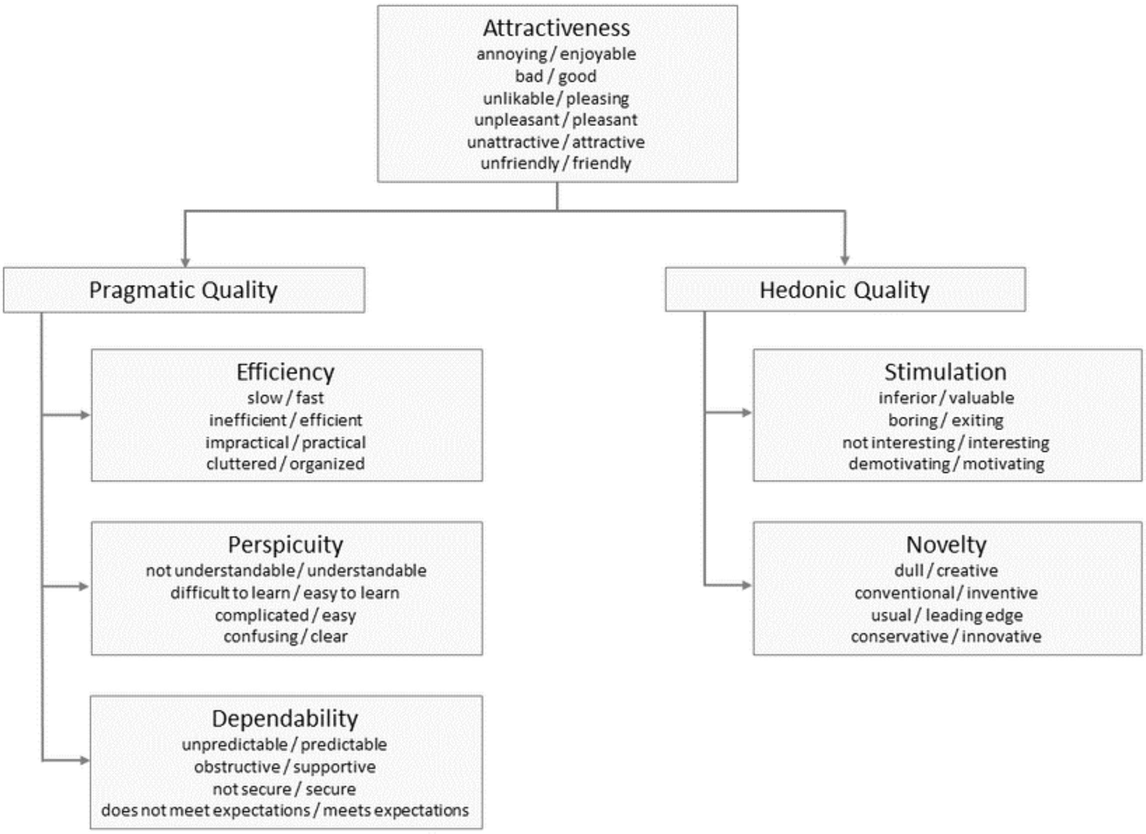


Figure 1.16: *The 26 items composing the UEQ, subdivided into the 6 scales.*

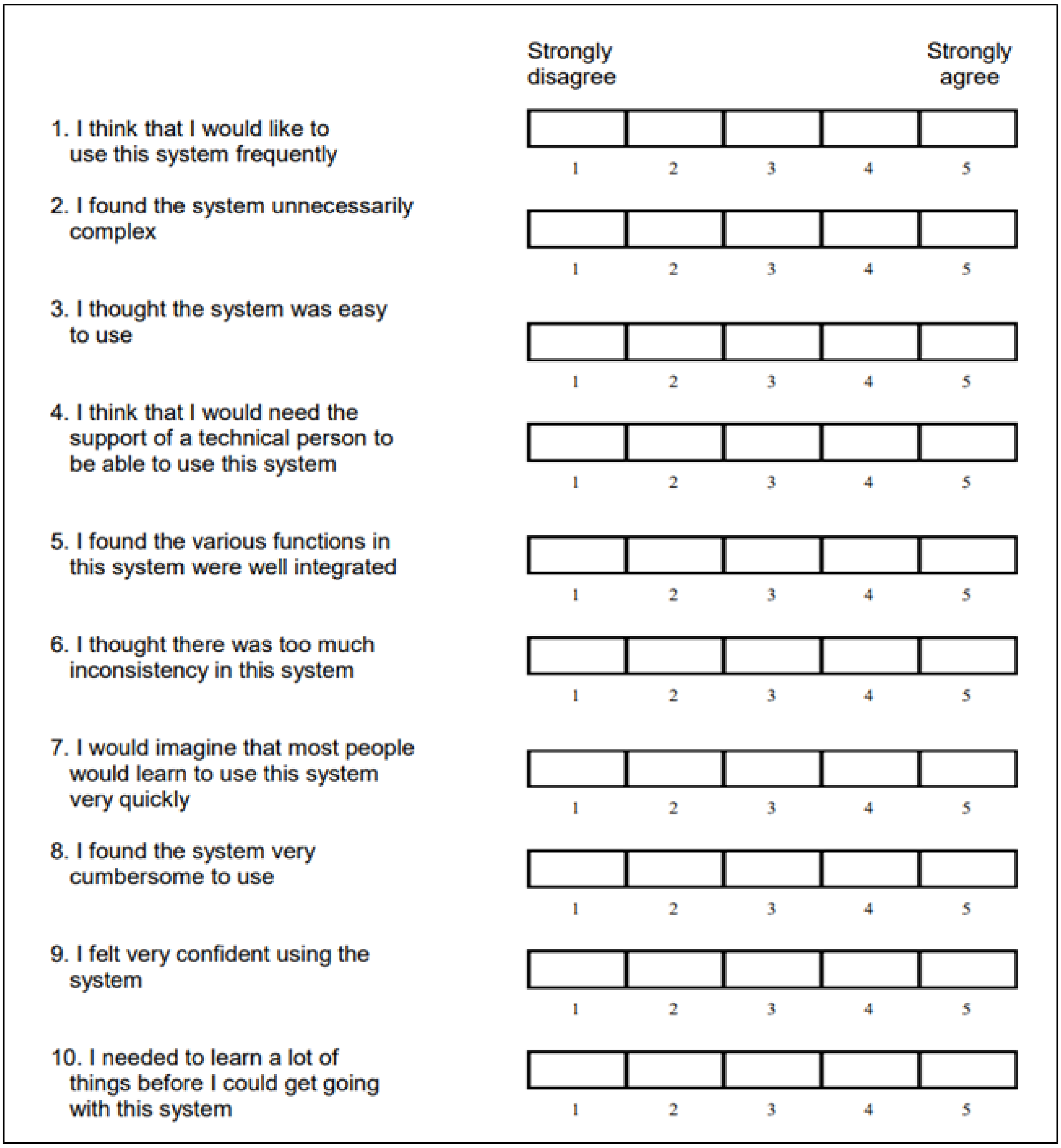
System Usability Scale:

The System Usability Scale (SUS) is a simple, reliable, and robust tool for the subjective assessment of usability, composed of 10 validated items. The user has to indicate his level of agreement with each sentence on a 5-point Likert scale ranging from "strongly agree" to "strongly disagree" (Figure 1.17).

The questionnaire has been developed starting from a pool of 50 potential items, which have been used by a group of 20 people to evaluate two software systems, one considered "very easy to be used", i.e., very usable, and the other "almost impossible to be used", i.e., unusable. Among these 50, the items leading to the most extreme responses were selected. The selected items also showed high intercorrelation. In order to avoid response biases, they were characterized by the fact that the common response to half of them was strong agreement and to the other half was strong disagreement. Moreover, the positive and negative items are alternated in the questionnaire: in this way, the respondent has to read each question and carefully think about his level of agreement or disagreement with the statement.

The selected statements cover different aspects of a system's usability, such as complexity, need for training, confidence, and willingness to use it, therefore, allowing for capturing users' satisfaction in using the system. This scale can be defined as "quick and dirty" since it allows for obtaining the subjective perception of the usability of a system in a simple way and in a short time.

The final score is computed considering that the expression of agreement with a positive sentence is equivalent to the expression of disagreement with a negative one. For the positive sentences, the contribution is computed as (scale position - 1) while for the negative ones as (5 - the scale position). In this way, the contribution of each item's score ranges from 0 to 4. The scores are then summed up and the result is multiplied by 2.5 in order to obtain a very intuitive value for the SUS score, ranging from 0 to 100. The higher the score, the greater the usability perceived by the user with respect to the product under evaluation. The drawback of having the results on a 0-100 scale is that there is a tendency to perceive them as percentages though they are not. [11]



	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	□ 1	□ 2	□ 3	□ 4	□ 5
2. I found the system unnecessarily complex	□ 1	□ 2	□ 3	□ 4	□ 5
3. I thought the system was easy to use	□ 1	□ 2	□ 3	□ 4	□ 5
4. I think that I would need the support of a technical person to be able to use this system	□ 1	□ 2	□ 3	□ 4	□ 5
5. I found the various functions in this system were well integrated	□ 1	□ 2	□ 3	□ 4	□ 5
6. I thought there was too much inconsistency in this system	□ 1	□ 2	□ 3	□ 4	□ 5
7. I would imagine that most people would learn to use this system very quickly	□ 1	□ 2	□ 3	□ 4	□ 5
8. I found the system very cumbersome to use	□ 1	□ 2	□ 3	□ 4	□ 5
9. I felt very confident using the system	□ 1	□ 2	□ 3	□ 4	□ 5
10. I needed to learn a lot of things before I could get going with this system	□ 1	□ 2	□ 3	□ 4	□ 5

Figure 1.17: *System Usability Scale (SUS) questionnaire template.*

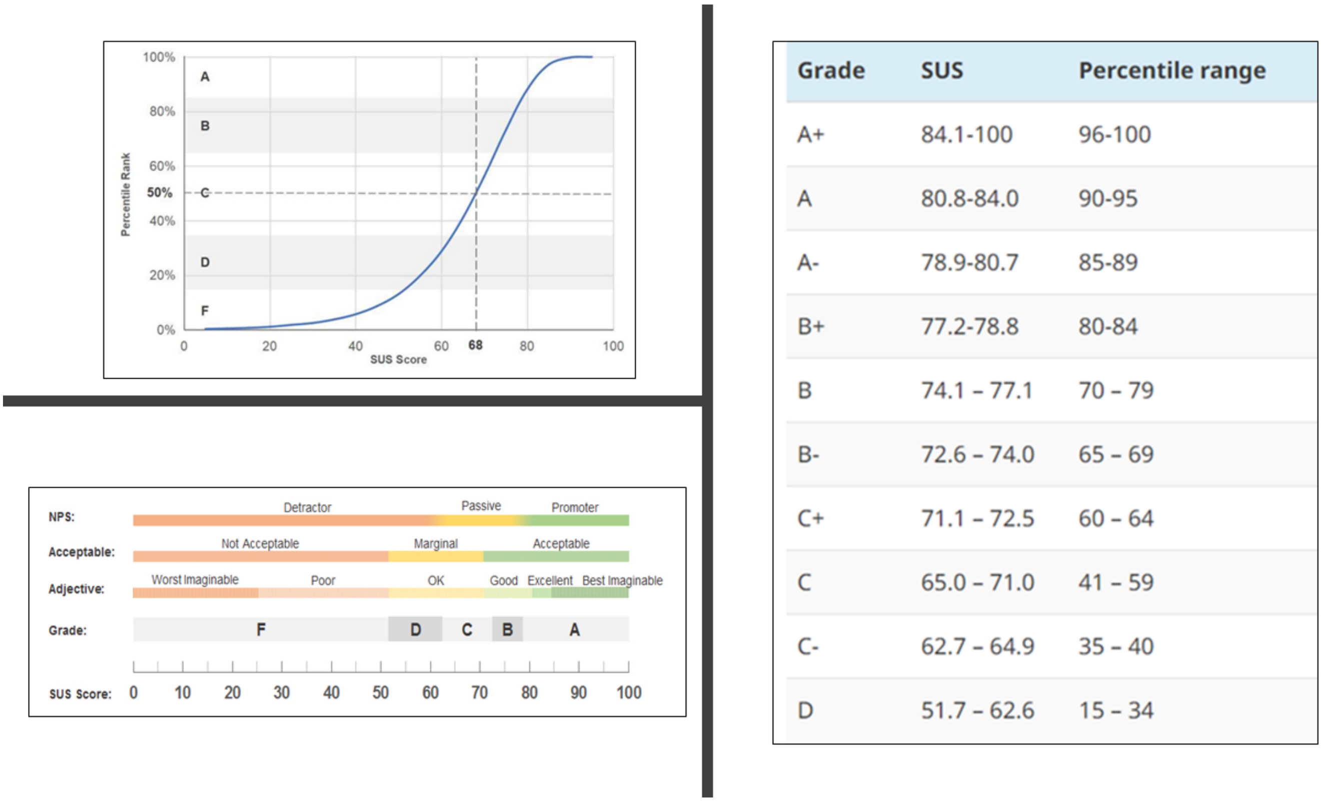
Even if the methodology is simple, the interpretation of the score obtained may be problematic, particularly if no comparators are included in the analysis. For example, it may be difficult to understand if an intermediate score close to 50 is average or good, or if a high score close to 90 is realistic or not. For this reason, several methods to interpret SUS scores are proposed in the literature.

Bangor et al. [8] have associated 1'000 SUS scores with a 7-point adjective scale, demonstrating a close correlation between SUS scores and qualitative adjectives such as “poor”, “good”, and “excellent”.

Sauro et al. [30] investigated the relationship between SUS scores and Likelihood to Recommend Scores (LTR). The LTR distinguishes between three classes of recommenders based on the response to a single question: “How likely is it that you would recommend our company/product to a friend or colleague?”. The respondent can rate his likelihood from 0 (not likely at all) to 10 (extremely likely). People scoring between 9 and 10 are considered “promoters”, meaning they are likely to recommend the product to a friend, the ones scoring between 7 and 8 are “passives”, i.e., neutral, and the ones scoring 6 or below are “detractors”, meaning they are likely to discourage. The research group collected SUS and LTR data from 2’200 respondents and computed the regression equation binding these values, which resulted in the following expression: LTR = 1.33 + 0.08*(SUS). Anyway, the interpretation of the SUS score with the promoter/detractor approach is ambiguous since in the literature there are contrasting opinions about the precise usability scores which distinguish between the three classes of recommenders.

Another interpretation approach consists in converting the scores into percentile ranks. This approach has been proposed again by Sauro et al. [31] and it allows for comparing the result of the product under evaluation with a large dataset of SUS scores. The dataset is composed of data from over 10’000 responses and hundreds of products, collected over more than 30 years of usage of the questionnaire. The 50° percentile corresponds to a score of 68, meaning that 50% of the products in the dataset score below 68 and 50% score above 68. Therefore, a score higher than 68 can be considered above the average.

The curve showing the percentile ranks of the SUS scores, together with a summarization of other approaches applicable in the interpretation of the scores is shown in Figure 1.18.



Grade	SUS	Percentile range
A+	84.1-100	96-100
A	80.8-84.0	90-95
A-	78.9-80.7	85-89
B+	77.2-78.8	80-84
B	74.1 – 77.1	70 – 79
B-	72.6 – 74.0	65 – 69
C+	71.1 – 72.5	60 – 64
C	65.0 – 71.0	41 – 59
C-	62.7 – 64.9	35 – 40
D	51.7 – 62.6	15 – 34

Figure 1.18: *Methods to be used in the interpretation of SUS scores: percentile ranking (top left and right), promoters-detractors approach and association of an adjective (bottom left).*

Another advantage of SUS questionnaire is that it allows for obtaining reliable conclusions even with a small sample size: a high percentage of "correct" conclusions, i.e., a high level of consistency between responders, is reached with just 8-12 responders. This has been demonstrated by Tullis et al. [37] starting from 123 evaluations of the two reference software and then computing, for different sample sizes, the percentage of tests that reached the correct conclusion, i.e., the same conclusion obtained by analyzing the whole dataset (Figure 1.19).

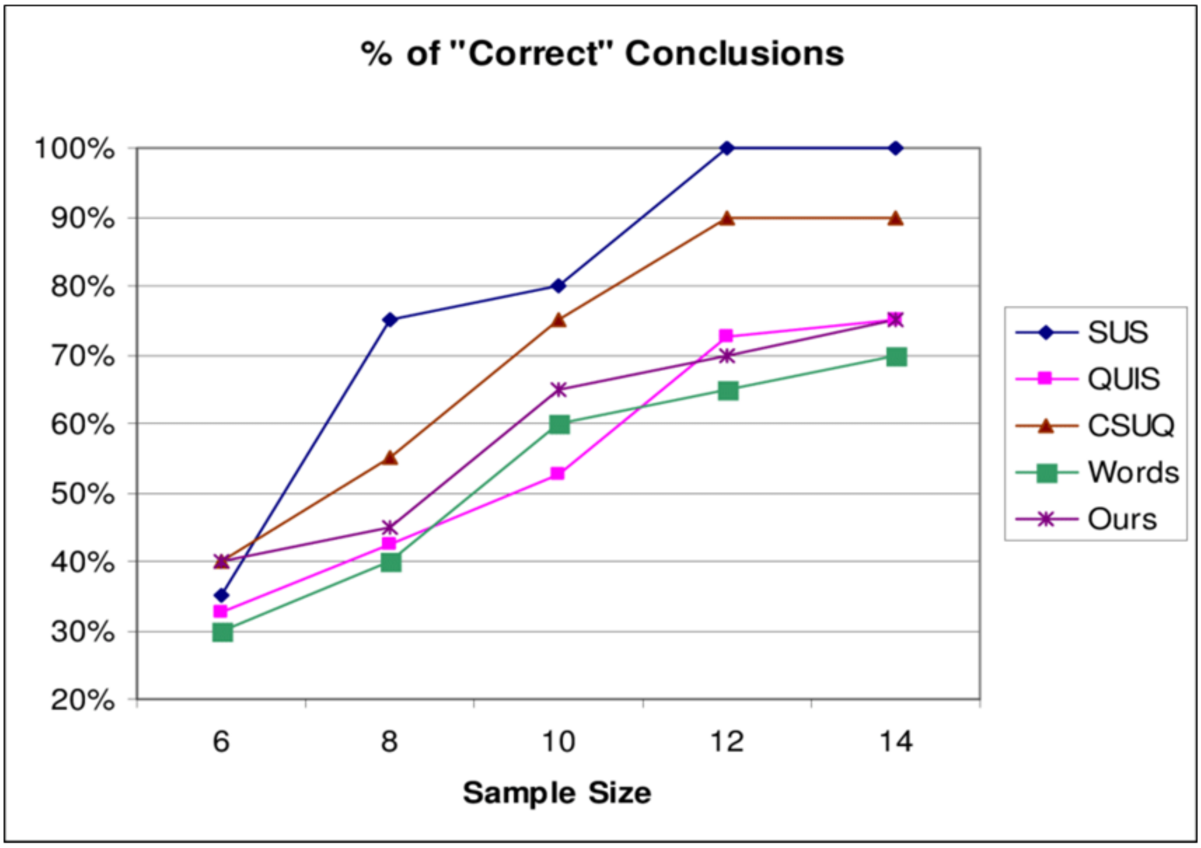


Figure 1.19: *Sample size necessary to reach consistency between respondents with different questionnaires.*

Other studies demonstrated that the SUS scores obtained after the user has been shortly exposed to the system (5 seconds) is very similar to the one obtained after they have used it for an extended period of time, demonstrating that usability is strongly affected by the first impression. [29]

Lastly, it is important to underline that the items of the SUS are general and do not relate to a specific feature of a particular system: as a consequence, the questionnaire is not diagnostic, since it does not provide information on why a system is usable or not, but it is a technology-neutral tool which can be used as technology evolves over time.

1.5. Aim of the work

This work aims to the summative evaluation of a module of the MR platform ARTICOR, namely the module dedicated to the navigation of pre-operative clinical images and corresponding 3D anatomies obtained via segmentation. This module of the MR platform has already been analyzed through formative evaluations ([18]), which identified some usability problems and drove the partial redesign of the user interface as well as the implementation of new tools within it. Moreover, a previous usability study was performed during the certification process of the whole ARTICOR platform. Nevertheless, due to the novelty of the technology, the number of usability studies on the MR platform is very limited and the range of aspects relevant to usability considered so far still needs to be extended to obtain a comprehensive evaluation. On this basis, this work aims to analyze usability-related aspects of the aforementioned ARTICOR module that were not considered before, through a standardized approach, i.e., through assessment methods among those described in this section, as well as ad hoc developed assessment tools. Moreover, another added value of this study is taking the perspective not only of MR platforms' end users and manufacturers, but also of those responsible for technology evaluation, with the goal of standardizing the assessment process of similar technologies through the definition of appropriate evaluation criteria.

2 | State of the art

2.1. Usability evaluation of XR applications

Focusing on the context of interest, i.e., extended reality in medical applications, the *usability engineering process* has been applied to develop and validate the design of different products. In the following section, some examples from the literature are presented. Some of them have the same purpose as my study, some others share the tools and methodologies exploited. Anyway, they describe some inspiring works performed in the context of interest to which my study aims to add something in terms of tools exploited or aspects evaluated.

As previously mentioned, a relevant application of MR is in the context of electrophysiology. The current workflow of radiofrequency ablation results in successful procedures, but there are still some limitations that increase complexity and medical effort: first, the electroanatomic maps are displayed on a 2D screen, requiring the electrophysiologist to mentally recreate the 3D image; second, many operators are needed for the procedure, each of them being in charge of controlling a different workstation, and the control of data is decentralized from the operator performing the procedure, thus generating potential communication issues. To solve these two problems, the already mentioned Enhanced Electrophysiology Visualization and Interaction System (ELVIS) has been developed.
In a study by Silva et al.[34], formative evaluations of the system have been conducted during the design and development of the platform, as required by IEC 62366-1. These tests provided feedback about the preferred method of interaction, menu legibility, and potential use errors. In addition, the final version of the platform was tested through an in-human study in which 3 physicians were asked to perform some tasks on electroanatomic mapping images of 16 patients using both a standard mapping system and ELVIS. The tasks consisted of: 1) creation of a single, high-density cardiac chamber; 2) sequential point navigation within the generated chamber. The number of interactions between the electrophysiologist and the mapping technicians was recorded and physicians were asked to answer 7 questions about the usability of the system. The results showed a significant reduction in personnel interaction in task 2 when performed with ELVIS, which may

improve efficiency and team dynamics. The questionnaires concerning usability reported positive scores about comfort, ease of use, tools accessibility, and improved capability in interpreting the information obtained.

Moving to another application field, the study conducted by Glas et al. [19] investigated the use of a MR visualization platform for image-guided surgery (IGS). They connected a surgical navigation system, i.e., Brainlab (Brainlab AG, Munich), to Microsoft HoloLens (Figure 2.1). The tool allows the operator to visualize and interact with the surgery plan during the procedure. In particular, beyond the evaluation of the accuracy in performing typical navigation tasks, the user-friendliness and usability of the system have been investigated by a formal user study that compared the MR platform with the gold standard setup for a perioperative navigation system.

The standard workflow in IGS consists of the use of preoperative images (CT, MRI) to generate a 3D virtual surgical plan, which is later registered with the patient through the navigation system itself. This registration is accomplished by mapping predetermined landmarks on the image to the patient's actual position. In this way, the operator can obtain real-time information about the relative position of surgical instruments and anatomical structures and can place virtual landmarks on the surgical field. Once again, the main limitation of this technology is the fact that information from the navigation system is provided on physical 2D screens, leading to ergonomic problems and increased mental burden for the physician in reconstructing the 3D features of the patient's anatomy. In turn, this can result in increased surgical time, deviation of surgeon attention, and a higher probability of errors. Moreover, the lack of direct control of the operator on the data does not allow for careful data exploration and increases workflow complexity.

The proposed solution gives the surgeon the possibility to visualize anatomical data in 3D maintaining depth perception, superimpose it onto the patient, and directly interact with the image through vocal commands or user-defined gestures. From this comparative study involving 12 participants, the performance and usability of the MR platform against a traditional navigation interface have been evaluated. The defined tasks consisted of the search for 3 physical landmarks and 3 trajectories on a human phantom. The time required to complete the pre-defined tasks and the accuracy in reaching a target landmark have been measured, and the results were in favour of the new technology. In particular, the overall completion time of all tasks with HoloLens was 1.71 times faster than with Brainlab alone ($p = 0.034$). The measurement of the accuracy in performing the tasks, calculated as the Euclidian distance between the final position of the instrument's tip and the target landmark, resulted in smaller deviations from the planned trajectories when using the HoloLens ($p < 0.001$). After having performed the tasks, participants were also asked to fill in a questionnaire to rate the usability in terms of ease of use, efficiency, intuitiveness of interactions, and real-time experience. Even if half of the participants were not familiar with MR, all of them reported that the tasks became easier to be performed (difficulty with Brainlab rated 3.25/5, with HoloLens 2.4/5), the interaction was reported as intuitive, and the real-time experience was rated above the average. In conclusion, the system was proven to enable a reduction of the workload, improvement of visual feedback, and enhanced eye-hand coordination.

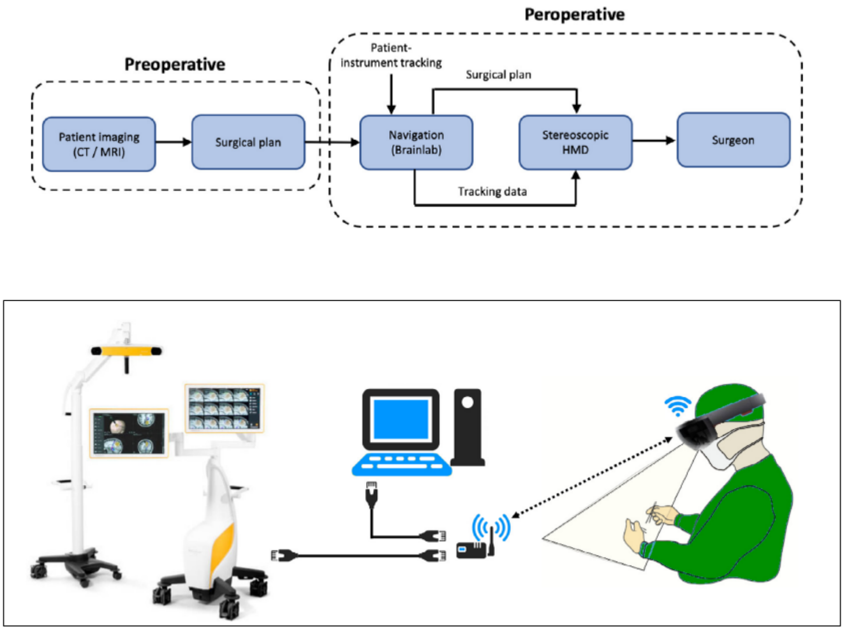


Figure 2.1: *Top: Schematization of the workflow for a MR platform for image-guided surgery: starting from preoperative patient imaging and patient-specific surgical plan, the surgical plan is then loaded on the navigation system (Brainlab) and eventually visualized on the HMD. Bottom: hardware components. From left to right: the conventional Brainlab navigation system, a PC running the application, and a surgeon wearing the HoloLens. All hardware communicates through a dedicated router.*

Considering the preoperative use of MR for diagnostic and morphological analysis, Brun H. et al [13] investigated the feasibility of 3D MR holograms. Cardiac computed tomography angiogram (CTA) images of a pediatric patient with double-outlet right ventricle and transposition of the great arteries [1] were segmented to create a 3D model of the heart. 36 members of the heart team, with different levels of expertise, visualized the hologram

[1] Double-outlet right ventricle and transposition of the great arteries is a rare congenital heart defect in which both the aorta and the pulmonary artery arteries arise from the right ventricle. The only outflow from the left ventricle is a ventricular septal defect, which diverts blood toward the right ventricle.

through the HMD Microsoft HoloLens. In the first part of the study, participants had to recognize some anatomical landmarks and perform a diagnosis of a specific heart defect by viewing the hologram. In the second part, users had to fill out a questionnaire concerning anatomy identification, diagnostic output, and 3D experience of the model. Also, they had to rate the quality of the interactions with the hologram (e.g., s moving, rotating, scaling, and slicing). The questionnaire was constituted by 1-6 rating scales. All the participants were able to identify the selected landmarks and all but two performed the correct diagnosis. The ratings were all high, in particular among female and younger users. The overall hologram experience resulted in mean scores from 5.32 to 5.46 for all variables. The quality of the interactions resulted in a lower rating but was still closer to the maximum than to a neutral score. This study demonstrated that MR models can have a significant diagnostic value when used as a surgical planning tool, in particular in the case of complex and abnormal anatomies.

Another system that deserves attention is RealView Holographic Display system [Realview Imaging Inc., Yokneam, Israel], a platform that creates and displays holographic models starting from 3D rotational angiography (3DRA) coupled with 3D transoesophageal echocardiography (3DTEE). In this case, virtual models are not generated through stereoscopy as it happens in most MR systems. They are instead generated by feeding specific algorithms with 3D data, which are transformed into interference patterns. Coherent light at a defined wavelength then passes through these interference patterns, which lead to image formation by inducing phase distribution. The particular feature of this system is that it does not require any human-mounted device or 2D display since the hologram is visualized through the Holoscope (Figure 2.2).
The feasibility of this system inside a catheterization lab has been investigated in a study [12] involving 8 patients and 4 specialists, with two aims: i) demonstrating that all the anatomical landmarks identified on standard imaging can be similarly identified using dynamic and static holographic images, ii) demonstrating the usability of interactions with the hologram (marking, cropping, zooming, rotating, slicing, and moving). 4 specialists in cardiology were involved in the study. Each of them had to identify specific landmarks and rate the image quality with respect to the standard display on a 1-5 rating scale (where a rate of 5 meant "as good as" the standard display). The usability of the interactions was investigated using again a 1-5 rating scale, where 1 meant "very hard interaction" and 5 meant "very easy interaction". All the chosen anatomical landmarks were identified both through the holographic display and through conventional 3DTEE and 3DRA with the same level of difficulty. Usability was rated with the maximum score since all 4 specialists were able to perform the above-mentioned imaging interactions "very easily".

Figure 2.2: *Holographic reconstruction of a heart obtained through the RealView Holographic Display system and displayed under the Holoscope.*

From the methodological standpoint, the usability study by Sternini et al. [36] is particularly interesting. The authors investigated the usability of a medical device software intended to assist intraoperative planning through the visualization of a 3D reconstruction of patient's anatomy. Starting from 2D medical images of the patient, a 3D model is reconstructed and visualized on a screen. The user can zoom, move, and rotate the model, annotate it and hide some elements. In this case, every action is performed through a touchless user interface based on Leap Motion sensors (Figure 2.3), which are able to detect and track the hands of the user and visualize them as a virtual model on the same screen where the model is displayed.

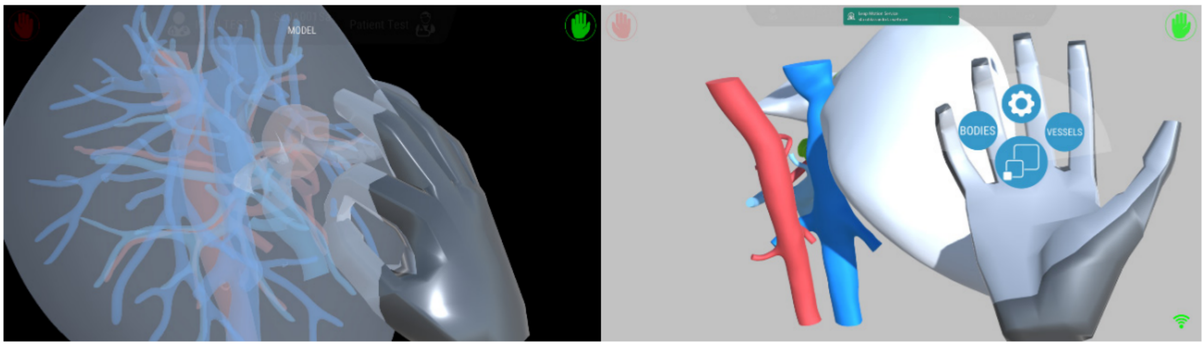


Figure 2.3: *Device interaction through Leap Motion sensors. Left: rotation of the model. Right: menu opening.*

The usability of the platform has been assessed starting from a formative evaluation divided in turn into two phases. The first one was desk-based: designers and usability experts used a quick and dirty approach such as brainstorming, cognitive walkthrough, fault tree analysis (FTA), and standard review [2], to define the primary operating functions and the position of the sensor and of the screen to guarantee ergonomics for the user. The second phase has been conducted through focus groups with real users to confirm the outputs of the previous stage and to identify possible additional issues. The focus groups were organized as a first training session, whose contents have been developed based on the outputs of the desk-based stage, followed by sessions in which the users had to perform specific tasks. These led to the identification of some issues with functions of the device such as zooming and rotating the model, and with aspects of the user interface such as position and visibility of the menu. The users were then asked to provide an evaluation of the primary operating functions and to report usability issues by compiling a questionnaire. Subsequently, they were invited to join a discussion with designers and usability experts. The formative evaluation led to modifications of the position of the sensor and of the screen and to the insertion of a tutorial section (Figure 2.4).

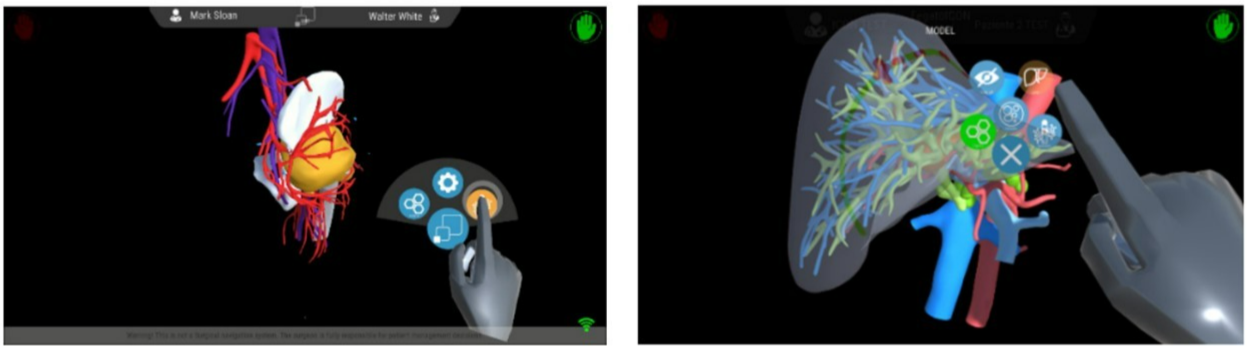


Figure 2.4: *Menu visualization of the device version used during the formative (left) and summative (right) evaluation.*

The subsequent step has been the summative evaluation, intended to confirm the usability of the final version of the device. In this phase, the users were involved in simulations

[2] Cognitive walkthrough: researchers try to determine what is expected by the user by walking through a preliminary design, completing the tasks the user will complete, and leading through these tasks both experts in the matter and representative users. In this way, researchers investigate whether users understand what they have to do for each task and when a correct or incorrect use has been done.
FTA: "top-down" approach that focuses on one failure event (defined by its frequency of occurrence, the severity of the harm that can result from it, and the effectiveness of the risk control measures applied) and tries to determine its causes (i.e., failure modes) in successive levels of detail.
Standard reviews: usability specialists assessment of a user interface according to established usability engineering practices.

of the real use in a setting representative of an OR, with a phantom simulating a patient undergoing a laparoscopic intervention. The summative evaluation was articulated in a training phase, where the users were explained how the device works and how to interact with it, followed by task analysis, during which the performance of the users in each task has been recorded and classified as correct, use error, technical error, or critical error. The results from this phase showed a decreasing rate of use errors from the first to the last task, suggesting an improvement in the user performance during the use of the device and therefore a steep learning curve. Further analysis was carried out in the form of heuristic evaluation: a questionnaire was designed to identify the elements that violate some usability heuristics in order to detect usability problems in the user interface. The results showed user difficulties in the management of the model, particularly when the user needed to modify the visualization status. This was explained considering that these actions required very precise interaction while, in the beginning, it was difficult for the user to have fine control of the movement of the hand in the virtual space. Other questionnaires were administered to the user in order to redefine the primary operating functions, investigate the risk of the device, and describe the overall usability of the user interface. The last one was the User Experience Questionnaire (UEQ) (see subsection 1.4.2), a validated tool designed to evaluate different aspects of the user interface: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. The high score obtained in all these aspects showed that the device was considered very good by the users. Additional questionnaires were provided to the user to evaluate the usability of a different visualization system, based on stereoscopic displays and a virtual reality visor. No statistical differences have been identified by comparing the scores of the UEQ of the general and stereoscopic visualization. The greatest difference has been found in the evaluation of perspicuity, which describes how easy it is for the user to learn how to use the device, meaning that the tridimensionality provided by the stereoscopic visualization may help the user to perceive his hands' position in the virtual environment.

An example of the application of another validated questionnaire is provided in the already mentioned paper by S. Moosburner et al. [24] (see subsection 1.2.2), whose study aims at comparing the usability of two AR head-mounted displays, namely Microsoft HoloLens and Meta 2. The questionnaire used in this study is a modified version of the System Usability Scale (SUS) where, to the 10 original items about usability, 5 more questions about ergonomics, uncomfortable sensations, visual clarity, field of view, and gesture control have been added. 15 medical students were asked to fill in the questionnaire after having interacted with both devices and visualized a 3D model of a liver created from a CT scan. From the analysis of the scores, ergonomics, ease of use and visual clarity did not report statistical differences; the smaller field of view of Microsoft Hololens led a significant number of users to feel limited by it but its superior object stability and the improved mobility due to its wireless and stand-alone functioning, which does not require cables and additional hardware, resulted in higher evaluation score. Therefore, the study has highlighted the overall superiority of Microsoft HoloLens, a usable device in surgical settings.

Again, Long Qian et al. [27] compared the usability of three optical see-through head-mounted displays (OST-HMDs) to enable MR experiences during surgical procedures. The evaluated criteria included text readability, contrast perception, task load, frame rate, and system lag. The three devices under investigation were Microsoft HoloLens, ODG R-7, and Epson Moverio BT-200 (Figure 2.5).

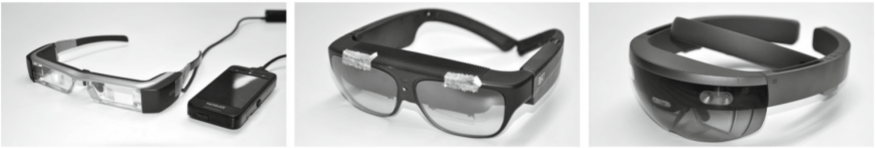
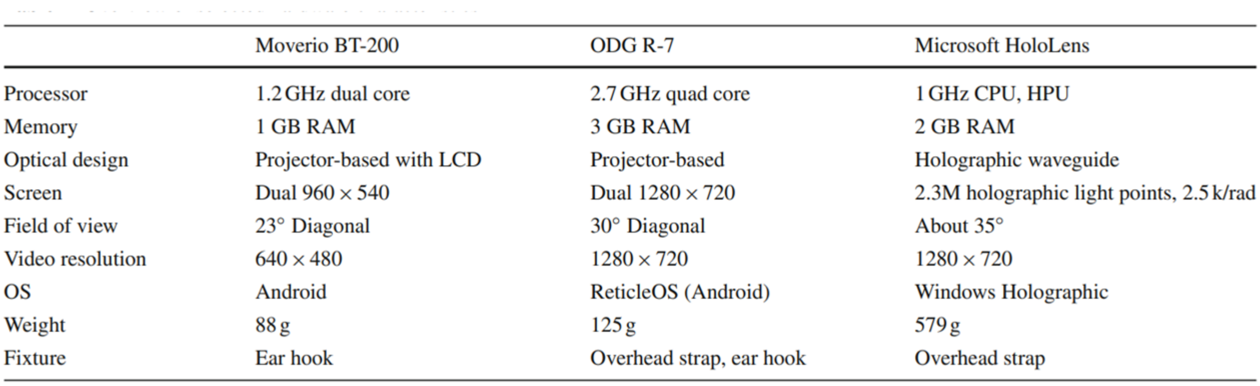


Figure 2.5: *Three optical see-through head-mounted displays compared in the study by Long Qian et al. [27]: Epson Moverio BT-200 (left), ODG R-7 (middle), Microsoft HoloLens (right).*

The display technology of Epson Moverio BT-200 is based on a binocular LCD projector, and it is very lightweight (88 g) and affordable for non-professional users. ODG R-7 has binocular projector-based optics, with a higher refresh rate (80 Hz) than the BT-200 (60 Hz). It is suitable for professional use due to its processing power, and it weighs 125 g. Microsoft HoloLens optical design is based on holographic waveguides, it has the largest FOV and weighs 579 g. A summary of the hardware characteristics of the three devices is reported in Figure 2.6.

The clinical scenario of this study is the use of OST-HMDs for object-anchored 2D-display: in this case, the visor allows for the visualization of a virtual 2D monitor which, instead of being anchored to the head of the operator, is anchored to a fixed object. Therefore, the monitor can be placed close to the surgical site without risking that the superimposition of the screen on the surgical site affects the clear view of both. With head-anchored 2D displays instead, unexpected movements from the operator can move the hologram into uncomfortable or even dangerous positions. This scenario, differently from HMDs which



	Moverio BT-200	ODG R-7	Microsoft HoloLens
Processor	1.2 GHz dual core	2.7 GHz quad core	1 GHz CPU, HPU
Memory	1 GB RAM	3 GB RAM	2 GB RAM
Optical design	Projector-based with LCD	Projector-based	Holographic waveguide
Screen	Dual 960 × 540	Dual 1280 × 720	2.3M holographic light points, 2.5 k/rad
Field of view	23° Diagonal	30° Diagonal	About 35°
Video resolution	640 × 480	1280 × 720	1280 × 720
OS	Android	ReticleOS (Android)	Windows Holographic
Weight	88 g	125 g	579 g
Fixture	Ear hook	Overhead strap, ear hook	Overhead strap

Figure 2.6: *Summary of the hardware characteristics of the three OST-HMDs compared in the study.*

display 3D objects, is easily implementable both from a technical and organizational point of view and can provide ergonomic benefits in image-guided surgery. The comparative study included a multi-user study for the evaluation of subjective criteria (text readability, contrast perception, task load) and an offline experiment for the evaluation of the system's performance (frame rate and system lag). Microsoft HoloLens outperformed the other two devices for all the evaluation criteria. In particular, the task workload has been measured through the NASA-Task Load Index (NASA TLX) (see subsection 1.4.2), a validated questionnaire that takes into account mental demand, physical demand, temporal demand, performance, effort, and frustration perceived during the task. The significantly lower score of HoloLens shows that its heavier weight is very well tolerated thanks to the adjustable design and that eye fatigue due to the vergence-accommodation conflict is reduced by its multiscopic display design.

3 | Materials and methods

The whole study was carried out in collaboration with Artiness, the already mentioned start-up company which develops ARTICOR (section 1.3), the MR platform under evaluation, and with hospital IRCCS Fondazione Cà Granda Ospedale Maggiore Policlinico di Milano (OMPM). In particular, two operative units (OU) of the hospital were involved: the clinical engineering unit and the cardiac surgery unit, which was recently integrated into the hospital. The clinical engineering unit is composed of experts in technology evaluation and regulation; hence, it played a key role in defining the methodology to be applied in this usability study and in providing the normative support to sustain it. The cardiac surgery unit is one of the divisions that could benefit the most from the technology being assessed; together with cardiac surgeons, also cardiologists, hemodynamic cardiologists, and trainees were involved in the study as potential users of the technology.

In particular, the work focused on the pre-procedural planning software ARTICOR, namely on the evaluation of its usability. In conducting the tests, I have not exploited the certified version of the software but a very similar one currently used by the developing team to perform corrective actions and implement further functionalities. Moreover, I have focused on a specific module of the R&D version of ARTICOR, that is the one dedicated to medical imaging navigation, which allows obtaining specific image planes or identifying specific anatomical structures on the 3D holographic model. Thus, I have not considered the modules dedicated to 3D model generation, image retrieval and storage, data security, and data transmission.

3.1. Usability evaluation

Criteria for the evaluation test design

In order to define the test protocol, the opinion of users and experts in the technology under evaluation was collected through interviews: the experts were the CTO and a scientific advisor of the start-up company developing the ARTICOR platform; the user was a cardiac surgeon from OMPM.

The usability test was focused on end-user interaction with medical images and 3D anatomical reconstructions through the ARTICOR platform in the pre-procedural phase and was designed as comparative. These features drove the choice of the comparator and hence the definition of the test protocol. The selected comparator consisted of a standard DICOM viewer software used for pre-operatory planification since it was characterized, with respect to the tested MR technology, by opposite features in terms of rendering (on a physical 2D screen as opposed to fully 3D holograms) and user interaction (through mouse and keyboard instead of direct model manipulation), while allowing for similar operations relevant to pre-operative planning (navigating volumetric data, defining cut-planes, cropping, etc). Also, the choice of the gold standard technology for pre-operative planning as the comparator allowed for the investigation of how the level of confidence in using it could influence the acceptance of the new technology.
In particular, the RadiAnt DICOM image viewer was chosen (`https://www.radiantviewer.com/`): it runs on almost any type of Windows PC even if it does not have any medical certification. As a consequence, it is not intended to be used for diagnosis, but it is typically exploited by students and residents to study medical images. Beyond the basic tools (zoom, negative mode, rotation, flip, image filters), it also allows for the measurement of lengths, areas, perimeters, and angles of the regions of interest. The software does not offer storage space, but it offers several additional features such as 3D multiplanar reconstruction (MPR), maximum and minimum intensity projections (MIP), and image fusion, along with the possibility of exporting images to JPEG, PNG, and other formats. Since the tasks performed in this study were not meant to provide real diagnosis nor surgical planning, the lack of medical certification was not a problem. Furthermore, the 3D multiplanar reconstruction was a fundamental functionality because it allowed for a comparison vs. the direct 3D visualization offered by ARTICOR.

Subsequently, the following research questions and endpoints were identified:

1. Report of participants' comments and opinions, and descriptions of the observed use errors and difficulties encountered during the test;

2. Comparative evaluation of the usability of ARTICOR with respect to RadiAnt DICOM viewer;

3. Differences in the time required to fulfill some tasks with the two technologies;

4. Influence of the level of experience with MR on usability evaluation and time performances for both the technologies;

5. Influence of the level of experience in using DICOM viewer software on usability

evaluation and time performances for both the technologies;

6. Absolute evaluation of the usability of the ARTICOR platform.

Enrolled participants

In order to answer these questions, 16 users were enrolled. Two inclusion criteria were applied when selecting the participants: i) professional background in cardiovascular anatomy since the models developed by Artiness find application in this field; ii) general confidence in the management of CT and echocardiographic images. Moreover, users with different levels of experience with DICOM viewer software were enrolled in order to investigate the influence of these factors on the performance and perception of the users with respect to the technology, thus answering research question **5)**. No exclusion criteria were considered. As a result, 6 cardiology residents, 1 vascular resident, 1 emergency medicine resident, 1 medical student, 1 interventional cardiologist, 2 hemodynamic cardiologists, and 4 cardiac surgeons were enrolled. A numerical code was assigned to each of them in order to protect their privacy. Their level of confidence with MR and DICOM viewer software was investigated through an online questionnaire before the test. No participants declared to be experts in MR technology nor to use it in their working environment; 4 of them stated they had previously tried the technology in specific tests or simulations; 2 participants declared to use MR outside the working environment (e.g., to play videogames or playing sport) (Figure 3.1).

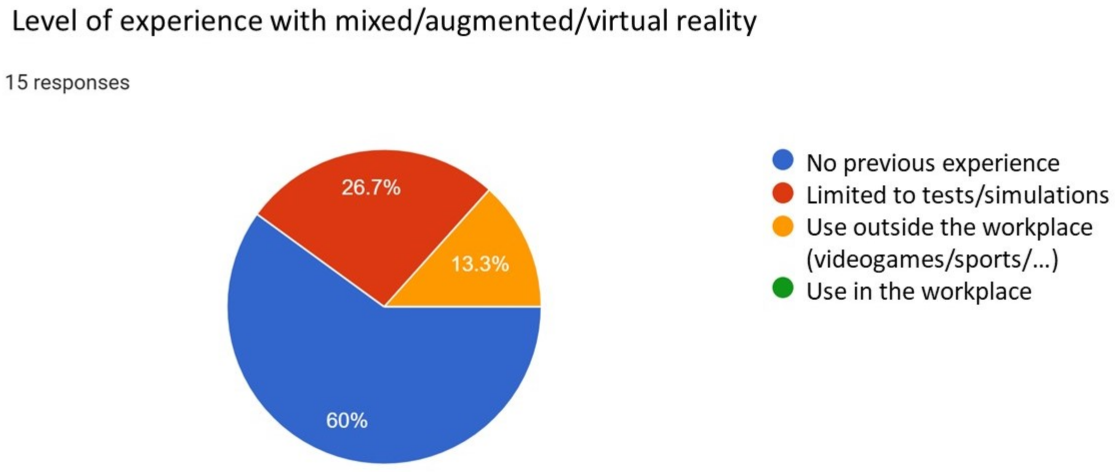


Figure 3.1: *Pie chart representing the level of participants' experience with XR technologies.*

The general lack of significant experience with MR did not allow to analyze the influence of this factor on usability evaluation and time performance, leading to the exclusion of question **4)** from the list of endpoints. Following the assessment of participants' experience with DICOM viewer software, users were clustered into 2 groups: group 1 (n=7), which included skilled DICOM viewer software users, i.e., users with more than 5 years of experience; group 0 (n=9), which included newbies in using DICOM viewers, i.e., users with less than 5 years of experience. Participant 16 was the only exception to this general criterion: despite having used DICOM viewers for less than 5 years, he claimed to be well-experienced and skilled, therefore he was included in group 1. Also, through the same questionnaire, information concerning the working position of the participants was collected to confirm the frequency with which DICOM viewer software is used in their working routine and therefore their experience. The information collected for each participant is schematically shown in Table 3.1:

Table 3.1: *Information on the enrolled participants: identification code (1° column); job category (2° column); membership group (0 = newbies in using DICOM viewer software, 1 = experts in using DICOM viewer software) (3° column); previous experience with XR technologies (4° column).*

PARTICIPANT CODE	JOB	GROUP	EXPERIENCE WITH XR
PART1	interventional cardiologist	0	-
PART2	cardiology resident	0	no previous experience
PART3	hemodynamic cardiologist	1	no previous experience
PART4	hemodynamic cardiologist	1	limited to tests/simulations
PART5	cardiology resident	0	no previous experience
PART6	cardiology resident	0	no previous experience
PART7	cardiac surgeon	1	use outside the working place (videogames/sport/...)
PART8	cardiac surgeon	1	limited to tests/simulations
PART9	cardiac surgeon	1	limited to tests/simulations
PART10	cardiac surgeon	1	no previous experience
PART11	vascular resident	0	limited to tests/simulations
PART12	emergency medicine resident	0	no previous experience
PART13	medical student	0	use outside the working place (videogames/sport/...)
PART14	cardiology resident	0	no previous experience
PART15	cardiology resident	0	no previous experience
PART16	cardiology resident	1	no previous experience

Evaluation test protocol

Two groups of participants worked on two separate test days; the same protocol was applied each day.

Participants were asked to perform three ad hoc defined tasks. These consisted in obtaining 3 echographic-like views of the heart of a patient: 4-chambers (T1), 3-chambers LVOT (left ventricular outflow tract) (T2), and ventricular short axis (T3), which are shown in figure 3.2.

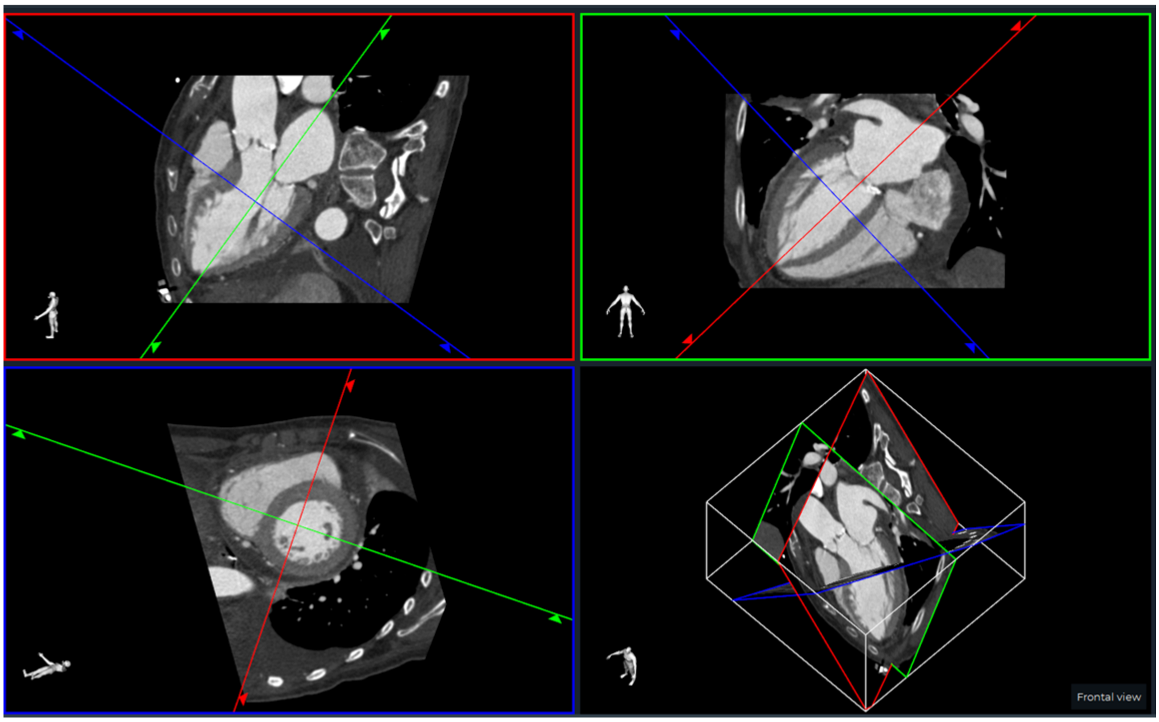


Figure 3.2: *Echographic-like views of the heart: 3-chambers LVOT (top left), 4-chambers (top right), ventricular short axis (bottom left).*

These tasks were chosen to satisfy three criteria: i) not being specific to a particular medical discipline since participants had different backgrounds and levels of experience; ii) avoiding excessive time expense, so as to comply with the busy schedule of clinical participants; iii) avoiding discomfort or annoyance in the participants. These three criteria resulted in low-difficulty tasks. Also, since usability is a very immediate feature to be perceived when using a device, more complex tasks would not have been necessary.

The same three tasks were performed both with the ARTICOR platform and with the RadiAnt software. The order of performance was randomized. The tasks were carried out in a meeting room at OMPM premises; this environment was deemed fully adequate for the defined tasks as it was characterized by sufficient space to move around the hologram when using the MR technology, and by suitable lighting conditions.

The holographic model of the heart has been developed by Artiness from 4D contrast-enhanced TC images (20 time-frames/cardiac cycle, pixel spacing: 0.78125x0.78125 mm, slice thickness: 0.6999 mm) of a patient with aortic stenosis, resulting in a dynamic model where also the heartbeat has been reproduced. On ARTICOR, the model can be navigated, and its rendering can be edited through the tools available on a holographic menu shown next to the model. For example, it is possible to show or hide some anatomical components (chambers, valves, coronaries) in order to focus on the remaining ones, to obtain sections and explore the inside of the model by cutting it with a plane or a cube.

At the beginning of the test, the features, functioning, and purpose of the MR technology were briefly explained collectively to the participants in order to give all of them a basic knowledge of the device they would have tried. Subsequently, each participant was tested individually.

After wearing the headset and regulating it to fit the head size, the model was shown to the user. The calibration of the headset, which would optimize the visualization of the holograms based on the interpupillary distance of the user, was skipped because participants had to interact for a very limited time with the device and the tasks they had to perform were not meant to provide real diagnostic information. Furthermore, in this way, it was possible to speed up the performance of the test. Anyway, calibration is a quick procedure that should be performed prior to the use of the device to help reducing possible annoying feelings and general malaise (e.g., headaches, nausea, eye fatigue) by correcting for the vergence-accommodation conflict.

At this stage of the test, a further explanation was provided individually to the user wearing the headset: he/she was instructed on how to grab, move, and zoom the model and was shown the possibility of looking at the model from different perspectives just by moving around it. The hologram in fact is not anchored to the headset and therefore it remains fixed in the 3D physical space even if the user moves around it.

Also, some specific tools controlled from the menu were explained. For example, the possibility to show or hide some parts of the model or the activation of the cutting plane to be used for navigating the 3D holographic model; this last feature was the pivotal tool to be used in accomplishing the tasks. After activating it from the corresponding menu button, a holographic plane appears in the field of view of the headset. The plane can be grabbed with the same pinch movement used to grab the other holograms and can be moved in order to superimpose it onto the anatomical model. In this way, the model is cut into two parts and one of them is hidden so that it is possible to see its interior. Given that it can be difficult to grab and move the plane in the correct position without also moving the heart, there is a locking button that allows for blocking the anatomical model and moving only the plane in order to obtain the desired view of the interior. Also, some useful shortcuts have been implemented by the developing team, such as the possibility of inverting the cut-view. This function is very useful when the cutting plane is placed in a way to show an inner view but the desired view is the symmetrical one with respect to the plane. In order to simplify their tasks, participants were also instructed about these two functions.

Afterwards, each participant was asked to obtain the three previously described echographic-like views, and the time required to accomplish the tasks was recorded in order to answer research question **3)**. While running the test, the streaming of the holographic content on a PC monitor was active so that the developer team and I had the possibility to visualize in real time the actions performed by the participant on the virtual model as well as to evaluate if the tasks were correctly executed.

Some users performed the tasks firstly with ARTICOR and then with RadiAnt while some others in the opposite sequence. For each participant, the randomized order of the tasks was maintained for both technologies.

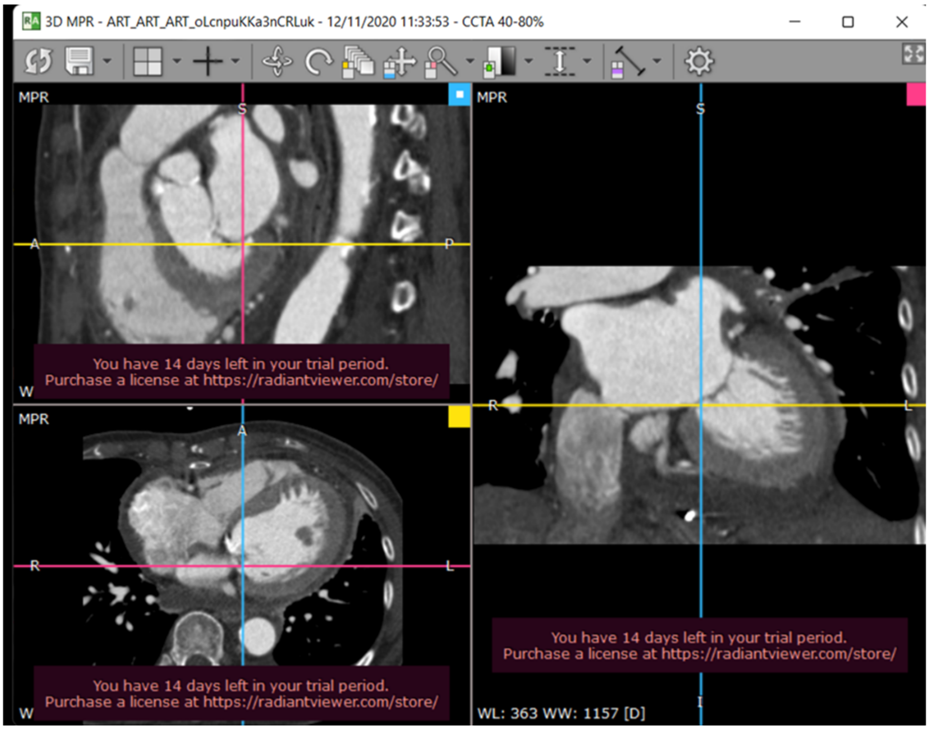
For what concerns the use of RadiAnt viewer, after uploading the CT dataset, the 3D MPR tool provided the visualization of the associated anatomy on the three mutually orthogonal anatomical cut-planes: sagittal, transverse, and coronal (Figure 3.3). The volumetric dataset was then navigated by the user by moving the two cursors present on each of the three cut-plane views and the target view of each task was manually identified. Again, the time required to obtain each view was recorded.

In addition, in order to answer research question **1)**, I took note of participants' comments and difficulties they encountered during the whole trial session.

The other participants were not blind to the performance of the one carrying out the test, since they were physically present inside the test room; yet, they could not visualize the PC monitor where the holographic content was streamed nor the one with RadiAnt software running.

At the end of the test, the users filled in three validated questionnaires and one questionnaire conceived ad hoc for this work.

The validated questionnaires consisted of Surgery Task Load Index, User Experience Questionnaire, and System Usability Scale (see subsection 1.4.2). These are all open source and freely available on the web and can be applied both in absolute and comparative studies. Participants were asked to complete the questionnaires both for ARTICOR and RadiAnt software. In this way, the additional information about the relative perception each user had of the two technologies allowed for answering research questions **2)** and **5)**.



3D MPR - ART_ART_ART_oLcnpuKKa3nCRLuk - 12/11/2020 11:33:53 - CCTA 40-80%

You have 14 days left in your trial period.
Purchase a license at https://radiantviewer.com/store/

You have 14 days left in your trial period.
Purchase a license at https://radiantviewer.com/store/

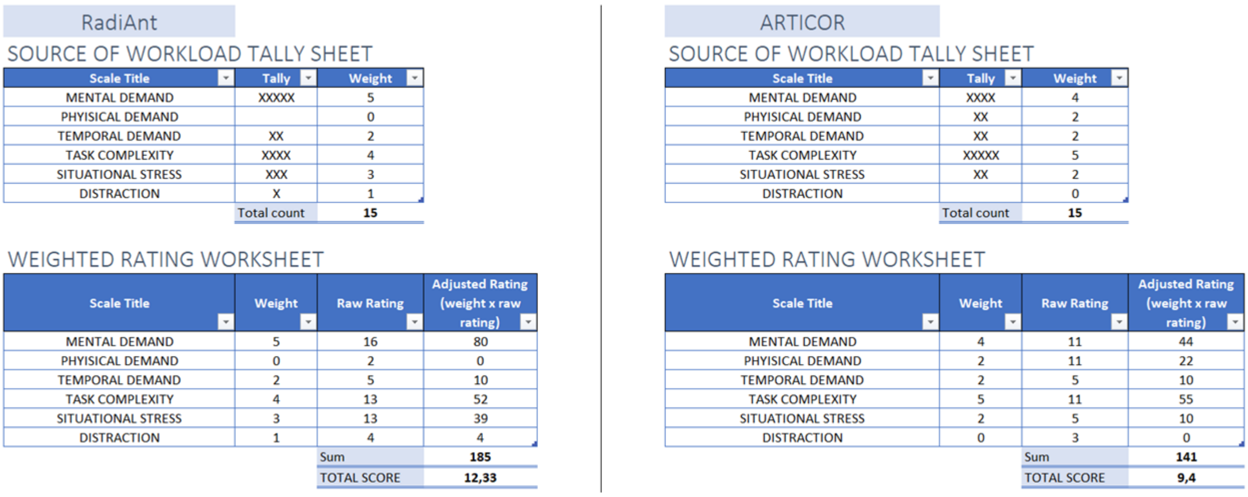
You have 14 days left in your trial period.
Purchase a license at https://radiantviewer.com/store/

WL: 363 WW: 1157 [D]

Figure 3.3: *RadiAnt graphical interface showing 3D MPR of the thoracic CT used in this study. The three anatomical planes are displayed (top left: sagittal; bottom left: transverse; right: frontal). The yellow, pink, and blue cursors allow image navigation.*

3.1.1. Surgery Task Load Index

For each participant, an Excel worksheet containing the collected data was created as exemplified in Figure 3.4; it was used to compute the weight and rating of each source of workload as well as the cumulative workload score, both for Radiant and for ARTICOR. In the process, four tables were generated: two of these summarized the weight assigned to each dimension contributing to the perceived workload, computed considering the tally of the pairwise comparisons constituting the first part of the questionnaire (Figure 3.4, top row). In the two remaining tables, the weights were combined with the raw ratings obtained from the second part of the questionnaire to compute the weighted ratings and, eventually, the total workload score (Figure 3.4, bottom row).



RadiAnt

SOURCE OF WORKLOAD TALLY SHEET

Scale Title	Tally	Weight
MENTAL DEMAND	XXXXX	5
PHIYSICAL DEMAND		0
TEMPORAL DEMAND	XX	2
TASK COMPLEXITY	XXXX	4
SITUATIONAL STRESS	XXX	3
DISTRACTION	X	1
	Total count	15

WEIGHTED RATING WORKSHEET

Scale Title	Weight	Raw Rating	Adjusted Rating (weight x raw rating)
MENTAL DEMAND	5	16	80
PHIYSICAL DEMAND	0	2	0
TEMPORAL DEMAND	2	5	10
TASK COMPLEXITY	4	13	52
SITUATIONAL STRESS	3	13	39
DISTRACTION	1	4	4
		Sum	185
		TOTAL SCORE	12,33

ARTICOR

SOURCE OF WORKLOAD TALLY SHEET

Scale Title	Tally	Weight
MENTAL DEMAND	XXXX	4
PHIYSICAL DEMAND	XX	2
TEMPORAL DEMAND	XX	2
TASK COMPLEXITY	XXXXX	5
SITUATIONAL STRESS	XX	2
DISTRACTION		0
	Total count	15

WEIGHTED RATING WORKSHEET

Scale Title	Weight	Raw Rating	Adjusted Rating (weight x raw rating)
MENTAL DEMAND	4	11	44
PHIYSICAL DEMAND	2	11	22
TEMPORAL DEMAND	2	5	10
TASK COMPLEXITY	5	11	55
SITUATIONAL STRESS	2	5	10
DISTRACTION	0	3	0
		Sum	141
		TOTAL SCORE	9,4

Figure 3.4: *Excel sheet created for the analysis of the Surgery Task Load Index questionnaire.*

3.1.2. User Experience Questionnaire

The data obtained with this questionnaire were analysed through a pre-set Excel file, which allowed for obtaining statistical information and trends relative to both the 26 single items and the 6 scales. The Excel file was organized into four worksheets, each one dedicated to a specific aspect of the analysis:

1. in the first sheet, the data mean value and standard deviation were computed. Participants evaluated each item on a 7-point Likert scale and the grades inserted were mapped onto the interval [-3; +3] where –3 and +3 univocally indicated the worst and the best evaluation, respectively. The results from this analysis were also represented graphically through coloured arrows: green, red, and yellow arrows indicate scores > 0.8 (overall positive evaluation), scores < 0.8 (negative evaluation), and scores in the range [–0.8; 0.8] (neutral evaluation), respectively.

2. In the second sheet, the 5% confidence intervals (CI) for the items and scale means were computed. CI measures the reliability of the estimation of the scale mean: the narrower the CI, the greater the reliability of the results. CI width depends on the sample size and on the consistency of responses: the higher the number of data acquired and the more consistent the opinions, the smaller the confidence interval. An auxiliary worksheet also allowed for knowing how much data would be needed to obtain the desired reliability.

3. The third sheet was the benchmark worksheet. The mean values of the scales for the

product under evaluation were compared vs. the means of an existing dataset. This one has been formed by the results obtained from the evaluation of products such as software, web pages, web shops, and social networks in 468 studies involving 21'175 people. Through this comparison, it was possible to have an idea of the relative quality of the product by classifying its scales into 5 categories: excellent (in the range of the 10% best results), good (10% of the results in the benchmark dataset are better and 75% of the results are worse), above average (25% of the results in the benchmark are better than the result for the evaluated product, 50% of the results are worse), below average (50% of the results in the benchmark are better, 25% of the results are worse), bad (in the range of the 25% worst results).

4. The fourth worksheet allowed for detecting inconsistencies, i.e., random answers provided by the participants. Since all the items of a scale measure more or less the same user experience quality aspect, the responses of a participant to these items should be not too different. If the difference between the best and the worst evaluation of an item is higher than 3, the scale is problematic, but the responses of a subject should be eliminated from the dataset only if the inconsistent scales are more than 2.

In this work, I focused mostly on the first, third, and fourth worksheet.

3.1.3. System Usability Scale

The total usability score was numerically computed as reported in the literature [11]. For each participant, the cumulative score was obtained both in reference to ARTICOR and RadiAnt. The results were stored in an Excel table for further processing.
For what concerns the interpretation of the SUS scores through the "promoters/detractors" approach, which is inherent to this test, I computed the LTR corresponding to each SUS score through the regression equation:
LTR = 1.33+0.08*SUS [30].
Based on the LTR score, I assigned each participant to a class: promoters if LTR >= 9, passives if 9>LTR>=6, and detractors if LTR<6.

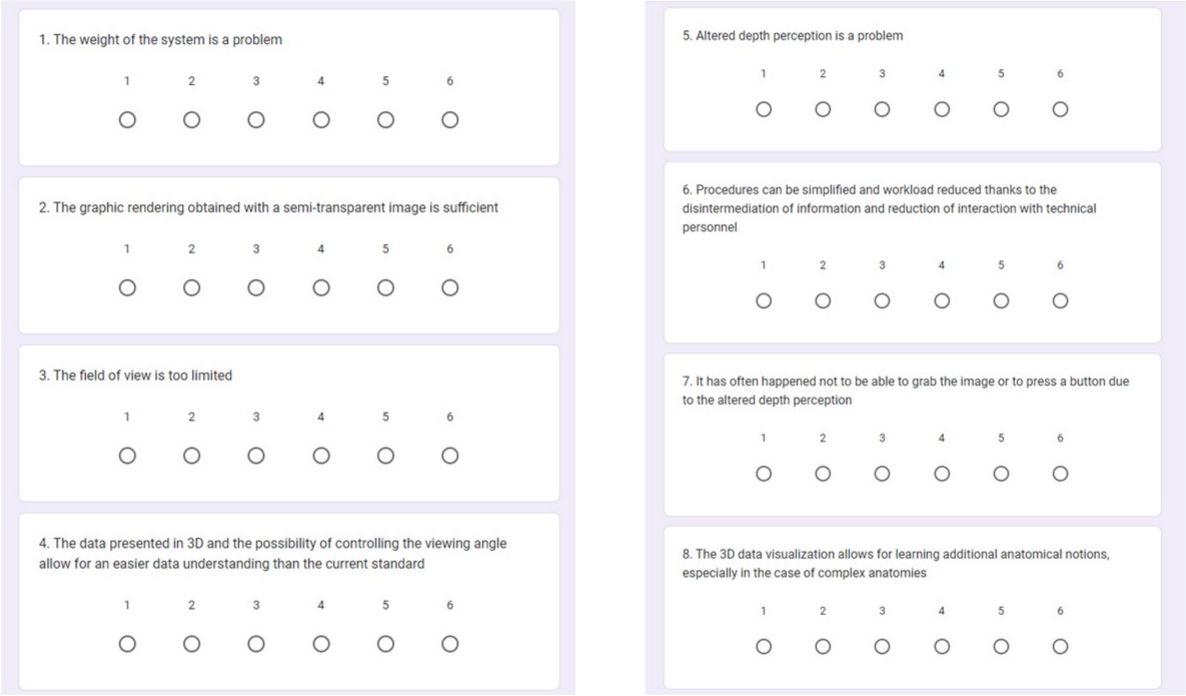
3.1.4. Ad hoc created questionnaire

Given that the above-described questionnaires are not technology-specific, they did not allow for the evaluation of specific aspects and features of the MR platform ARTICOR. Therefore, in order to answer research question **6)**, a more specific questionnaire was developed. The aspects to be investigated through this additional tool were derived from

literature: they consisted of features, functionalities, and potential issues concerning the technology under evaluation and its application in a real-case scenario.
Following the methodology proposed for the SUS (see subsection 1.4.2), respondents had to indicate their level of agreement with 8 sentences where positive and negative items were alternated to avoid response biases. Differently from the 5-point Likert scale of the SUS questionnaire, I decided to apply a 6-point Likert scale (Figure 3.5). In this way, respondents were not offered the possibility to provide a neutral response at the center of the scale. This fact increased the commitment to be applied when filling in the questionnaire, but it also allowed for a clearer distinction between positive and negative user experiences. Among these 8 questions, 4 were "positive", meaning that a high score was representative of a strength of the system, and the other 4 were negative, meaning that a high score was representative of a weakness of the system.
The 8 questions composing the questionnaire were:

1. The weight of the system is a problem;
2. The graphic rendering obtained with a semi-transparent image is sufficient;
3. The field of view is too limited;
4. The data presented in 3D and the possibility of controlling the viewing angle allow for an easier data understanding than the current standard;
5. Altered depth perception is a problem;
6. Procedures can be simplified and workload reduced thanks to the disintermediation of information and reduction of interaction with technical personnel;
7. It has often happened not to be able to grab the image or to press a button due to the altered depth perception;
8. The 3D data visualization allows for learning additional anatomical notions, especially in the case of complex anatomies.

Following again the methodology used to compute the total SUS score, the contribution of the positive elements was computed as (scale position - 1) while the contribution of the negative ones as (6 - scale position). Therefore, each question's contribution ranged from 0 to 5. The contribution of each question was then added up and the result was multiplied by 2.5. In this way, I obtained again a usability score ranging from 0 to 100, which is a very intuitive scale.



1. The weight of the system is a problem

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

2. The graphic rendering obtained with a semi-transparent image is sufficient

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

3. The field of view is too limited

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

4. The data presented in 3D and the possibility of controlling the viewing angle allow for an easier data understanding than the current standard

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

5. Altered depth perception is a problem

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

6. Procedures can be simplified and workload reduced thanks to the disintermediation of information and reduction of interaction with technical personnel

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

7. It has often happened not to be able to grab the image or to press a button due to the altered depth perception

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

8. The 3D data visualization allows for learning additional anatomical notions, especially in the case of complex anatomies

1 2 3 4 5 6

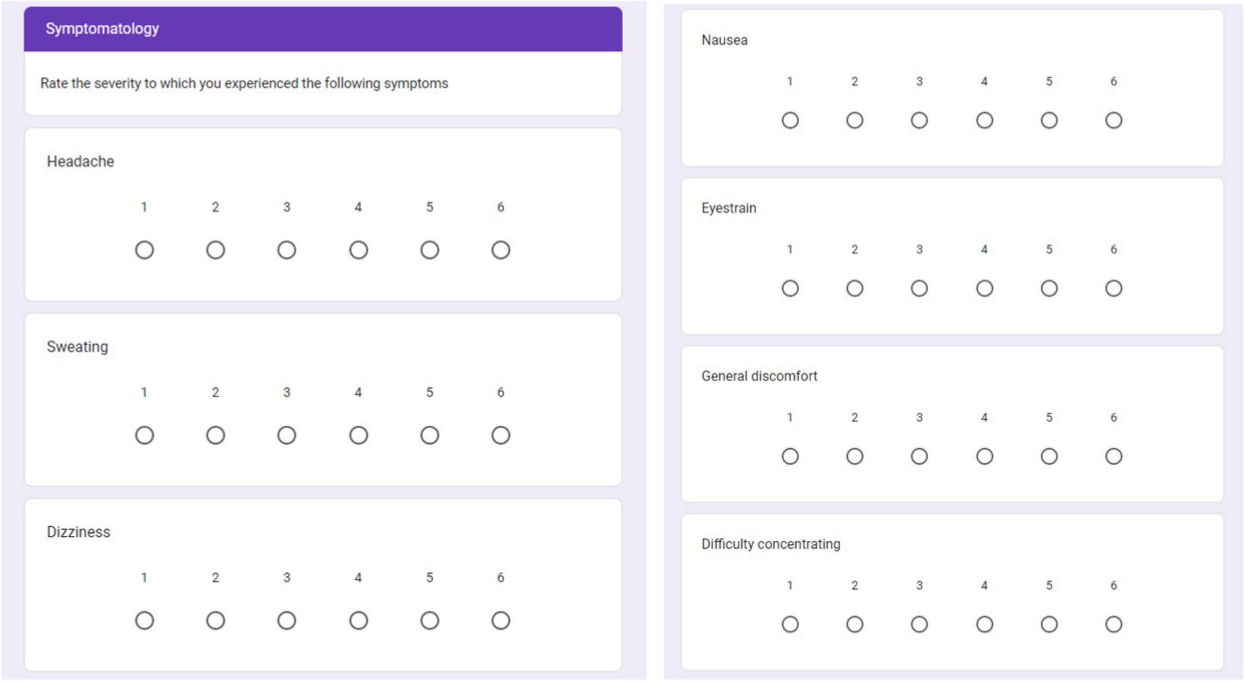
○ ○ ○ ○ ○ ○

Figure 3.5: *Ad hoc developed questionnaire template.*

The second part of the developed questionnaire concerned the symptomatology experienced while interacting with the MR headset during the test phase. Again, respondents had to indicate the severity of different physical symptoms perceived (headache, nausea, sweating, . . .) on a scale from 1 to 6. The dimensions considered in this last part of the questionnaire were derived from a validated tool used to quantify the sickness provoked by VR systems, referred to as the Simulator Sickness Questionnaire (SSQ) [10].

In the SSQ, participants, after the exposure, are asked to rate on a scale from 0 to 3 the severity of 16 symptoms. Higher scores indicate stronger perceptions of the underlying sickness symptoms during the test and are therefore undesired. The scores attributed to each symptom are then combined in a total sickness score which allows for the evaluation of the goodness of a VR simulator.

In the study, I decided not to administer the whole questionnaire because, being designed for immersive VR applications, some of the considered symptoms did not apply to this context. Moreover, it was important for me to keep the test as short as possible to maintain the attention and concentration of the participants; the administration of an additional questionnaire would have been detrimental to this criterion. Therefore, among the 16 symptoms considered in the SSQ, I extracted those potentially most significant in this application context. Eventually, 7 questions were added to the ad-hoc developed questionnaire so that the participants rated the severity of headache, nausea, dizziness, sweating, eyestrain, difficulty in concentrating, and general discomfort (Figure 3.6).



Symptomatology

Rate the severity to which you experienced the following symptoms

Headache

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

Sweating

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

Dizziness

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

Nausea

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

Eyestrain

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

General discomfort

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

Difficulty concentrating

1 2 3 4 5 6

○ ○ ○ ○ ○ ○

Figure 3.6: *Questions investigating the symptomatology in the ad hoc developed questionnaire.*

In addition to these two sections, an open question was also provided to detect which may be the most promising field of application of the technology according to the users. The question was formulated in the following way: "In which application could the greatest added value be achieved by using the technology?". Respondents were offered the possibility to write their answer in a blank space.

The questionnaire was administered in Italian through Google Forms.

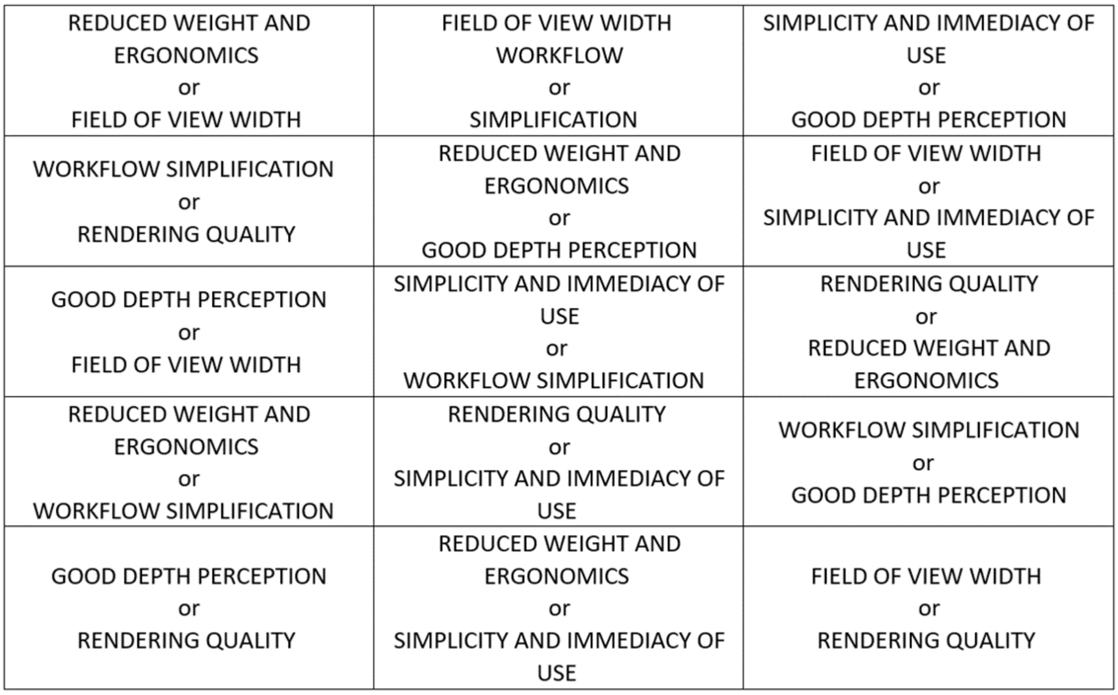
3.2. Technology assessment methodology

Up to now, this study has been conducted with the purpose of providing benefits to manufacturers and final users of the technology of interest. An additional goal has been defined considering the perspective of another important player involved in the technology's lifecycle: those responsible for technology evaluation and acquisition.
In this context, I focused on the definition of a methodology to be applied when, in the healthcare sector, devices like the one under investigation have to be compared in order to decide which one to acquire.

In public healthcare, medical devices are acquired through tender procedures. During standard tender procedures, technology evaluations take into account both technical specifications and economic information about the device to be acquired. Typically, a total of 70/100 points is attributed to the technical features while the remaining 30/100 to economic considerations. These 70 points are given by the sum of the scores assigned to each technical evaluation criterion. Moreover, for each criterion, a maximum assignable score is defined based on its relative relevance. The criteria are technology-specific, meaning that they strictly depend on the device under evaluation. Because of the novelty of the technology of interest, no previous examples of purchasing processes may be exploited to have an idea about the possible criteria and their relative relevance. The need for their definition is further highlighted by the likely progressive spread of similar technologies in the next future.
Following these considerations, the clinical engineering department of Policlinico di Milano Hospital reported to me the necessity to gather some information that might help them in the evaluation of similar technologies. Therefore, I tried to settle possible evaluation criteria in terms of features and functionalities of the platform as well as the weight to be assigned to each one.

Based on literature searches as well as opinions of developers, experts, and potential users, 6 dimensions of evaluation were defined, namely reduced weight and ergonomics, field of view width, good depth perception, rendering quality, workflow simplification, and simplicity and immediacy of use.
Their relative relevance, i.e., the weight of each of them in the computation of the final quality score, was assigned by asking the participants, in an indirect way, which dimensions would have been the most relevant in their opinion. In fact, beyond economic evaluations, the best technology to be acquired strictly depends on its performance related to the technical aspects that are more relevant to the end-users. The methodology followed to derive this information was the already mentioned pairwise comparison. This approach was very similar to the one used in the validated S-TLX questionnaire to compute the relative relevance of the different factors contributing to the workload experienced during a specific task (see subsection 1.4.2). The 6 selected criteria were combined two by two in all the possible ways, resulting in a total of 15 couples (Figure 3.7). For each couple, the respondent had to select the item which, in his opinion, is the most relevant between the two and whose improvements would deserve most of the investments. In particular, the question was articulated as follows: “between the items of each couple, which functionality of the technology do you think is more important to be guaranteed in order to boost its effectiveness and added value?”. The tally of the number of times each item was selected

in the couples was a maximum of 5/15 and this number was converted into a percentage so that it was more intuitive to understand the relative relevance.



REDUCED WEIGHT AND ERGONOMICS or FIELD OF VIEW WIDTH	FIELD OF VIEW WIDTH WORKFLOW or SIMPLIFICATION	SIMPLICITY AND IMMEDIACY OF USE or GOOD DEPTH PERCEPTION
WORKFLOW SIMPLIFICATION or RENDERING QUALITY	REDUCED WEIGHT AND ERGONOMICS or GOOD DEPTH PERCEPTION	FIELD OF VIEW WIDTH or SIMPLICITY AND IMMEDIACY OF USE
GOOD DEPTH PERCEPTION or FIELD OF VIEW WIDTH	SIMPLICITY AND IMMEDIACY OF USE or WORKFLOW SIMPLIFICATION	RENDERING QUALITY or REDUCED WEIGHT AND ERGONOMICS
REDUCED WEIGHT AND ERGONOMICS or WORKFLOW SIMPLIFICATION	RENDERING QUALITY or SIMPLICITY AND IMMEDIACY OF USE	WORKFLOW SIMPLIFICATION or GOOD DEPTH PERCEPTION
GOOD DEPTH PERCEPTION or RENDERING QUALITY	REDUCED WEIGHT AND ERGONOMICS or SIMPLICITY AND IMMEDIACY OF USE	FIELD OF VIEW WIDTH or RENDERING QUALITY

Figure 3.7: *Pairwise comparison applied to define the relative relevance of the 6 selected evaluation criteria.*

Furthermore, the pairwise comparison was performed both before and after the test, so to investigate whether the a priori ideas the users had on the technology were consistent with the ones developed after having tried it. The first assessment, in fact, was performed after having shown the headset to the user and briefly described what a MR platform is, but without offering him/her the possibility to interact with the technology. The second assessment, instead, was run after performing the tasks in the usability test and after possible additional interactions the user wanted to have to explore other functionalities of the platform.

3.3. Data analysis

The data collected through the different questionnaires were organized in Excel. The spreadsheet was also exploited to compute the cumulative score resulting from the S-TLX and SUS questionnaires. Data from the UEQ were analyzed instead through the Excel file attached to the questionnaire itself, which has already been described in the corresponding subsection (see 3.1.2).

For what concerns the management of the results from the ad hoc developed questionnaire, the plots automatically computed by Google Form were quite self-explaining but, in addition, a cumulative score was computed following the same methodology used to compute the SUS score (see 3.1.4).

Moreover, a statistical analysis of some results was performed by using SPSS Statistics (`https://www.ibm.com/products/spss-statistics`).
The statistical analysis aimed to investigate whether the S-TLX and the SUS scores statistically differed between the two technologies (i.e., ARTICOR platform and RadiAnt software) and between the two groups of users (i.e., experts and newbies in using DICOM viewer software).
Since scores were not normally distributed and given the small sample size, non-parametric tests were applied. In case of missing data, i.e., some participants did not fill in some fields of a questionnaire, the subject has been excluded from the statistical analysis concerning the relative questionnaire.
The results of these tests must be cautiously considered since the sample size was small, hence any statistical significance found could lack reliability.

The S-TLX and the SUS scores obtained by the two technologies were compared via **paired-sample sign test**, which is an alternative to the paired-sample t-test and to the Wilcoxon signed-rank test. It is used to determine whether there is a median difference between matched observations when the distribution of the differences between paired observations is neither normal nor symmetrical. The null hypothesis, which states that the median difference between the two matched variables is equal to 0, can be rejected when statistical significance is found. This test is applied to paired variables, i.e., when participants are tested at two time points or under two different conditions. In this case, the same individuals were subjected to two different conditions, since they were asked to fulfill the defined tasks both with the MR platform and with the traditional computer software.
Thus, in this study, two different dependent variables were considered: the SUS score and the S-TLX score. For each, the two conditions to which participants have been exposed were: performing the tasks with ARTICOR and with RadiAnt. As a result, the two paired-sample sign tests implemented can be summarized as follow: `SUS_ARTICOR` vs `SUS_RadiAnt`; `STLX_ARTICOR` vs `STLX_RadiAnt`. The lack of symmetry in the difference between the considered variables is evident from the distributions shown in Figure 3.8.

Differences between the two groups of users were analyzed by the **Mann-Whitney U test**. This test is used to compare two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed. It is sometimes considered the

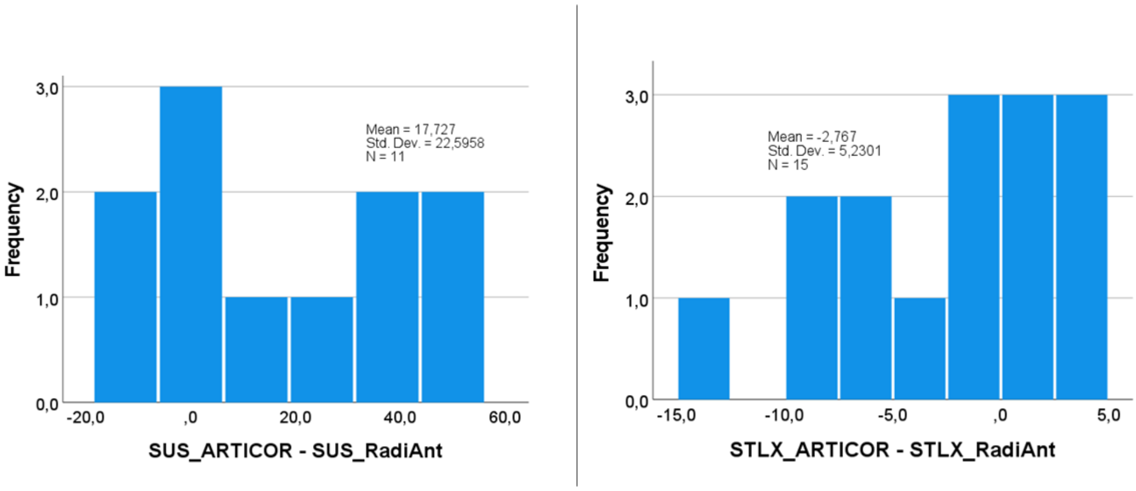


Figure 3.8: *Histograms of the distribution of the variables SUS_ARTICOR - SUS_RadiAnt (left), STLX_ARTICOR - STLX_RadiAnt (right). The lack of symmetry is evident.*

non-parametric alternative to the independent t-test since, being based on ranks instead of means, can be applied when normality cannot be assumed. Before applying it, some hypotheses have to be verified: the dependent variable must be ordinal or continuous; the independent variable must consist of two categorical, independent groups; the observations must be independent, i.e., no relationship can exist between the observations in each group or between the groups themselves. In the present case, all these assumptions were confirmed.

Additionally, even if this type of test is used with not normally distributed variables, it is useful to investigate whether the two distributions have the same shape. If this is the case, the test can be used to investigate whether there is a statistical difference between the medians of the dependent variable for the two groups. Otherwise, only mean ranks can be compared.

The null hypothesis states that there is not any tendency for one of the two populations in presenting a higher score with respect to the other, i.e., the distribution of the dependent variable is the same in the two groups. When statistical significance is verified, it is possible to reject the null hypothesis, thus affirming with sufficient evidence that the two considered groups of participants come from populations with different values of the dependent variable.

In this case, the test was implemented to investigate whether the scores representing the usability (SUS score) and the workload (S-TLX score) relative to ARTICOR and RadiAnt were statistically different between the two groups of users. Therefore, the Mann-

Whitney U test was applied to the following two dependent variables: **V1** = `SUS_ARTICOR - SUS_RadiAnt` , **V2** = `STLX_ARTICOR - STLX_RadiAnt`.

The decision to consider the differential scores rather than the absolute ones was guided by the fact that participants' judgement on the usability and workload perceived when using the two technologies depended not only on the features of the single technologies, but also (and maybe even more) on their own personal features. For example, some operators might rate both systems with high usability scores and low workload scores because of their greater experience in the cardiologic field, better stress management capabilities, or higher self-confidence. Some others, instead, might rate the systems with low scores because of low self-esteem or lack of confidence with the tasks themselves, and not because the system was not perceived as usable. By considering the differential score, the influence of other factors beyond the level of expertise with DICOM viewer software, which was the discriminating variable between the two groups, did not bias the results.

Cases in which V1 or V2 were equal to zero were excluded from the statistical analysis since they were synonyms of indifference of opinion, a case that was not considered in the present study.

Furthermore, in order to improve the robustness of the results, the same statistical analysis was applied also to the absolute value of the two variables. The reason for this additional test was that the inter-group difference for each score was assumed to be positive in one group and negative in the other one. Therefore, statistical significance would be quite easy to be found. Considering instead the absolute value of the difference, it was possible to evaluate more reliably whether the SUS and S-TLX scores of the two technologies significantly differ in the two groups of participants.

In all cases, the significance level was set to 0.05.

4 | Results and discussion

4.1. Usability test

4.1.1. Report from participants' observation

Given that most of the participants did not have prior experience with MR technologies, their first reaction after wearing the headset was amazement. In particular, the sense of surprise was generated by the modalities of interaction with the model. The possibility of zooming and rotating, of cutting it in order to visualize a section, of looking from different perspectives just by walking around it, was unexpected to the users. Many of them showed curiosity about how the technology had been developed and they often asked questions about the possible application scenarios. The enthusiasm of the participants was confirmed when, after completing the tasks, many of them asked to keep the headset on in order to explore the other functionalities of the platform and to gain competence and expertise in using it. Participants' comments were positive in most cases, including the ones from participants experienced in the use of standard DICOM viewer software and potentially less prone to appreciate the new MR platform.

No significant complaints were reported by the participants during the tasks with the MR platform, and all of them managed to fulfill the three tasks without relevant problems.

The most observed difficulty was related to the altered depth perception. This caused 75% of the participants to fail in grabbing the hologram or in pressing a holographic button more than twice; still, everyone managed to complete the action after few extra attempts.

Also the loss of the hologram, i.e., the impossibility to locate the hologram in the 3D space while using the MR headset, was experienced by 75% of the participants, but no participant experienced it more than once. Moreover, only in one case my intervention was needed, while in all the other situations participants managed to solve the problem on their own.

While navigating the 3D hologram to find the three echographic views, another difficulty emerged: many users struggled in moving the cutting plane when this was superimposed on the 3D heart model. Even if informed of the “locking function” (i.e., the possibility of blocking the movement of the model and moving only the plane) at the beginning of the session, many of them did not recall this information and my suggestion was required to facilitate task completion.

In the second phase of the test, when participants had to perform the same tasks with the DICOM viewer software, the comments and performance were very different depending on the level of previous experience with the tool. Newbies often felt in trouble and asked for my help. Some of them also showed embarrassment because of the longer time they took to complete the task. In some cases, they gave up before finding the required view. On the other hand, participants experienced in the use of the software completed the three tasks without any issue and reported them as very easy to be performed.

4.1.2. Time performances

The times required by the participants to perform the three tasks with the two technologies are summarized in Table 4.1, 4.2, 4.3. In particular, table 4.1 shows the time performances of all the participants, while in tables 4.2 and 4.3 participants have been subdivided into the two groups (group 0 and 1, respectively).

Table 4.1: *Time performances and average completion time (last row) for all participants, with ARTICOR and Radiant. The cardinal numbers 1°, 2°, and 3° refer to the order in which the tasks were performed upon randomization. Empty table boxes represent the failure of the user in performing that task.*

	ARTICOR			RadiAnt		
PARTICIPANT (GROUP)	**1°**	**2°**	**3°**	**1°**	**2°**	**3°**
PART1 (0)	00:01:00	00:00:25	00:00:45	00:01:00	00:02:00	00:02:00
PART2 (0)	00:00:17	00:03:00	00:03:00	00:01:00	00:05:00	00:05:00
PART3 (1)	00:00:15	00:01:30	00:00:30	00:00:10	00:00:10	00:00:10
PART4 (1)	00:01:30	00:01:30	00:00:20	00:00:10	00:00:40	00:00:10
PART5 (0)	00:00:34	00:00:30	00:00:40	00:02:00	00:03:15	00:00:55
PART6 (0)	00:01:15	00:00:10	00:00:08	00:01:30	00:02:45	00:00:15
PART7 (1)	00:01:10	00:01:10	00:01:44	00:00:10	00:01:10	00:00:13
PART8 (1)	00:01:07	00:00:22	00:00:25	00:00:10	00:00:15	00:00:20
PART9 (1)	00:00:30	00:00:30	00:01:10	00:00:18	00:00:20	00:00:25
PART10 (1)	00:01:10	00:04:00	00:01:00	00:01:05	00:00:15	00:00:15
PART11 (0)	00:00:55	00:00:05	00:00:13	00:01:15	X	X
PART12 (0)	00:01:06	00:01:04	00:00:40	00:01:30	00:01:10	00:00:30
PART13 (0)	00:00:20	00:00:30	00:00:08	00:00:50	00:01:00	00:00:50
PART14 (0)	00:01:16	00:00:40	00:00:40	00:00:50	00:01:10	00:00:30
PART15 (0)	00:00:20	00:00:30	00:00:40	00:01:05	00:01:40	00:00:40
PART16 (1)	00:00:50	X	00:00:10	00:00:53	00:00:55	00:00:30
AVERAGE	**00:00:51**	**00:01:04**	**00:00:46**	**00:00:52**	**00:01:27**	**00:00:51**

Table 4.2: *Time performances and average completion time (last row) for participants of group 0, with ARTICOR and Radiant. The cardinal numbers 1°, 2°, and 3° refer to the order in which the tasks were performed upon randomization. Empty table boxes represent the failure of the user in performing that task.*

	ARTICOR			RadiAnt		
PARTICIPANT (GROUP)	**1°**	**2°**	**3°**	**1°**	**2°**	**3°**
PART1 (0)	00:01:00	00:00:25	00:00:45	00:01:00	00:02:00	00:02:00
PART2 (0)	00:00:17	00:03:00	00:03:00	00:01:00	00:05:00	00:05:00
PART5 (0)	00:00:34	00:00:30	00:00:40	00:02:00	00:03:15	00:00:55
PART6 (0)	00:01:15	00:00:10	00:00:08	00:01:30	00:02:45	00:00:15
PART11 (0)	00:00:55	00:00:05	00:00:13	00:01:15	X	X
PART12 (0)	00:01:06	00:01:04	00:00:40	00:01:30	00:01:10	00:00:30
PART13 (0)	00:00:20	00:00:30	00:00:08	00:00:50	00:01:00	00:00:50
PART14 (0)	00:01:16	00:00:40	00:00:40	00:00:50	00:01:10	00:00:30
PART15 (0)	00:00:20	00:00:30	00:00:40	00:01:05	00:01:40	00:00:40
AVERAGE	**00:00:47**	**00:00:46**	**00:00:46**	**00:01:13**	**00:02:15**	**00:01:20**

Table 4.3: *Time performances and average completion time (last row) for participants of group 1, with ARTICOR and Radiant. The cardinal numbers 1°, 2°, and 3° refer to the order in which the tasks were performed upon randomization. Empty table boxes represent the failure of the user in performing that task.*

	ARTICOR			RadiAnt		
PARTICIPANT (GROUP)	**1°**	**2°**	**3°**	**1°**	**2°**	**3°**
PART3 (1)	00:00:15	00:01:30	00:00:30	00:00:10	00:00:10	00:00:10
PART4 (1)	00:01:30	00:01:30	00:00:20	00:00:10	00:00:40	00:00:10
PART7 (1)	00:01:10	00:01:10	00:01:44	00:00:10	00:01:10	00:00:13
PART8 (1)	00:01:07	00:00:22	00:00:25	00:00:10	00:00:15	00:00:20
PART9 (1)	00:00:30	00:00:30	00:01:10	00:00:18	00:00:20	00:00:25
PART10 (1)	00:01:10	00:04:00	00:01:00	00:01:05	00:00:15	00:00:15
PART16 (1)	00:00:50	X	00:00:10	00:00:53	00:00:55	00:00:30
AVERAGE	**00:00:56**	**00:01:30**	**00:00:46**	**00:00:25**	**00:00:32**	**00:00:18**

Considering the whole group of participants, the average time to perform the three tasks with ARTICOR was slightly shorter with respect to the time required by the same tasks with RadiAnt (reduction of 00:00:01, 00:00:23, and 00:00:05 seconds when performing the three tasks with ARTICOR) (Table 4.1).
It is interesting to note the presence of some outliers inside the participant group. Participant 2 for example took 3 minutes to fulfill the second and third tasks with ARTICOR and 5 minutes for the same tasks with RadiAnt. Thus, it is possible that he had trouble understanding and implementing the tasks themself. Participant 10 took an above-average time to fulfill the second task with ARTICOR, but this was due to the fact that he lose the image while performing it and my intervention was necessary to tackle the problem. When removing these two participants from the computation of the average performance time, the difference between the two technologies is even less significant: the average times with ARTICOR resulted in 00:00:52, 00:00:41, 00:00:35 seconds and with RadiAnt in 00:00:51, 00:01:16, 00:00:34 seconds.

More significant differences can be observed when the two groups are analysed separately (Tables 4.2 and 4.3)
For group 0, the average completion time with ARTICOR was below 1 minute for each task; with RadiAnt instead, it was above 1 minute for each task (reduction of 00:00:26, 00:01:29, 00:00:34 seconds respectively when performing the three tasks with ARTICOR). On the other hand, for group 1, the time to fulfill the tasks with the standard software was shorter than the one required with the MR platform (reduction of 00:00:31, 00:00:58, and 00:00:28 seconds respectively when performing the three tasks with RadiAnt).
These data suggest that ARTICOR resulted in being significantly more intuitive and easier to use for users with no experience in any of the two considered technologies. Instead, for users experienced in using traditional DICOM viewer software, RadiAnt resulted in faster task completion.
No significant decrease in time-efficiency was observed when comparing the last task (3°) to the first one (1°). Thus, it was not possible to infer about the learning curve associated with the use of the MR platform. More complex and longer tasks might be useful in providing insights into this aspect.

Beyond these preliminary and qualitative considerations, data from the usability test were collected through the already described questionnaires (3.1.1, 3.1.3, 3.1.2, 3.1.4). The result of their analysis is summarized in the next subsections, each of which focuses on a single questionnaire.

4.1.3. Surgery Task Load Index

The S-TLX questionnaire was correctly filled in by all the participants; therefore, no missing data was detected.

The trend of the weighted ratings (weight*raw rating) of each workload dimension was investigated in order to gain insight into the relative importance of the different stressors in the defined tasks. From the plots shown in Figure 4.1 and 4.2, some considerations can be derived.
For what concerns the use of ARTICOR, *task complexity* was the most relevant stress source for most of the participants. Also, *mental* and *temporal demand* played a significant role in contributing to the perceived total workload.

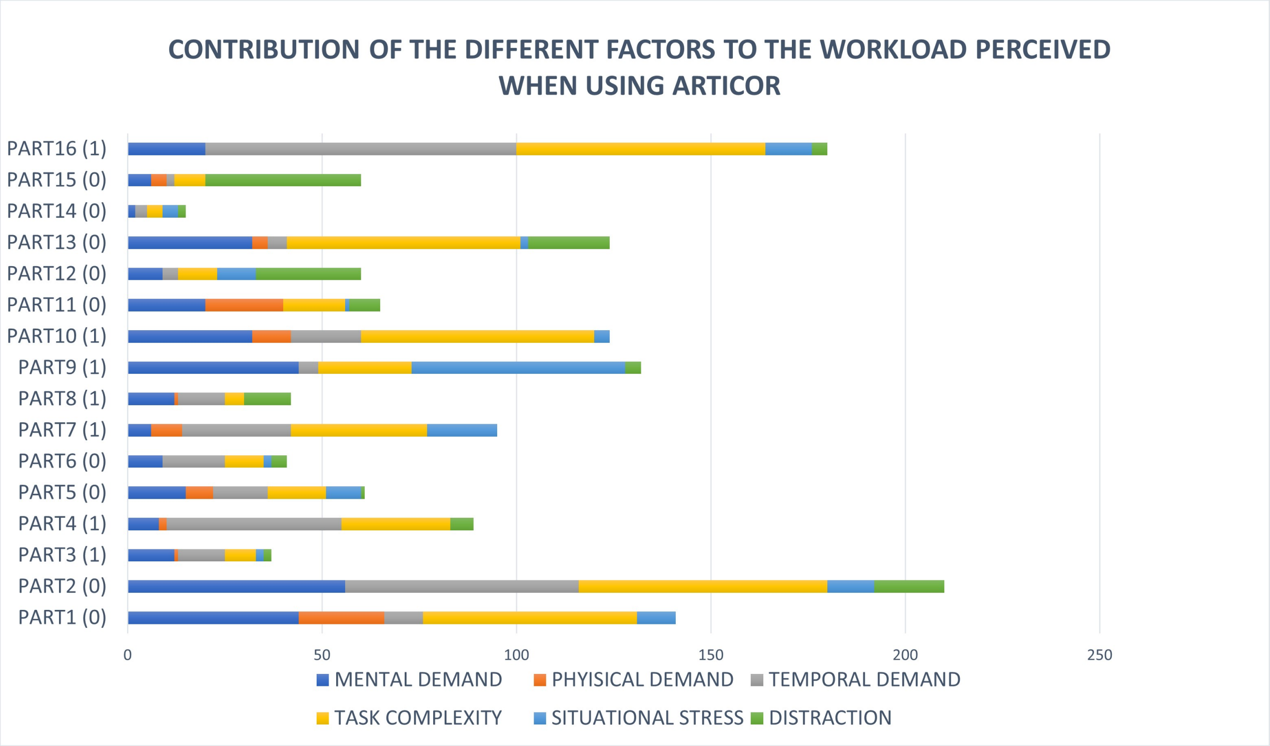


Figure 4.1: *Contribution of the different stressors to the workload perceived when using ARTICOR platform.*

When using RadiAnt software, *mental demand*, *temporal demand*, and *task complexity* were again the dominating dimensions but the contribution of *mental demand* gained relative importance. Also, the contribution of *situational stress* was more significant in this case, confirming that for newbie participants the difficulties encountered in performing the tasks with the standard software made the situation stressful and anxious.

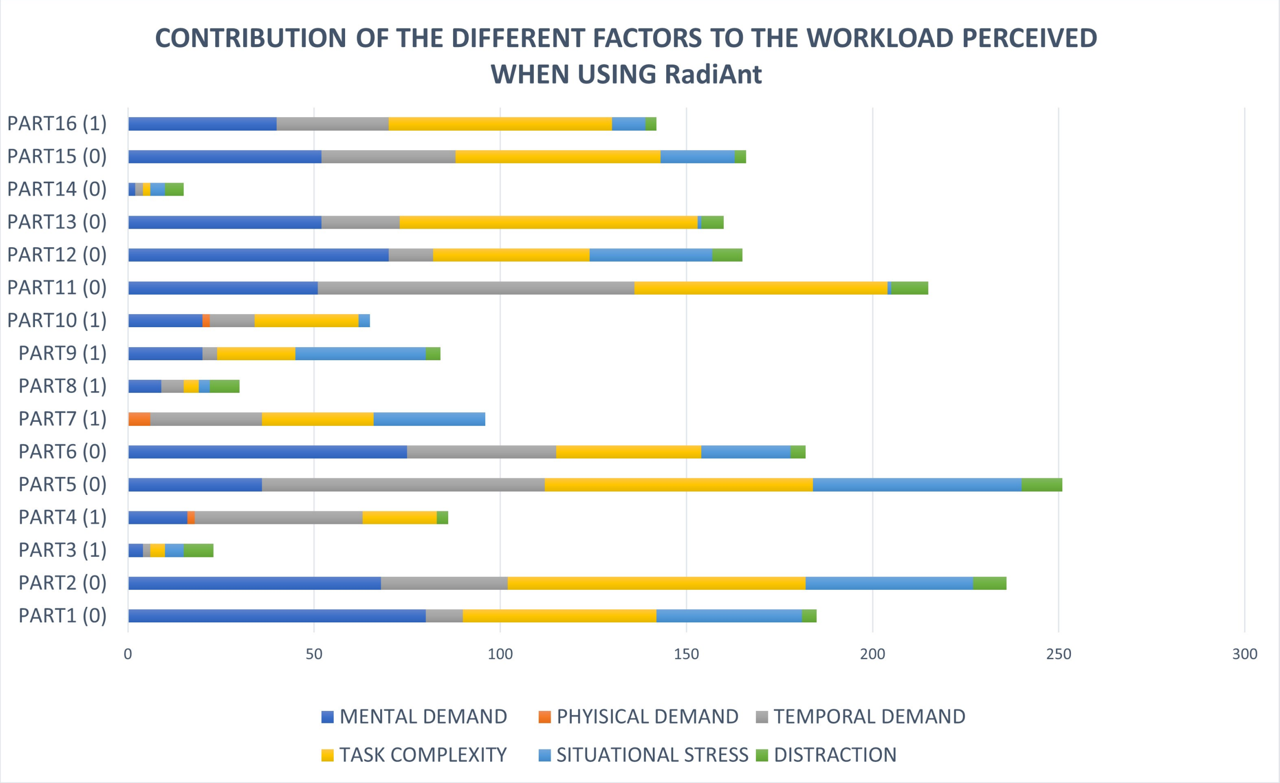


Figure 4.2: *Contribution of the different stressors to the workload perceived when using RadiAnt DICOM viewer.*

Distraction and *physical demand* were, in both cases, the least significant sources of workload.
Despite its low contribution, it is interesting to note the trend of the *physical demand* dimension: when using the gold standard, i.e., Radiant DICOM viewer, it was hardly ever perceived while, when using the MR, its importance grew. This is immediately explainable by the fact that RadiAnt is delivered through a PC, therefore the only movement required by the user is the shift of the mouse. On the other hand, the interaction with ARTICOR requires the user to grab the virtual model and the virtual devices with his own hands, move them in the space as if they were physically present, and walk around the model in order to see it from different perspectives. The physical effort required by these tasks is a source of workload which cannot be ignored when referring to MR.

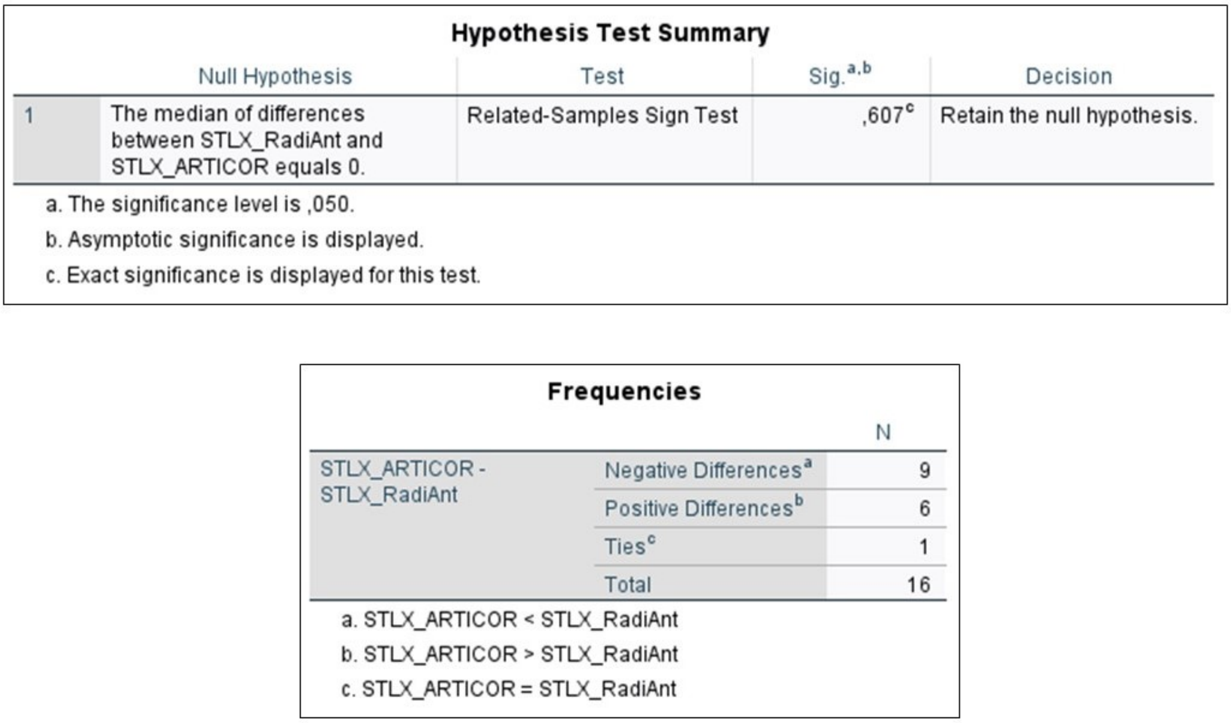
As described in the Methods section, the S-TLX scores associated to the two technologies by the two groups of participants were statistically analyzed to check for differences between technologies and between participants with different levels of experience in the use of DICOM viewers.

The total S-TLX score relating to RadiAnt and ARTICOR, computed considering the weights and the ratings of all the six workload dimensions, is summarized in Table 4.4.

Table 4.4: *Summary of the Surgery Task Load Index scores concerning RadiAnt software and ARTICOR platform. Higher scores indicate higher level of workload and stress perceived when using the related technology.*

PARTICIPANT CODE	GROUP	STLX_RadiAnt	STLX_ARTICOR
PART1	0	12,3	9,4
PART2	0	15,7	14,0
PART3	1	1,5	2,5
PART4	1	5,7	5,9
PART5	0	16,7	4,1
PART6	0	12,1	2,7
PART7	1	6,4	6,3
PART8	1	2,0	2,8
PART9	1	5,6	8,8
PART10	1	4,3	8,3
PART11	0	14,3	4,3
PART12	0	11,0	4,0
PART13	0	10,7	8,3
PART14	0	1,0	1,0
PART15	0	11,1	4,0
PART16	1	9,5	12,0

The results of the Paired-sample Sign test applied to the paired variables `STLX_ARTICOR` and `STLX_RadiAnt` considering the whole participants' group are shown in Figure 4.3. When comparing RadiAnt and ARTICOR, no statistically significant difference in the median S-TLX score of the two distributions was found (p=0.607). The lack of statistically significant difference in the median S-TLX score between the two technologies may be due to the heterogeneity of the selected population in terms of previous experience in using standard DICOM viewers, of attitude towards new technologies, of clinical background and medical experience. However, analyzing the frequency of the variable V2 = `STLX_ARTICOR - STLX_RadiAnt` it is possible to observe that the negative differences were more numerous than the positive ones, indicating that for more participants the workload perceived when using the MR platform was lower than the one perceived when using the traditional computer software.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The median of differences between STLX_RadiAnt and STLX_ARTICOR equals 0.	Related-Samples Sign Test	,607[c]	Retain the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

c. Exact significance is displayed for this test.

Frequencies

		N
STLX_ARTICOR - STLX_RadiAnt	Negative Differences[a]	9
	Positive Differences[b]	6
	Ties[c]	1
	Total	16

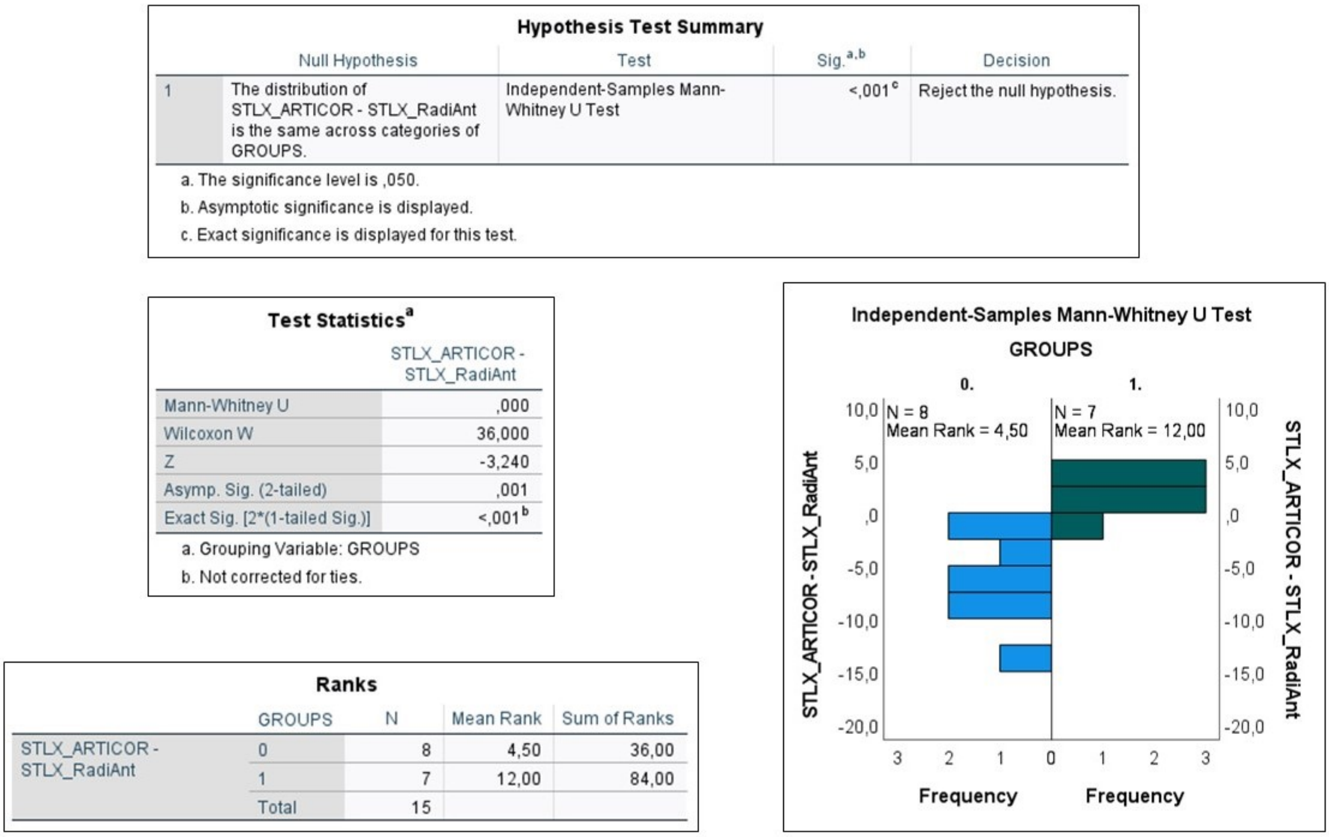
a. STLX_ARTICOR < STLX_RadiAnt

b. STLX_ARTICOR > STLX_RadiAnt

c. STLX_ARTICOR = STLX_RadiAnt

Figure 4.3: *Result of the Paired-sample Sign test applied to the paired variables STLX_ARTICOR and STLX_RadiAnt. Top: hypothesis test summary. Bottom: frequencies of positive and negative differences.*

The Mann-Whitney U test referring to S-TLX indicated a statistically significant difference (p=0.01) in perceived workload between group 0, which perceived a lower workload score when using ARTICOR (V2<0), and group 1, which instead perceived a lower workload when using RadiAnt (V2>0) (Figure 4.4). It is worth noting that the differential workload score has been considered instead of the absolute scores individually. Since for one participant the differential score was equal to 0, the total number of observations included in the Mann-Whitney U test was 15.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.^{a,b}	Decision
1	The distribution of STLX_ARTICOR - STLX_RadiAnt is the same across categories of GROUPS.	Independent-Samples Mann-Whitney U Test	<,001^{c}	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

c. Exact significance is displayed for this test.

Test Statistics^{a}

	STLX_ARTICOR - STLX_RadiAnt
Mann-Whitney U	,000
Wilcoxon W	36,000
Z	-3,240
Asymp. Sig. (2-tailed)	,001
Exact Sig. [2*(1-tailed Sig.)]	<,001^{b}

a. Grouping Variable: GROUPS

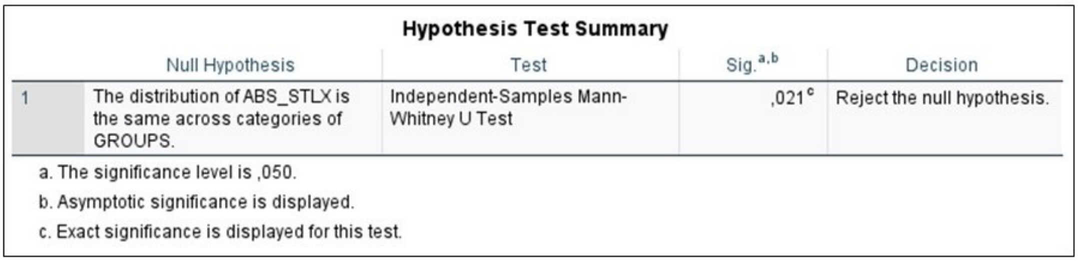
b. Not corrected for ties.

Ranks

	GROUPS	N	Mean Rank	Sum of Ranks
STLX_ARTICOR - STLX_RadiAnt	0	8	4,50	36,00
	1	7	12,00	84,00
	Total	15		

Figure 4.4: *Results of the Mann-Whitney U test applied to the variable STLX_ ARTICOR - STLX_ RadiAnt. Top: hypothesis test summary; middle left: test statistics; bottom left: ranks; right: frequency distribution of the variable over the two groups.*

The same considerations can be applied to the variable `ABS_STLX`, representing the absolute value of the variable V2 = `STLX_ARTICOR - STLX_RadiAnt`: a statistically significant difference (p=0.021) was observed between the two groups (Figure 4.5).
This analysis suggests that the absolute value of the difference in the workload perceived when using ARTICOR with respect to the one perceived when using RadiAnt is significantly higher for the newbies' group (group 0). It means that participants in group 0 perceived much more the advantages of the MR platform in terms of workload reduction. The expert group (group 1) instead rated in a similar way physical, mental, temporal demand, situational stress, distraction, and task complexity perceived with the two technologies. It means that, despite being in favour of the standard technology, they confirmed that the workload perceived when using the MR platform for the first time was comparable to the one perceived when using the standard software after several years of experience.



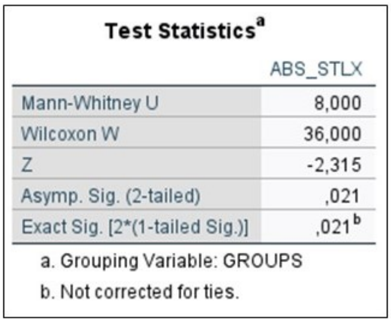
Hypothesis Test Summary

	Null Hypothesis	Test	Sig.^{a,b}	Decision
1	The distribution of ABS_STLX is the same across categories of GROUPS.	Independent-Samples Mann-Whitney U Test	,021^{c}	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

c. Exact significance is displayed for this test.

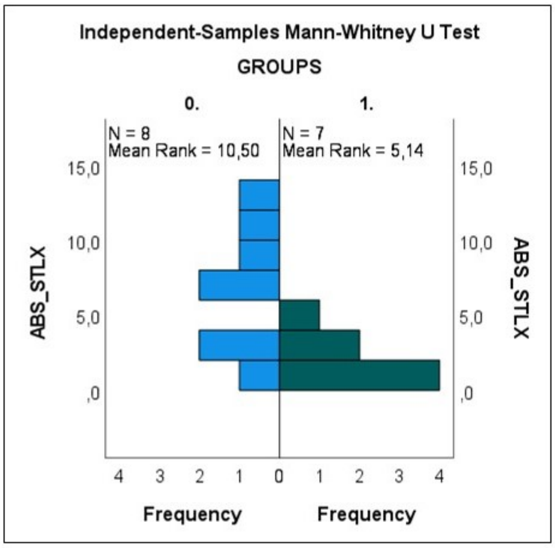


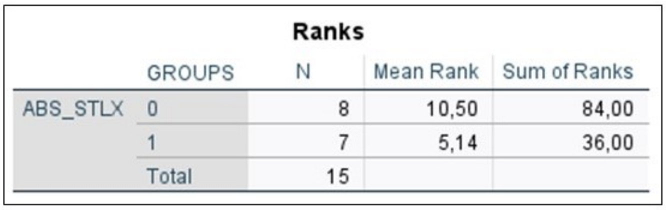
Test Statistics^{a}

	ABS_STLX
Mann-Whitney U	8,000
Wilcoxon W	36,000
Z	-2,315
Asymp. Sig. (2-tailed)	,021
Exact Sig. [2*(1-tailed Sig.)]	,021^{b}

a. Grouping Variable: GROUPS

b. Not corrected for ties.





Ranks

	GROUPS	N	Mean Rank	Sum of Ranks
ABS_STLX	0	8	10,50	84,00
	1	7	5,14	36,00
	Total	15		

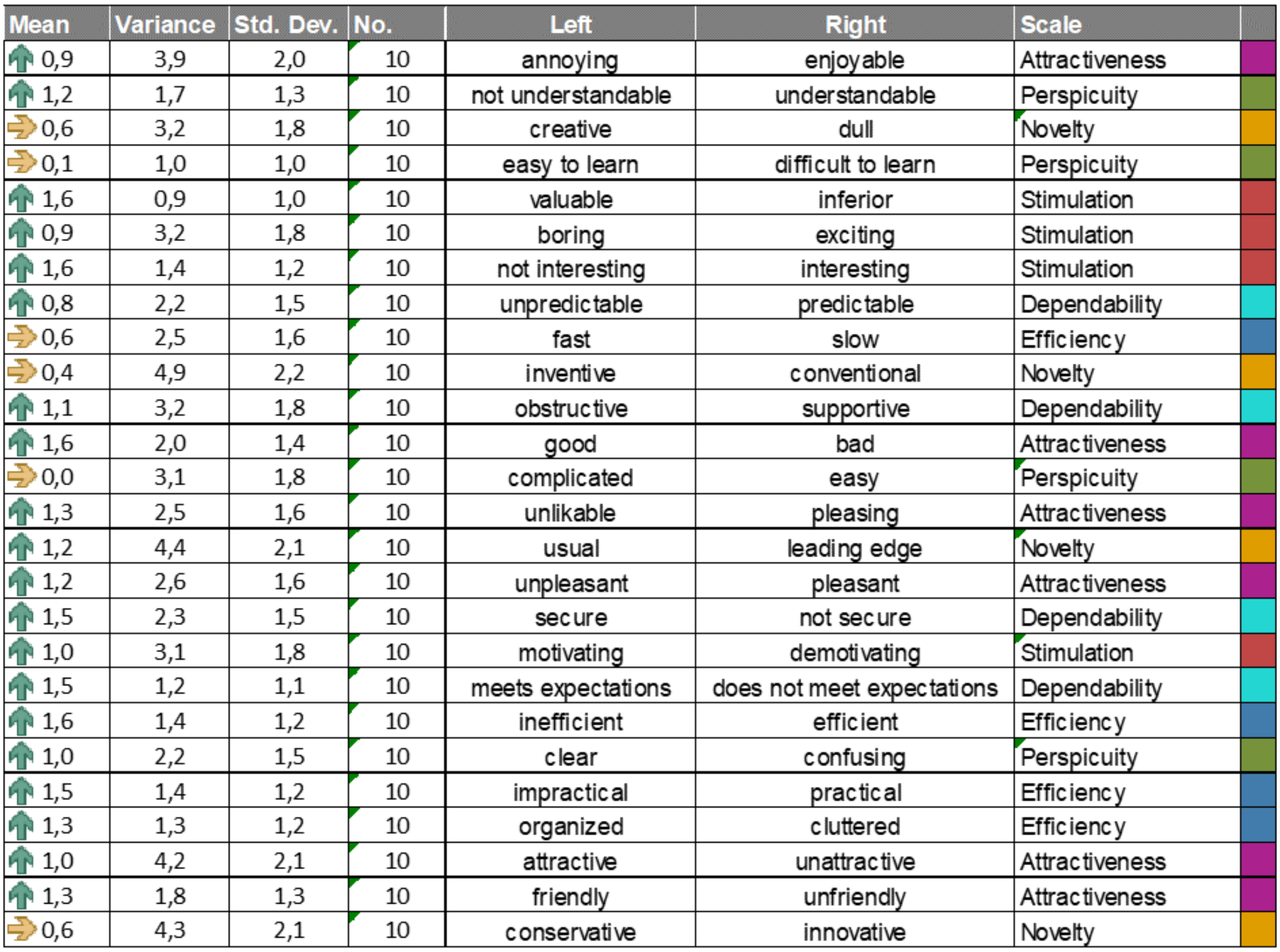
Figure 4.5: *Results of the Mann-Whitney U test applied to the variable ABS(STLX_ARTICOR - STLX_RadiAnt). Top: hypothesis test summary; middle left: test statistics; bottom left: ranks; right: frequency distribution of the variable over the two groups.*

4.1.4. User Experience Questionnaire

For what concerns the UEQ referring to RadiAnt, the first observation is the presence of several missing data since only twelve participants over sixteen filled in this questionnaire. Moreover, two of them were a posteriori excluded from the analysis since more than 3 inconsistent scales resulted from their responses. Among the remaining ten respondents, five of them belonged to group 1 (experts in using the viewer software) and the other five belonged to group 0 (newbies in using it). As a consequence, results from the UEQ suffer from the limited numerosity of gathered data, but are supposedly free from bias associated to the different levels of experience of the participants.

Considering the ten respondents, the mean and standard deviation of the 26 items are shown in Figure 4.6.
Taking into account that the different opinions of the raters and the answer tendencies

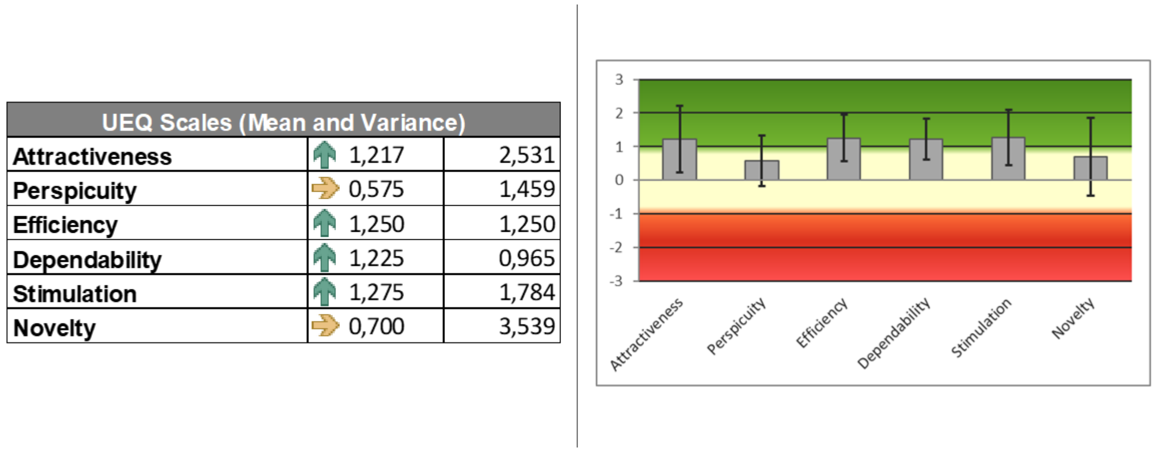
typically prevent extreme answers, in real applications it is unlikely to observe values above +2 or below -2. Therefore, mean values higher than +0.8 are considered as positive scores and mean values close to +1.5 as very positive. In this case, no red arrows indicating a negative mean score (i.e., < -0.8) are present and 20 items over 26 obtained an average positive score, i.e. > 0.8 (green arrows).



Mean	Variance	Std. Dev.	No.	Left	Right	Scale	
⇧ 0,9	3,9	2,0	10	annoying	enjoyable	Attractiveness	🟪
⇧ 1,2	1,7	1,3	10	not understandable	understandable	Perspicuity	🟩
⇨ 0,6	3,2	1,8	10	creative	dull	Novelty	🟧
⇨ 0,1	1,0	1,0	10	easy to learn	difficult to learn	Perspicuity	🟩
⇧ 1,6	0,9	1,0	10	valuable	inferior	Stimulation	🟥
⇧ 0,9	3,2	1,8	10	boring	exciting	Stimulation	🟥
⇧ 1,6	1,4	1,2	10	not interesting	interesting	Stimulation	🟥
⇧ 0,8	2,2	1,5	10	unpredictable	predictable	Dependability	🟦
⇨ 0,6	2,5	1,6	10	fast	slow	Efficiency	🟦
⇨ 0,4	4,9	2,2	10	inventive	conventional	Novelty	🟧
⇧ 1,1	3,2	1,8	10	obstructive	supportive	Dependability	🟦
⇧ 1,6	2,0	1,4	10	good	bad	Attractiveness	🟪
⇨ 0,0	3,1	1,8	10	complicated	easy	Perspicuity	🟩
⇧ 1,3	2,5	1,6	10	unlikable	pleasing	Attractiveness	🟪
⇧ 1,2	4,4	2,1	10	usual	leading edge	Novelty	🟧
⇧ 1,2	2,6	1,6	10	unpleasant	pleasant	Attractiveness	🟪
⇧ 1,5	2,3	1,5	10	secure	not secure	Dependability	🟦
⇧ 1,0	3,1	1,8	10	motivating	demotivating	Stimulation	🟥
⇧ 1,5	1,2	1,1	10	meets expectations	does not meet expectations	Dependability	🟦
⇧ 1,6	1,4	1,2	10	inefficient	efficient	Efficiency	🟦
⇧ 1,0	2,2	1,5	10	clear	confusing	Perspicuity	🟩
⇧ 1,5	1,4	1,2	10	impractical	practical	Efficiency	🟦
⇧ 1,3	1,3	1,2	10	organized	cluttered	Efficiency	🟦
⇧ 1,0	4,2	2,1	10	attractive	unattractive	Attractiveness	🟪
⇧ 1,3	1,8	1,3	10	friendly	unfriendly	Attractiveness	🟪
⇨ 0,6	4,3	2,1	10	conservative	innovative	Novelty	🟧

Figure 4.6: *Items' mean and standard deviation from the User Experience Questionnaire concerning RadiAnt viewer software. Green upwards arrow = mean > 0.8; yellow horizontal arrow = -0.8 < mean < 0.8; red downwards arrow = mean < -0.8. Columns "left" and "right" contain the adjectives that constitute the relative item on the UEQ.*

For what concerns the six scales in which the 26 items can be aggregated, it is possible to observe that all the mean values are quite close to a neutral evaluation. Even if no one of them reached a negative score, this result shows a lack of enthusiasm for the technology. In particular, the lowest means are the ones relative to the *perspicuity* and *novelty* dimensions, meaning that the users did not find RadiAnt easy-to-learn nor innovative and creative. The plots in Figure 4.7 are two ways of displaying this information.



UEQ Scales (Mean and Variance)		
Attractiveness	⬆ 1,217	2,531
Perspicuity	➡ 0,575	1,459
Efficiency	⬆ 1,250	1,250
Dependability	⬆ 1,225	0,965
Stimulation	⬆ 1,275	1,784
Novelty	➡ 0,700	3,539

Figure 4.7: *Scales' mean and standard deviation from the User Experience Questionnaire concerning RadiAnt. Left: numerical value and colour code (green upwards arrow = mean > 0.8; yellow horizontal arrow = -0.8 < mean < 0.8; red downwards arrow = mean < -0.8). Right: bar plots representing the mean and black line representing the confidence interval of each scale.*

The conclusions derived from the "Benchmark" spreadsheet are shown in Figure 4.8. Only the *perspicuity* dimension is classified as "bad", i.e., in the range of 25% worst results, meaning that one of the major problems for users was the difficulty in learning how to use the software, which is not intuitive nor straightforward. In this context, it is important to consider the fact that the time participants were allowed to interact with the technology during the test was very limited: in case of no previous experience, difficulties are understandable.

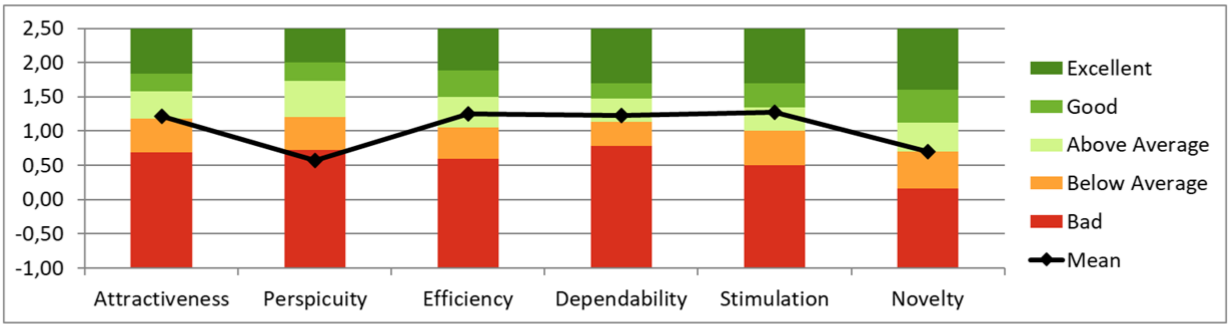
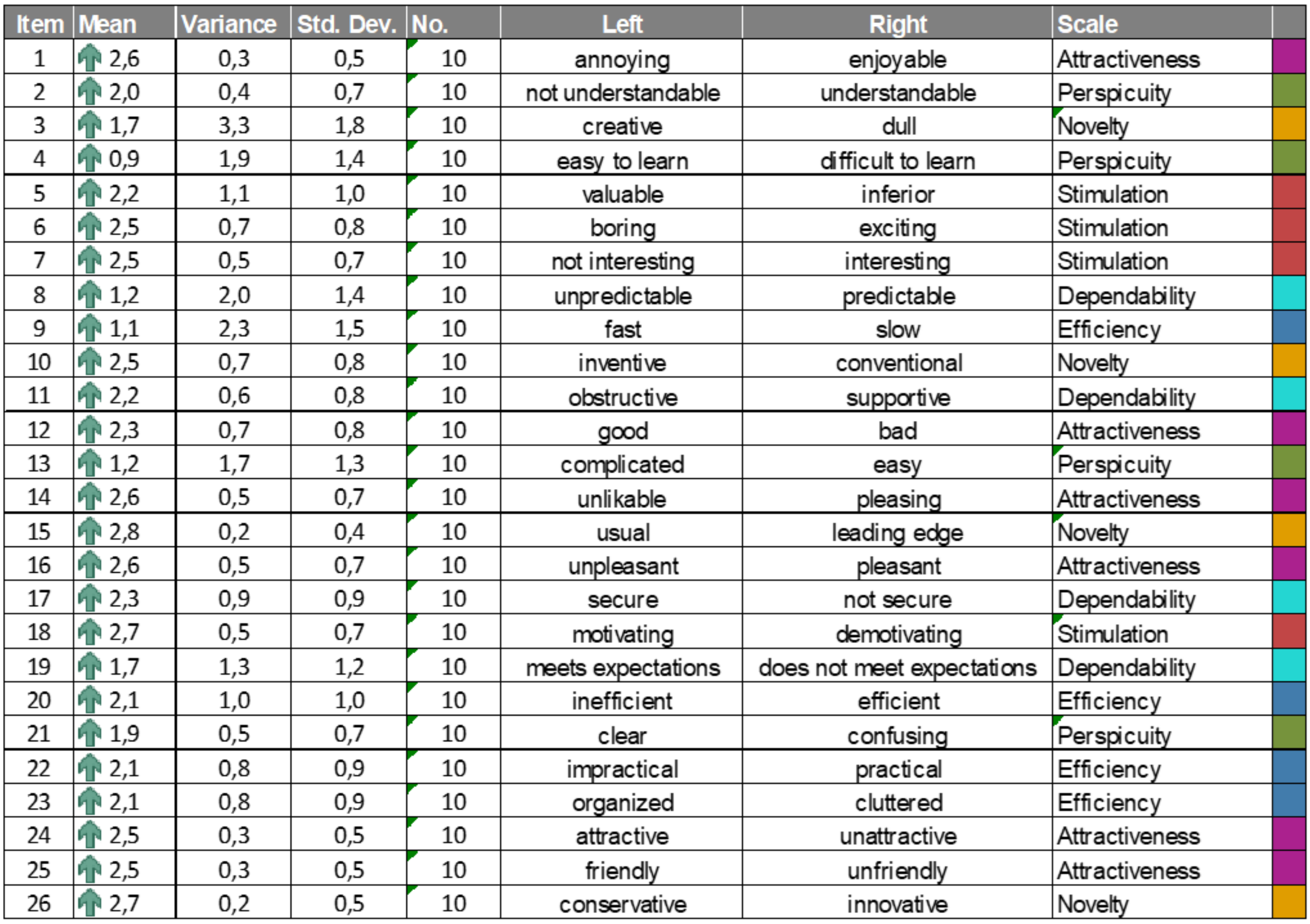


Figure 4.8: *Results from the benchmark spreadsheet concerning RadiAnt viewer software. The black line represents the performance of the product under evaluation on the 6 evaluation scales with respect to the benchmark dataset.*

For what concerns the analysis of the questionnaire relative to ARTICOR, no data were missing and only one inconsistency was detected. Therefore, fifteen responses could be analyzed (n=6 from participants of group 1, n=9 from participants of group 0). However, in order to perform a coherent comparison with the results obtained from the UEQ referring to RadiAnt, I decided to consider only the answers provided by the same ten participants who answered the latter.

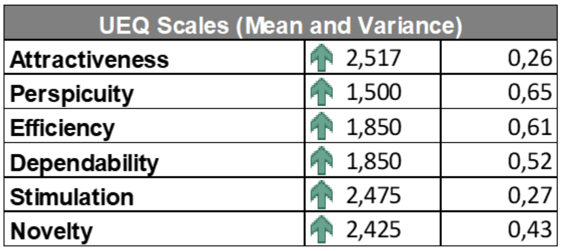
Again, the mean and standard deviation of each dimension are shown in Figure 4.9. From the colour of the arrows, an improvement in the performance of ARTICOR with respect to RadiAnt is evident, with positive scores in all 26 dimensions. Moreover, many of these resulted in a mean value higher than 2, which is a very positive score considering the answer tendency that typically prevents extreme answers (i.e., close to -3, +3).



Item	Mean	Variance	Std. Dev.	No.	Left	Right	Scale	
1	↑ 2,6	0,3	0,5	10	annoying	enjoyable	Attractiveness	🟪
2	↑ 2,0	0,4	0,7	10	not understandable	understandable	Perspicuity	🟩
3	↑ 1,7	3,3	1,8	10	creative	dull	Novelty	🟧
4	↑ 0,9	1,9	1,4	10	easy to learn	difficult to learn	Perspicuity	🟩
5	↑ 2,2	1,1	1,0	10	valuable	inferior	Stimulation	🟥
6	↑ 2,5	0,7	0,8	10	boring	exciting	Stimulation	🟥
7	↑ 2,5	0,5	0,7	10	not interesting	interesting	Stimulation	🟥
8	↑ 1,2	2,0	1,4	10	unpredictable	predictable	Dependability	🟦
9	↑ 1,1	2,3	1,5	10	fast	slow	Efficiency	🟦
10	↑ 2,5	0,7	0,8	10	inventive	conventional	Novelty	🟧
11	↑ 2,2	0,6	0,8	10	obstructive	supportive	Dependability	🟦
12	↑ 2,3	0,7	0,8	10	good	bad	Attractiveness	🟪
13	↑ 1,2	1,7	1,3	10	complicated	easy	Perspicuity	🟩
14	↑ 2,6	0,5	0,7	10	unlikable	pleasing	Attractiveness	🟪
15	↑ 2,8	0,2	0,4	10	usual	leading edge	Novelty	🟧
16	↑ 2,6	0,5	0,7	10	unpleasant	pleasant	Attractiveness	🟪
17	↑ 2,3	0,9	0,9	10	secure	not secure	Dependability	🟦
18	↑ 2,7	0,5	0,7	10	motivating	demotivating	Stimulation	🟥
19	↑ 1,7	1,3	1,2	10	meets expectations	does not meet expectations	Dependability	🟦
20	↑ 2,1	1,0	1,0	10	inefficient	efficient	Efficiency	🟦
21	↑ 1,9	0,5	0,7	10	clear	confusing	Perspicuity	🟩
22	↑ 2,1	0,8	0,9	10	impractical	practical	Efficiency	🟦
23	↑ 2,1	0,8	0,9	10	organized	cluttered	Efficiency	🟦
24	↑ 2,5	0,3	0,5	10	attractive	unattractive	Attractiveness	🟪
25	↑ 2,5	0,3	0,5	10	friendly	unfriendly	Attractiveness	🟪
26	↑ 2,7	0,2	0,5	10	conservative	innovative	Novelty	🟧

Figure 4.9: *Items' mean and standard deviation from the User Experience Questionnaire concerning ARTICOR. Green upwards arrow = mean > 0.8; yellow horizontal arrow = -0.8 < mean < 0.8; red downwards arrow = mean < -0.8. Columns "left" and "right" contain the adjectives that constitute the relative item on the UEQ.*

Consequently, also the analysis of the means of each scale reports very positive results (Figure 4.10). In particular, the scales of *attractiveness*, *stimulation*, and *novelty* registered a mean close to the maximum score, thus indicating that the users had a very positive general impression of the MR platform and appreciated especially its motivation capability, innovation, and creativity.



UEQ Scales (Mean and Variance)		
Attractiveness	↑ 2,517	0,26
Perspicuity	↑ 1,500	0,65
Efficiency	↑ 1,850	0,61
Dependability	↑ 1,850	0,52
Stimulation	↑ 2,475	0,27
Novelty	↑ 2,425	0,43

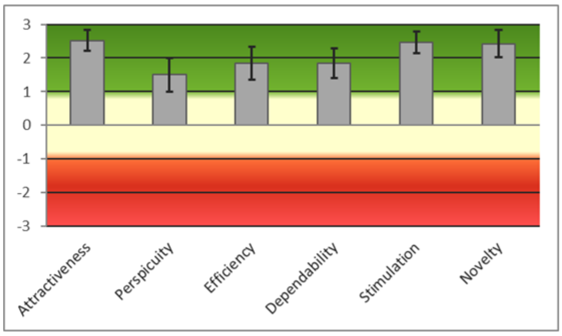


Figure 4.10: *Scales' mean and standard deviation from the User Experience Questionnaire concerning ARTICOR. Left: numerical value and colour code (green upwards arrow = mean > 0.8; yellow horizontal arrow = -0.8 < mean < 0.8; red downwards arrow = mean < -0.8). Right: bar plots representing the mean and black line representing the confidence interval of each scale.*

Comparing ARTICOR with the benchmark dataset (Figure 4.11), it is possible to observe that all mean scores are above the average and the technology outperforms. In particular, *attractiveness*, *dependability*, *stimulation*, and *novelty* dimensions are classified as "excellent", i.e., in the range of the 10% best results. The product induced a particularly positive overall feeling, the users felt completely in control of the interaction, they were very motivated in using the technology, and greatly appreciated its creativity and innovativeness.

The scale which reached the lowest score, which is however above the average, is *perspicuity*. The users encountered some difficulties in learning how to use the technology but, considering that most of them did not have any previous experience nor knowledge about MR working principle and that the interaction time was very short, this can be considered a very positive result. Moreover, comparing the perspicuity scores of ARTICOR and RadiAnt, the former was judged notably more intuitive and easy to use. These features make the MR technology a potential valuable competitor of current technologies.

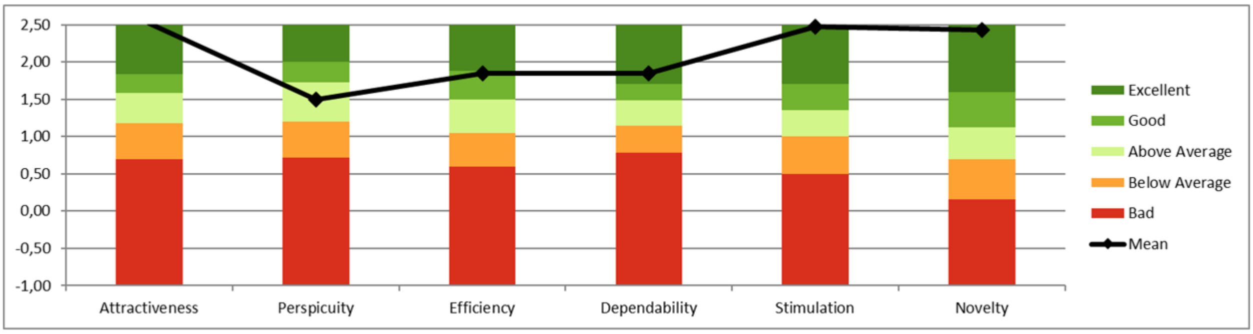


Figure 4.11: *Results from the benchmark spreadsheet concerning ARTICOR. The black line represents the performance of the product under evaluation on the 6 evaluation scales with respect to the benchmark dataset.*

4.1.5. System Usability Scale

The total usability scores obtained by the participants are summarized in Tables 4.5 and 4.6.

All but one participant filled in the SUS questionnaire concerning ARTICOR while 4 missing data resulted in the RadiAnt questionnaire.

Table 4.5: *Results from the SUS questionnaire concerning RadiAnt: contribution of the 10 items and total score for the 16 participants.*

	ITEM										
PARTICIPANT	1	2	3	4	5	6	7	8	9	10	**TOTAL SCORE**
PART1 (0)											/
PART2 (0)	2	4	2	3	2	3	2	1	2	4	**37,5**
PART3 (1)	5	1	5	2	5	2	4	2	4	3	**82,5**
PART4 (1)	5	2	3	5	3	1	3	2	4	3	**62,5**
PART5 (0)											/
PART6 (0)	2	4	3	4	3	4	3	2	2	4	**37,5**
PART7 (1)	4	2	4	2	4	2	5	2	4	3	**75**
PART8 (1)											/
PART9 (1)	4	2	3	3	4	2	4	1	4	3	**70**
PART10 (1)	4	2	3	2	4	1	3	1	3	3	**70**
PART11 (0)	3	4	2	5	3	2	2	2	2	5	**35**
PART12 (0)	2	4	2	4	1	1	1	4	2	4	**27,5**
PART13 (0)	4	4	3	3	4	2	3	1	5	4	**62,5**
PART14 (0)	5	1	5	3	3	4	4	2	5	3	**72,5**
PART15 (0)	3	3	2	4	3	2	4	3	3	3	**50**
PART16 (1)											/

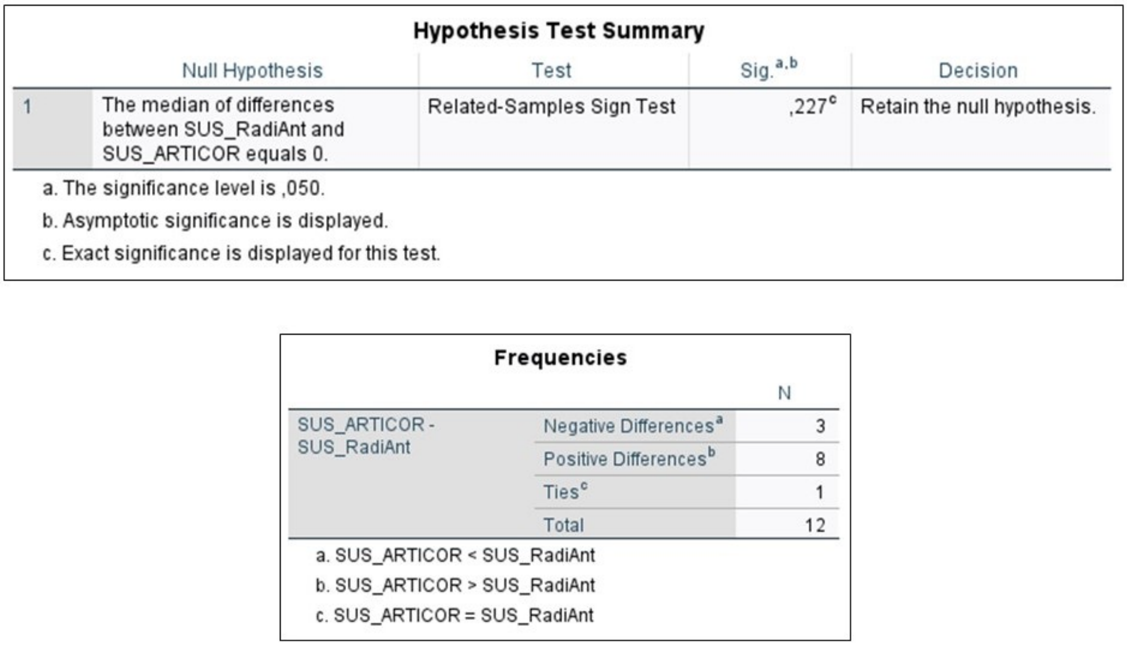
Table 4.6: *Results from the SUS questionnaire concerning ARTICOR: contribution of the 10 items and total score for the 16 participants.*

	ITEM										
PARTICIPANT	1	2	3	4	5	6	7	8	9	10	**TOTAL**
PART1 (0)											/
PART2 (0)	4	1	4	2	3	1	3	3	4	3	70
PART3 (1)	5	1	4	3	4	1	4	1	5	2	85
PART4 (1)	5	2	3	5	4	2	4	3	3	3	60
PART5 (0)	4	3	3	3	3	3	3	2	3	5	50
PART6 (0)	5	2	5	3	5	3	5	3	4	3	75
PART7 (1)	3	1	4	2	4	1	4	1	3	3	75
PART8 (1)	3	2	4	3	5	1	5	1	5	4	77,5
PART9 (1)	4	2	4	4	4	2	3	2	3	3	62,5
PART10 (1)	4	2	3	4	4	2	3	2	2	3	57,5
PART11 (0)	5	1	5	3	4	2	5	1	5	3	85
PART12 (0)	5	2	4	2	4	1	5	2	3	3	77,5
PART13 (0)	5	2	4	4	4	2	5	1	5	3	77,5
PART14 (0)	4	2	5	5	5	2	5	1	5	3	77,5
PART15 (0)	4	2	4	2	4	2	4	2	4	2	75
PART16 (1)	3	3	2	2	2	2	3	2	1	2	50

The same statistical analysis performed for the S-TLX score was applied to the SUS score. Since four people did not fill in the SUS questionnaire concerning RadiAnt, the answers of twelve participants were evaluated; out of these, seven belonged to group 0 and five belonged to group 1.

From the Paired-sample Sign test applied to the two paired variables `SUS_ARTICOR` and

`SUS_RadiAnt`, no statistical difference was found between the usability scores obtained by ARTICOR and RadiAnt (p = 0.227). Again, this result may be due to the heterogeneity of the participant sample, who expressed contrasting opinions concerning the usability of the two technologies. However, also in this case, the analysis of the frequencies reveals that feedbacks were not completely neutral: the higher number of positive frequencies for the variable V1 = `SUS_ARTICOR - SUS_RadiAnt` suggests that most of the participants found the MR platform more usable than the comparator (Figure 4.12).



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The median of differences between SUS_RadiAnt and SUS_ARTICOR equals 0.	Related-Samples Sign Test	,227[c]	Retain the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

c. Exact significance is displayed for this test.

Frequencies

		N
SUS_ARTICOR - SUS_RadiAnt	Negative Differences[a]	3
	Positive Differences[b]	8
	Ties[c]	1
	Total	12

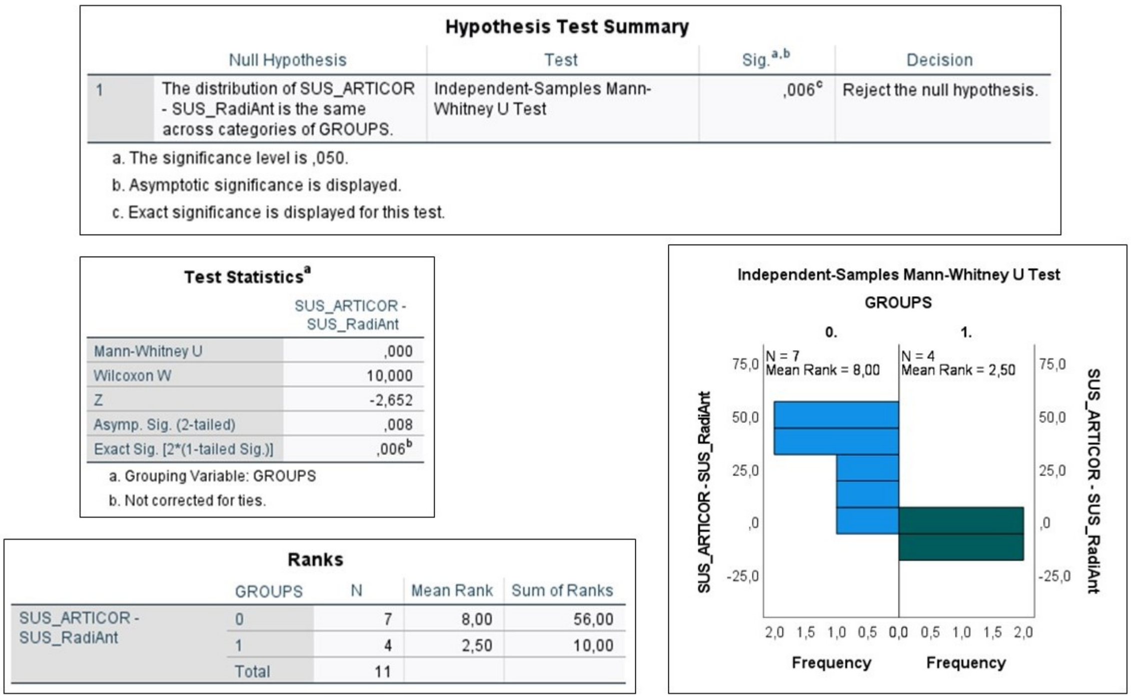
a. SUS_ARTICOR < SUS_RadiAnt

b. SUS_ARTICOR > SUS_RadiAnt

c. SUS_ARTICOR = SUS_RadiAnt

Figure 4.12: *Result of the Paired-sample Sign test applied to the paired variables SUS_ ARTICOR and SUS_ RadiAnt. Top: hypothesis test summary. Bottom: frequencies of positive and negative differences.*

Subsequently, the Mann-Whitney non-parametric test was applied to the dependent variable V1 to detect whether statistical differences exist between the SUS scores obtained by the two groups in which the sample population had been divided (experts and newbies in using DICOM viewer software). Again, it is worth noting that instead of considering the absolute SUS scores attributed to the two technologies separately, the differential usability score was considered. The total number of observations considered in this analysis is eleven instead of twelve since, for one participant, the differential score was equal to 0. As for the variable V2 = `STLX_ARTICOR - STLX_RadiAnt`, a statistically significant difference (p = 0.006) was found between the two groups (Figure 4.13). In particular, V1 was > 0 for group 0, meaning that the members perceived ARTICOR as more usable with respect to RadiAnt, and was <0 for group 1, which instead found RadiAnt more usable than ARTICOR.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of SUS_ARTICOR - SUS_RadiAnt is the same across categories of GROUPS.	Independent-Samples Mann-Whitney U Test	,006[c]	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

c. Exact significance is displayed for this test.

Test Statistics[a]

	SUS_ARTICOR - SUS_RadiAnt
Mann-Whitney U	,000
Wilcoxon W	10,000
Z	-2,652
Asymp. Sig. (2-tailed)	,008
Exact Sig. [2*(1-tailed Sig.)]	,006[b]

a. Grouping Variable: GROUPS

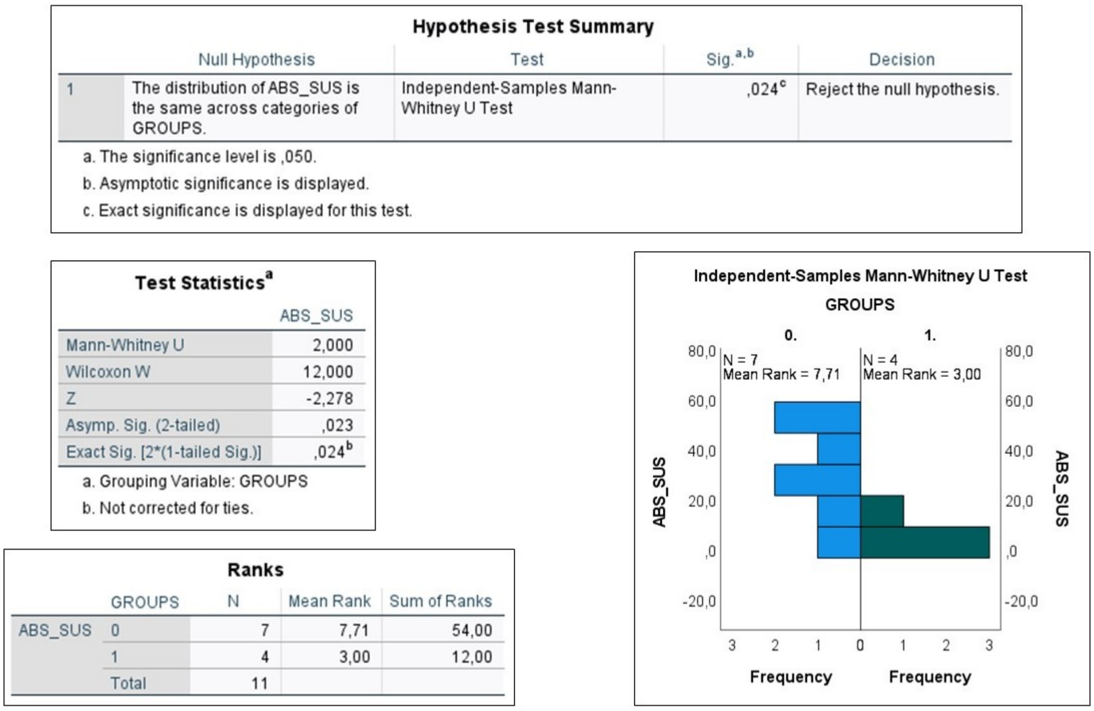
b. Not corrected for ties.

Ranks

	GROUPS	N	Mean Rank	Sum of Ranks
SUS_ARTICOR - SUS_RadiAnt	0	7	8,00	56,00
	1	4	2,50	10,00
	Total	11		

Figure 4.13: *Results of the Mann-Whitney U test applied to the variables SUS_ ARTICOR - SUS_ RadiAnt. Top: hypothesis test summary; middle left: test statistics; bottom left: ranks; right: frequency distribution of the variable over the two groups.*

The Mann-Whitney U test was also applied to the absolute value of the difference between the usability scores (Figure 4.14). Even in this case, as for the variable ABS(`STLX_ARTICOR - STLX_RadiAnt`), a statistically significant difference ($p = 0.024$) was found between the two groups. The participants in group 0 rated the usability of ARTICOR way better than the usability of Radiant, while those in group 1 rated the two technologies with more similar scores. This means that, despite their previous experience with the traditional software which lead them to prefer this solution, they rated the usability of ARTICOR as comparable to the one of RadiAnt.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.^{a,b}	Decision
1	The distribution of ABS_SUS is the same across categories of GROUPS.	Independent-Samples Mann-Whitney U Test	,024^{c}	Reject the null hypothesis.

a. The significance level is ,050.
b. Asymptotic significance is displayed.
c. Exact significance is displayed for this test.

Test Statistics[a]

	ABS_SUS
Mann-Whitney U	2,000
Wilcoxon W	12,000
Z	-2,278
Asymp. Sig. (2-tailed)	,023
Exact Sig. [2*(1-tailed Sig.)]	,024^{b}

a. Grouping Variable: GROUPS
b. Not corrected for ties.

Ranks

	GROUPS	N	Mean Rank	Sum of Ranks
ABS_SUS	0	7	7,71	54,00
	1	4	3,00	12,00
	Total	11		

Figure 4.14: *Results of the Mann-Whitney U test applied to the variable ABS(SUS_ARTICOR - SUS_DICOM). Top: hypothesis test summary; middle left: test statistics; bottom left: ranks; right: frequency distribution of the variable over the two groups.*

Table 4.7 shows the results obtained through the "promoters/detractors" interpretation method.

At first glance, the obtained results seem to be not very optimistic: for both technologies, no "promoters" were present among the respondents and most of them fell in the "passive" category. If instead a comparative analysis is performed, it is possible to draw more optimistic conclusions relatively to ARTICOR since the percentage of users who would likely discourage the use of ARTICOR is lower than the percentage of those who would discourage using Radiant. In particular, 5/12 participants were "detractors" with respect to RadiAnt software while only 3/15 with respect to ARTICOR.

Five participants from group 0, i.e., newbies in both technologies, resulted to be more prone to promote the use of the innovative technology, being "passive" toward ARTICOR and "detractor" toward RadiAnt. Only one participant was "detractor" to ARTICOR and "passive" to the gold standard. Not surprisingly, this participant belonged to group 1, i.e., he/she was an expert in using DICOM viewer software. The remaining participants were "passive" to both technologies.

Table 4.7: *SUS score, LTR score, and "passive/detractor" classification of the participants with respect to the two technologies.*

	RadiAnt			ARTICOR		
PARTICIPANT	**SUS**	**NPS**	**BEHAVIOUR**	**SUS**	**NPS**	**BEHAVIOUR**
PART1 (0)	\	\		**70**	6,93	PASSIVE
PART2 (0)	**37,5**	4,33	DETRACTOR	**85**	8,13	PASSIVE
PART3 (1)	**82,5**	7,93	PASSIVE	**60**	6,13	PASSIVE
PART4 (1)	**62,5**	6,33	PASSIVE	**50**	5,33	DETRACTOR
PART5 (0)	\	\		**75**	7,33	PASSIVE
PART6 (0)	**37,5**	4,33	DETRACTOR	**75**	7,33	PASSIVE
PART7 (1)	**75**	7,33	PASSIVE	**77,5**	7,53	PASSIVE
PART8 (1)	\	\		**62,5**	6,33	PASSIVE
PART9 (1)	**70**	6,93	PASSIVE	**57,5**	5,93	DETRACTOR
PART10 (1)	**70**	6,93	PASSIVE	**85**	8,13	PASSIVE
PART11 (0)	**35**	4,13	DETRACTOR	**77,5**	7,53	PASSIVE
PART12 (0)	**27,5**	3,53	DETRACTOR	**77,5**	7,53	PASSIVE
PART13 (0)	**62,5**	6,33	PASSIVE	**77,5**	7,53	PASSIVE
PART14 (0)	**72,5**	7,13	PASSIVE	**75**	7,33	PASSIVE
PART15 (0)	**50**	5,33	DETRACTOR	**50**	5,33	DETRACTOR
PART16 (1)	\	\		\	\	

The interpretation of these results should account for three aspects.

First, only participants whose answers generate a very high SUS score can be classified as "promoters". Considering that the 50° percentile of the distribution of SUS scores corresponds to 68, scores above this value are above the average. However, a SUS score higher than 90 is needed to be classified as a "promoter" according to the regression equation.

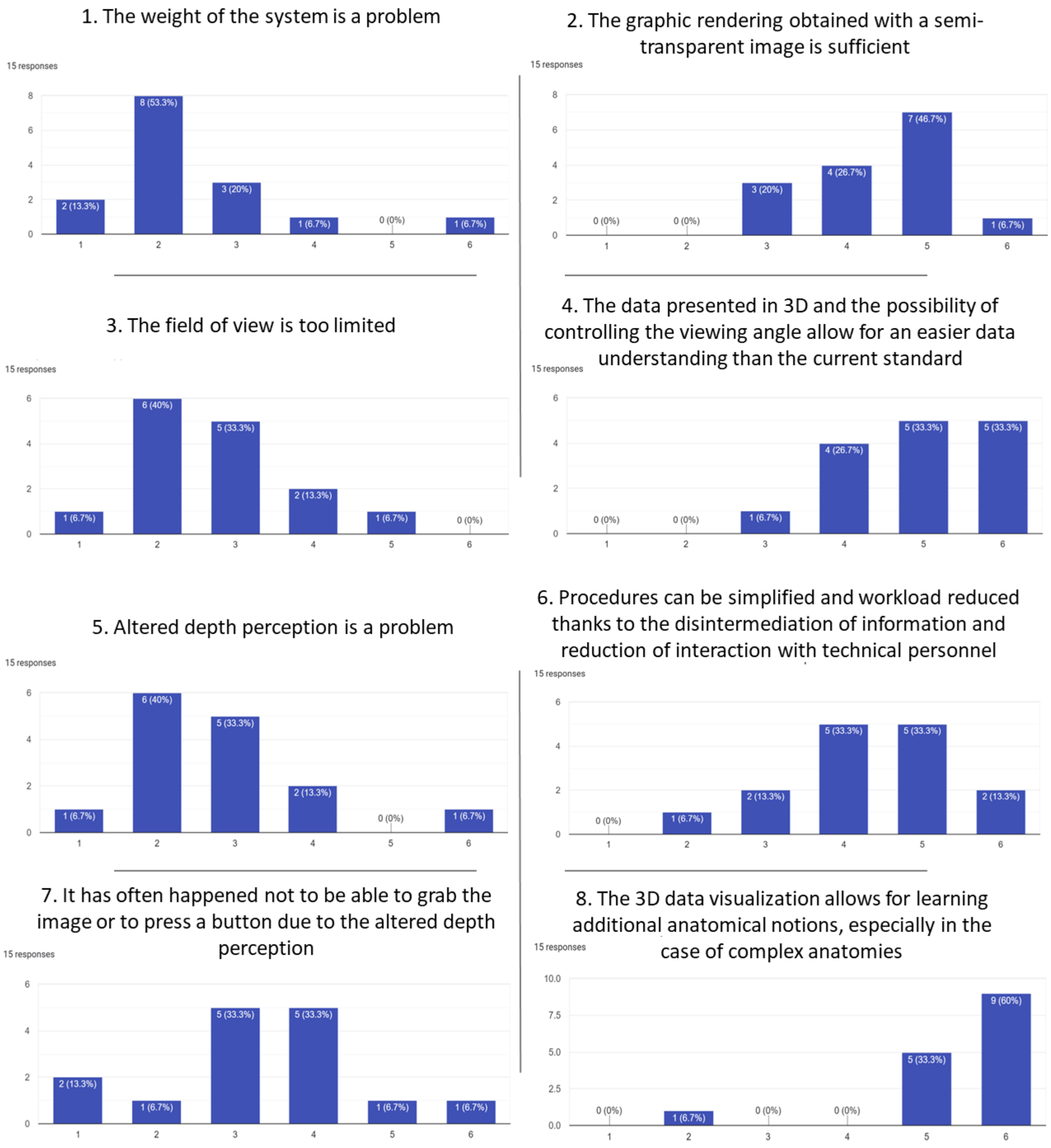
Second, participants were almost newbies to the MR technology under evaluation since

they had no previous experience, and they hardly knew its features and functions. During the test phase, they could not get used to the MR technology because they could interact with it for a limited amount of time.
Third, the application of the LTR score in Italy suffers from a cultural limitation since the Italian population is not prone to attribute the maximum score in a rating exercise. It means that, even if very satisfied with a product and available to promote it, it is difficult that users would attribute a LTR score higher than 9, thus falling in the "passive" category instead of in the "promoters" one.
These facts, the novelty of the technology, and the fact that it is designed to replace an already efficient one, clearly make it difficult to find promoters among the experimental sample. In this perspective, the fact that most participants were "passive" and not "detractors" with respect to ARTICOR still suggests that most of them perceived it in a positive way.

4.1.6. Ad hoc developed questionnaire

One of the participants did not fill in the ad hoc developed questionnaire. The remaining fifteen answers to each question were analyzed by building bar plots through the analytic tools available in Google Forms (Figure 4.15).

The bar plots in Figure 4.15 report the results obtained for the eight questions in the questionnaire: the distribution of the answers to statements 1, 3, 5, which were formulated to state something negative about the usability of the MR technology, resulted in being shifted toward the left. The distribution of the answers to statements 2, 4, 6, and 8, which were formulated to state something positive about the usability of the MR technology, resulted in being shifted toward the right. In other words, participants mostly disagreed with negative statements and agreed with positive ones, meaning that the usability of ARTICOR was positively evaluated. The only exception to this general remark consists in the answers to statement 7, which focused on failure in grabbing the image or in pressing a button because of altered depth perception. Twelve out of fifteen participants experienced this failure more than twice, meaning that the alteration of depth perception could generate most of the issues during real-case use.



1. The weight of the system is a problem

2. The graphic rendering obtained with a semi-transparent image is sufficient

3. The field of view is too limited

4. The data presented in 3D and the possibility of controlling the viewing angle allow for an easier data understanding than the current standard

5. Altered depth perception is a problem

6. Procedures can be simplified and workload reduced thanks to the disintermediation of information and reduction of interaction with technical personnel

7. It has often happened not to be able to grab the image or to press a button due to the altered depth perception

8. The 3D data visualization allows for learning additional anatomical notions, especially in the case of complex anatomies

Figure 4.15: *Bar plots showing participants' responses to the ad hoc developed questionnaire.*

The total usability score was hence computed as explained in subsection 3.1.4. Results are shown in Table 4.8:

Table 4.8: *Results from the hoc developed questionnaire: contribution of the different items and total usability score.*

	ITEM								
PARTICIPANT	1	2	3	4	5	6	7	8	**TOTAL SCORE**
PART1 (0)									/
PART2 (0)	4	5	4	5	2	4	3	5	**65**
PART3 (1)	1	5	2	6	2	4	2	6	**85**
PART4 (1)	1	3	4	5	4	5	4	6	**65**
PART5 (0)	3	4	3	6	3	6	5	6	**70**
PART6 (0)	2	3	3	6	3	4	4	6	**67,5**
PART7 (1)	2	5	2	4	2	5	3	6	**77,5**
PART8 (1)	6	6	2	5	4	5	4	6	**65**
PART9 (1)	2	5	2	6	3	5	3	6	**80**
PART10 (1)	3	5	2	5	2	4	3	5	**72,5**
PART11 (0)	2	4	1	6	3	6	4	6	**80**
PART12 (0)	2	5	5	5	1	5	1	5	**77,5**
PART13 (0)	2	4	3	4	2	3	3	6	**67,5**
PART14 (0)	2	5	2	4	2	2	6	2	**52,5**
PART15 (0)	2	4	3	4	3	3	1	5	**67,5**
PART16 (1)	3	3	3	3	6	4	4	5	**47,5**

The histogram in figure 4.16 represents the distribution of the usability scores computed from the answers to the ad hoc developed questionnaire. Most of the results are above 60, with a peak between 65 and 70. However, since the tool has been developed and applied for the first time in the context of this study, there are no statistics to compare the scores with and, differently from the SUS scores, a percentile curve to evaluate the goodness of the results is not available. The only conclusion that can be done is that, since the scale of the total scores ranges from 0 to 100, most of the results are above the middle of the scale.

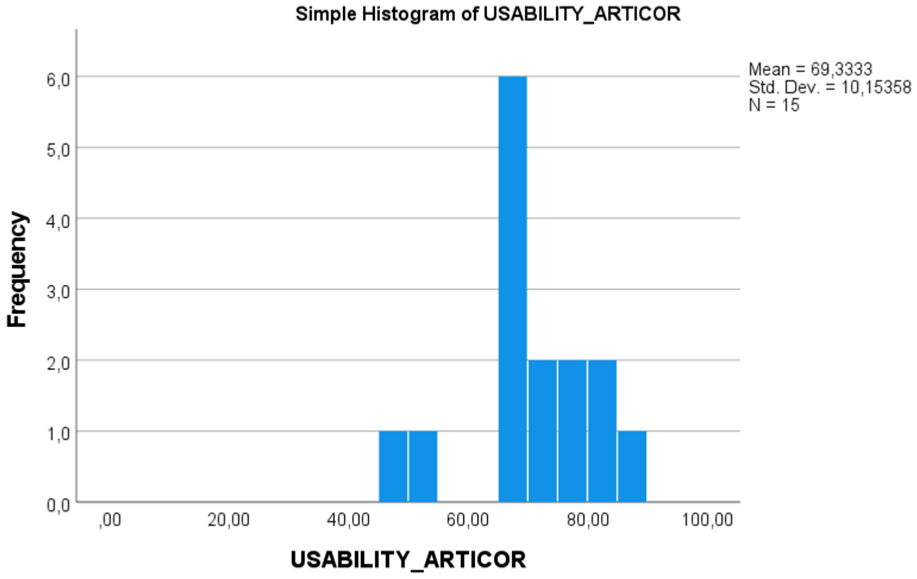
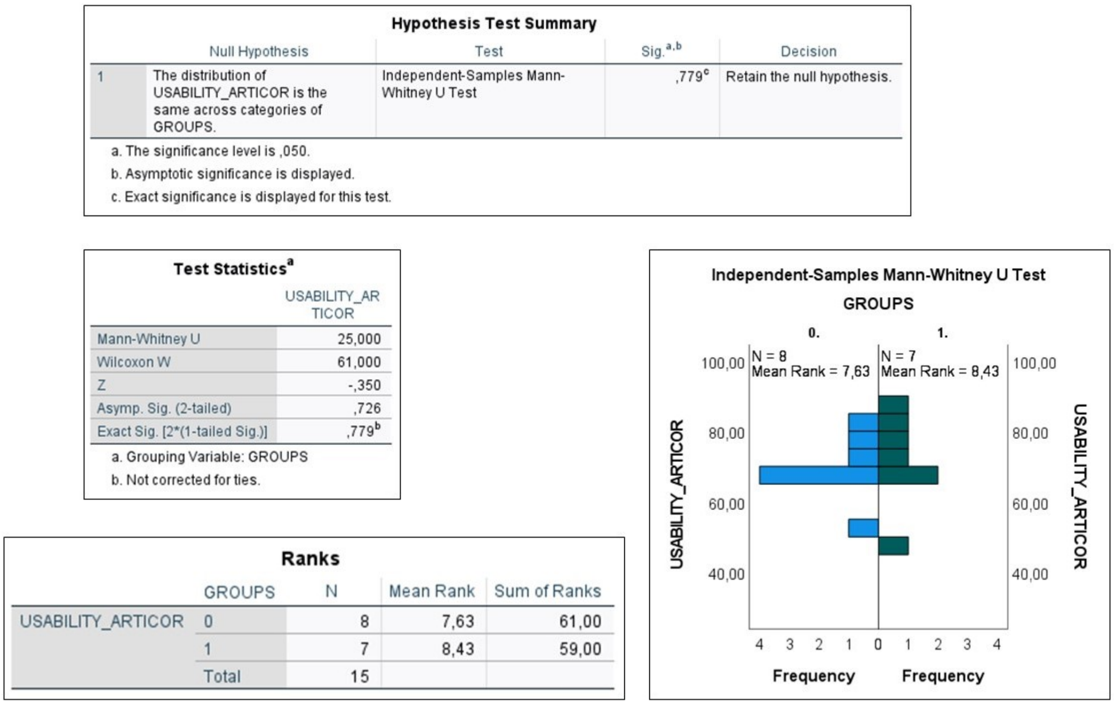


Figure 4.16: *Histogram of the distribution of the ARTICOR usability score obtained from the ad hoc developed questionnaire.*

Since the questionnaire was developed with reference to ARTICOR, it was not possible to perform a comparative analysis of the usability scores of the two technologies.

Conversely, from the Mann-Whitney U test applied to investigate differences between the answers provided by the participants in group 0 and group 1, no statistically meaningful difference was found (p=0.779) (Figure 4.17). This can be explained by the fact that this questionnaire was provided only with reference to ARTICOR and not to RadiAnt. Hence, the key difference between the two groups, i.e., the level of expertise in the use of standard DICOM viewers, did not significantly impact the answers. This result is encouraging since it suggests that the MR platform is positively judged by both user groups when it is not compared directly with the gold standard.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.^{a,b}	Decision
1	The distribution of USABILITY_ARTICOR is the same across categories of GROUPS.	Independent-Samples Mann-Whitney U Test	,779^{c}	Retain the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

c. Exact significance is displayed for this test.

Test Statistics^{a}

	USABILITY_ARTICOR
Mann-Whitney U	25,000
Wilcoxon W	61,000
Z	-,350
Asymp. Sig. (2-tailed)	,726
Exact Sig. [2*(1-tailed Sig.)]	,779^{b}

a. Grouping Variable: GROUPS

b. Not corrected for ties.

Ranks

	GROUPS	N	Mean Rank	Sum of Ranks
USABILITY_ARTICOR	0	8	7,63	61,00
	1	7	8,43	59,00
	Total	15		

Figure 4.17: *Results of the Mann-Whitney U test applied to the variable USABILITY_ARTICOR. Top: hypothesis test summary; middle left: test statistics; bottom left: ranks; right: frequency distribution of the variable over the two groups.*

The second part of the questionnaire investigated the extent of the symptomatology perceived by the participants when interacting with ARTICOR during the test phase. The evaluated symptoms were: headache, nausea, dizziness, sweating, eyestrain, difficulty in concentrating, and general discomfort.

From the histograms reported in Figure 4.18, it is possible to observe that the severity of all the symptoms was rated below 4 on a scale ranging from 1 to 6 by all the participants but one, who rated "headache" with 4. Most respondents rated all the symptoms with the lowest possible score, meaning that no one perceived them with such severity as to be annoying or problematic. However, it worth pointing out that users interacted with the technology for a short amount of time. Longer interaction may lead to a more significant perception of the same symptoms, and more realistic studies should be carried out to test this possibility.

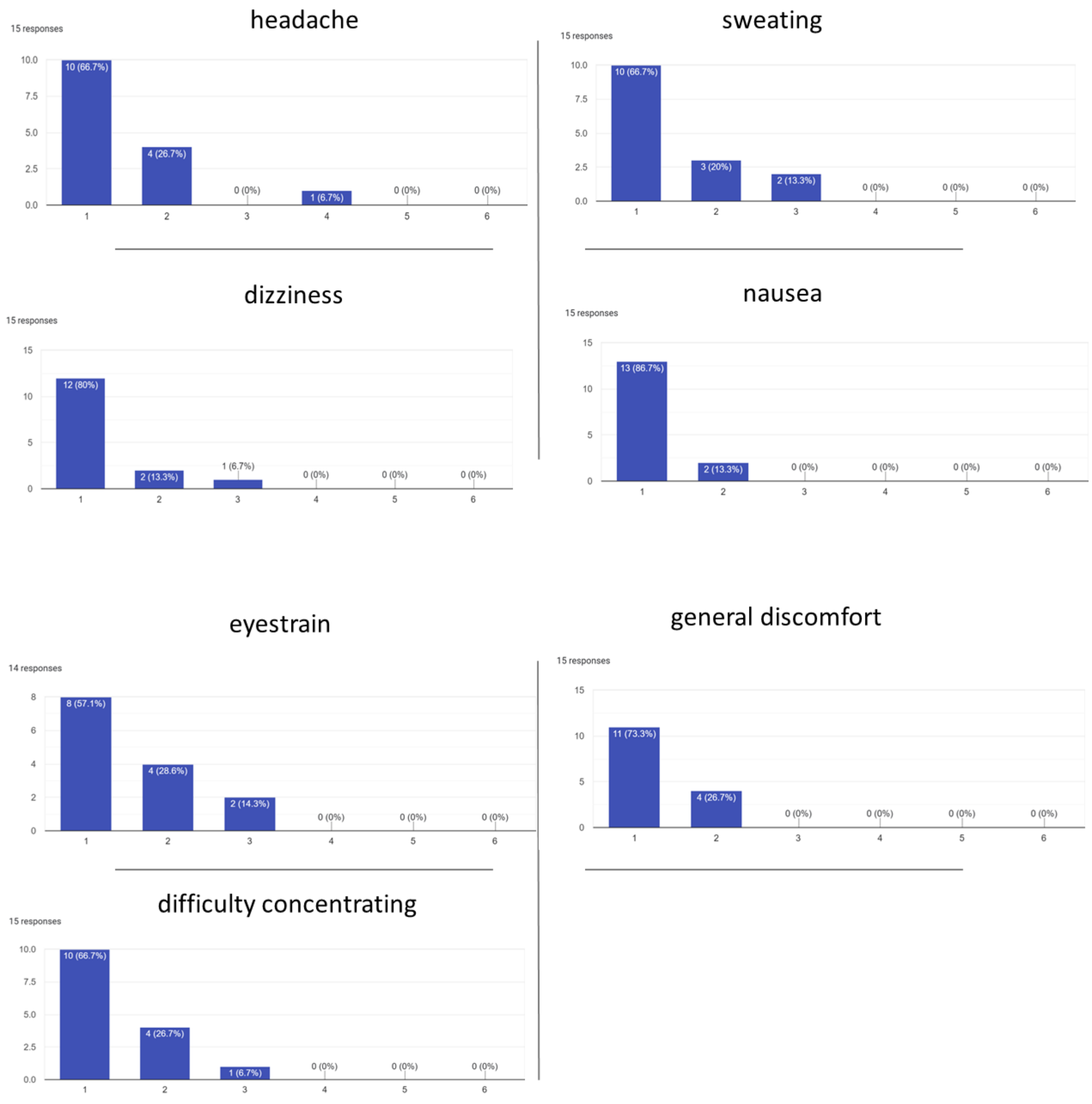


Figure 4.18: *Bar plots of the entity of the symptoms experienced during the test obtained from the second part of the ad hoc developed questionnaire.*

The open question reported at the end of the questionnaire ("In which application could the greatest added value be achieved by using the technology?") received the following answers:

- Mini-invasive surgery (2 responses)
- Teaching (1 response)
- Functional study analysis (1 response)
- Patient-specific surgical procedures planning and simulation (5 responses)
- Percutaneous vascular interventional surgery (1 response)
- Interventional cardiology (1 response)
- No answer (4 responses)

These answers clearly suggest that, despite the lack of experience and the short duration of the test, participants perceived an added value of MR technologies and identified applications that are indeed known for being the focus of current development and exploitation efforts. This is an important consideration because, for a new technology to be accepted, it is first required that end-users perceived an added value in it.

4.2. Technology assessment methodology

The 6 evaluation criteria extracted, as described in section 3.2, from literature review and expert and developer opinions, encompass hardware and software features of the technology as well as its organizational impact. In particular, the evaluation criteria have been articulated as follows:

- reduced weight and ergonomics
- field of view width
- good depth perception
- rendering quality
- workflow simplification
- simplicity and immediacy of use

The criterium concerning weight and ergonomics, i.e., the one referring to the hardware components, would play a relevant role primarily in intra-operatory applications. In this context, surgeons find themselves in a stressful condition where both mental and physical disturbances may have a large impact on the final outcome. Moreover, surgical interventions can last several hours, making it impossible for the operator to wear a very heavy and uncomfortable headset, which may hamper some necessary movements

or induce physical pain during or after the performance. This is still one of the major limitations to the application of the technology in intra-operative settings and the next challenge will be the miniaturization of the technological components.
The features concerning field of view width, depth perception, and rendering quality are referred to software capabilities. Great results have been obtained in this context by companies and startups developing XR platforms, even if steps forward can still be done. In the end, the capability of the technology to reduce the workflow and its learnability are factors relating to the organizational impact, but these will probably influence the acceptance among potential users at least as much as the technological aspects.

For what concerns the definition of the weight to be assigned to each of the considered features, the result of the pairwise comparisons performed both before and after the test are shown in the plots in Figure 4.19. Even if the opinion of the users was quite variable, it is possible to observe that in many cases the criterium referring to weight and ergonomics was considered the least important, being often even rated with 0% of relative relevance. This result should be interpreted carefully since, given the short duration of the test, it is possible that the users did not have enough time to perceive the physical encumbrance of the device as annoying or problematic. Also, according to many users, the most relevant dimensions were the ones relating to depth perception and rendering quality. This is a significant piece of information suggesting that a lot of importance is given to the quality of the image that is visualized, considered by the participants the most crucial characteristic the device should ensure in order to be efficiently exploited.

In addition, it is possible to observe that the opinion of the users changed after the performance of the test but not with a regular pattern, leading to the conclusion that the initial ideas and expectations they had about the technology did not necessarily correspond to the needs and priorities perceived when using it.

Again, in order to have more informative results about the priorities of the users with respect to the technology's features, it would be necessary to increase both the sample size and the interaction time, also allowing participants to perform some tasks in a real case setting.

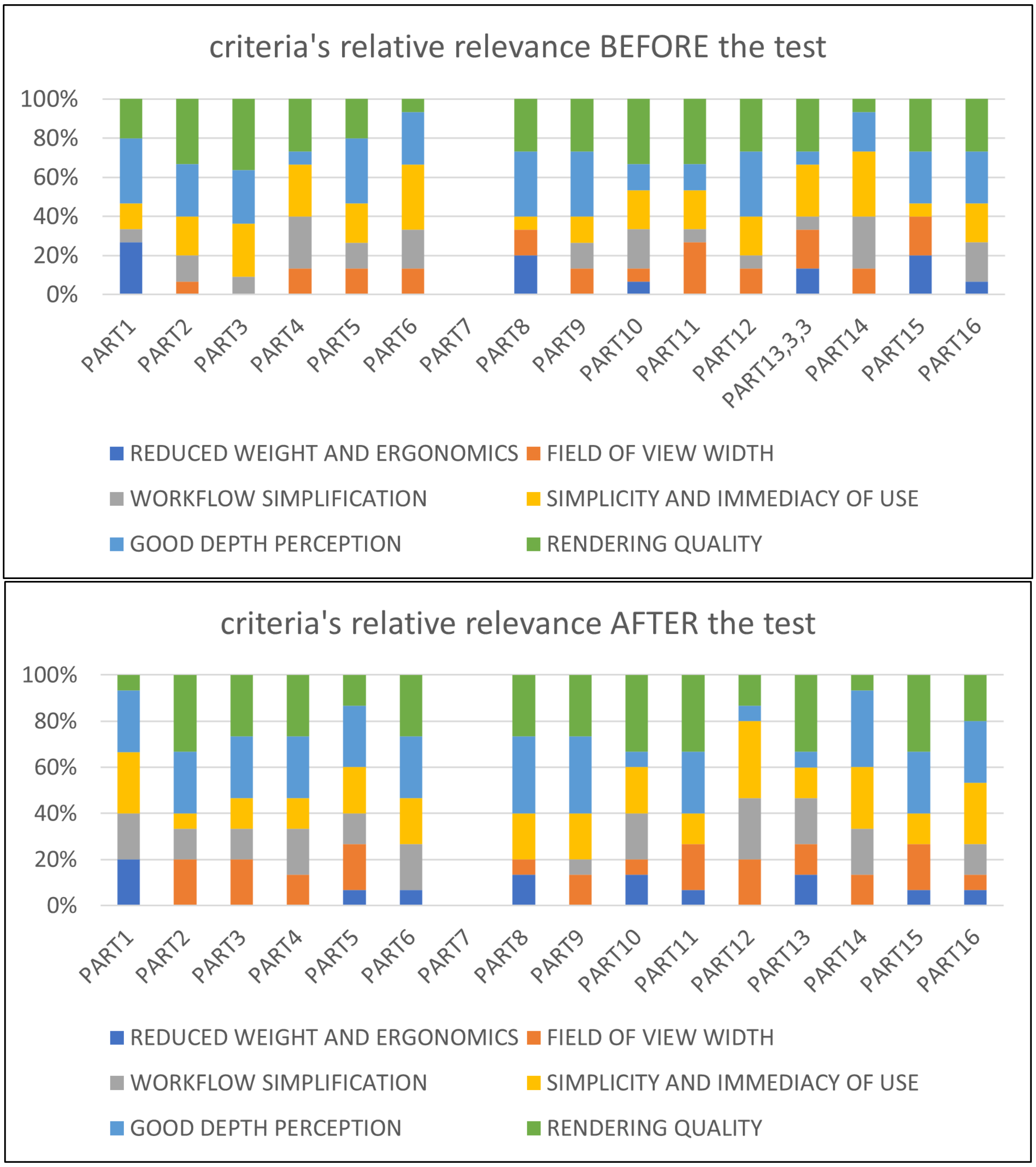


Figure 4.19: *Relative relevance of MR platform's evaluation criteria before (top) and after (bottom) the test with ARTICOR.*

5 | Conclusions and limitations

The numerous considerations deriving from the carrying out of this study, already presented in the previous section *Results and discussion*, are summarized in the following lines to facilitate the reader in retaining this information. In addition, the limitations of the current study are summed up.

From the observation of participants' performance during the test, no relevant issues concerning the interaction with ARTICOR platform were identified and everyone managed to complete the tasks even without previous experience in using MR technologies. The most relevant problem was related to the altered depth perception which, in some cases, caused the failure in grabbing or moving virtual objects.
On the other side, the interaction with RadiAnt was problematic for users with limited experience in using DICOM viewer software and some of them did not manage to fulfil the experimental tasks.
Newbies to DICOM viewer software took a shorter time (less than 1 minute per task) to perform the tasks with ARTICOR, requiring instead more than 1 minute to perform each same task with RadiAnt. Experts in using DICOM viewer software resulted in being faster in performing the tasks with RadiAnt, even though the average time required to fulfill each task using ARTICOR was below 1 minute as well.

The evaluation of the workload and stress perceived during the tasks was performed through the validated Surgery Task Load Index questionnaire. When considering the whole sample population, no statistical difference was found between the S-TLXs resulting from the use of the two technologies. Considering instead the two groups in which the sample population has been divided in the context of this study (group 0 = newbies in DICOM viewers; group 1 = experts in DICOM viewers), the differential workload scores (`STLX_ARTICOR` − `STLX_RadiAnt`) were statistically different: group 0 rated ARTICOR with a lower workload score with respect to the comparator, while group 1 perceived RadiAnt as less stressful than ARTICOR. Even the absolute value of this variable statistically differed between the two user groups: the workload reduction perceived by group 0 when using ARTICOR was more significant than the reduction perceived by group 1 when

using RadiAnt. This group rated both technologies with a similar workload score even if RadiAnt or a similar software had been used for many years in their working routine while ARTICOR was used for the first time.
In addition, the analysis of this survey highlighted that participants perceived a more significant physical effort when using the MR platform with respect to the viewer software.

The user experience, analyzed through the validated User Experience Questionnaire, was rated on average as positive for both technologies. ARTICOR however reported higher user experience scores with respect to RadiAnt and, when compared to a benchmark dataset, was ranked in the range of 10% best results across almost all the evaluation dimensions.

The evaluation of the usability of the two systems, addressed through the validated System Usability Scale, reported again optimistic results, even if no statistically significant differences were found in the comparison of the SUS scores relating to ARTICOR and RadiAnt on the entire sample of users. Comparing instead the differential usability score in the two groups of participants, the result showed significant differences in the variable `SUS_ARTICOR - SUS_RadiAnt`, which was positive for group 0 and negative for group 1. Even the absolute value of this variable was found to be significantly different between the two groups: group 0 found ARTICOR far more usable than RadiAnt while group 1, despite the long experience in using DICOM viewer software and inexperience in MR, perceived a similar usability level of the two technologies.

Even if both the user groups had a generally positive attitude toward the proposed MR technology, the general conclusion that can be derived from these results is that the level of confidence in using traditional DICOM viewer software could influence the acceptance of innovative MR technologies. Operators already experienced in the use of DICOM viewer software resulted in being less prone to perceive the advantages of the MR platform, and hence to learn how to use a new tool designed to replace their usual and already efficient one. Conversely, newbies in both technologies perceived the MR platform as significantly less stressful and more intuitive, informative and usable with respect to the gold standard, thus they might be more prone to accept it.

The ad hoc developed questionnaire referring to ARTICOR reported participants' general agreement with the positive statements (i.e., the ones expressing strengths of the system) and general disagreement with the negative ones (those reporting weaknesses of the system).
None of the considered symptoms was reported as annoying by the respondents.
Furthermore, even without previous knowledge and with a very short period of interaction

with the MR technology, the participants correctly understood the application contexts in which it might be more efficiently exploited.

In the end, the attempt to define a methodology to be applied in the assessment of MR technologies resulted in the selection of six evaluation criteria: reduced weight and ergonomics, field of view width, good depth perception, rendering quality, workflow simplification, and simplicity and immediacy of use. Among them, low weight and ergonomics were considered the least important features while the dimensions relating to depth perception and rendering quality as the most relevant. Given the short duration of the test, it is possible that the users did not have enough time to perceive the physical encumbrance of the device as annoying and problematic while the fact that a lot of importance is given to the quality of the image that is visualized suggests that, for most of the participants, it is the most important characteristic the device should ensure in order to be efficiently exploited. The fact that the opinion of the users changed after the performance of the test lead to the conclusion that the initial ideas and expectations they had about the MR technology did not necessarily correspond to the needs and priorities perceived when using it.

The study is affected by some limitations. The first one is the short duration of the testing phase: although usability is an immediate feature to be perceived, the evaluation of some relevant aspects of the proposed technology would require a longer interaction time. At first, the weight of the headset does not constitute a problem if it is worn only for a few minutes but, when using the technology for intra-procedural support, the wearing time can extend up to a few hours. In this case, if the design of the headset does not guarantee the correct distribution of weights and pressures over the surface of the head, pain and physical discomfort may be induced and, in turn, might negatively affect the performance of the operator. As a consequence, future studies should tackle this problem by making participants interact with the technology for a significantly longer time.

The second limitation is the small sample size: my study population consisted of a total of 16 participants, among whom someone did not correctly fill in some of the evaluation questionnaires, resulting in the presence of missing data that further reduced the sample size. The division of the participants into two groups based on their level of expertise in using standard DICOM viewer software (1 = experts, 0 = newbies) led to derive interesting conclusions but the significance of the statistical tests is questionable. The subdivision, in fact, resulted in 2 groups formed by 7 and 9 participants respectively. The fact of having heterogeneous participants is in itself positive since it allows for comparisons and analyses of how different factors affect users' performances and perceptions with respect to the technologies. However, the numerosity of each group should be increased in order to derive conclusions of general validity.

Other limitations of the study are related to the test itself, which was designed to include very simple and intuitive tasks. In future studies, the level of difficulty should be increased, and more realistic tasks should be defined. For example, remaining in the context of pre-operatory planning, it would be interesting to ask participants to perform diagnoses or to classify anatomical abnormalities by navigating the virtual model. However, participants should share a specific background for such detailed tasks to be suitable for all of them. Referring instead to the use of the technology in the intra-operatory context, in vitro simulations should be planned. Of course, this type of test would be more expensive to perform because of the need for simulating both the OR and the patient itself. Also, it would be far more complex to plan because it would require a protocol replicating the one followed in real OR contexts and of selected participants able to perform the surgical procedure that is simulated.

Bibliography

[1] NASA task load index. *Human Performance Research Group.*

[2] Directive 93/42 on medical devices. *Council of the European Communities, Official Journal of the European Communities*, June 1993.

[3] *IEC 62366-1:2015 Medical devices — Part 1: Application of usability engineering to medical devices*, 1° edition, February 2015.

[4] *IEC/TR 62366-2:2016 Medical devices — Part 2: Guidance on the application of usability engineering to medical devices*, 1° edition, April 2016.

[5] Regulation 2017/745 on medical devices. *European Parliament and Council of the European Union, Official Journal of the European Union*, April 2017.

[6] *ISO 9241-11:2018 Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts*, 2° edition, March 2018.

[7] *ISO 14971:2019 Medical devices — Application of risk management to medical devices*, 3° edition, December 2019.

[8] A. Bangor, P. Kortum, and J. Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. 2009.

[9] B. Battulga, T. Konishi, Y. Tamura, and H. Moriguchi. The effectiveness of an interactive 3-dimensional computer graphics model for medical education. July 2012. doi: 10.2196/ijmr.2172.

[10] P. Bimberg, T. Weissker, and A. Kulik. On the usage of the simulator sickness questionnaire for virtual reality research.

[11] J. Brooke. SUS - a quick and dirty usability scale.

[12] E. Bruckheimer, C. Rotschild, T. Dagan, G. Amir, A. Kaufman, S. Gelman, and E. Birk. Computer-generated real-time digital holography: first time use in clinical medical imaging. *European Heart Journal - Cardiovascular Imaging*, 2016. doi: 10.1093/ehjci/jew087.

[13] Brun, Bugge, Suther, Birkeland, Kumar, Pelanis, and Elle. Mixed reality holograms for heart surgery planning: first user experience in congenital heart disease. *European Heart Journal - Cardiovascular Imaging*, 2019. doi: 10.1093/ehjci/jey184.

[14] F. P. Chan, S. Aguirre, H. Bauser-Heaton, F. Hanley, and S. B. Perry. Head tracked stereoscopic pre-surgical evaluation of major aortopulmonary collateral arteries in the newborns. *Radiological Society of North America*, 2013.

[15] L. de Assis Pereira Cacau, G. U. Oliveira, L. G. Maynard, A. A. de Araújo Filho, W. M. da Silva Jr, M. L. C. Neto, A. R. Antoniolli, and V. J. Santana-Filho. The use of the virtual reality as intervention tool in the postoperative of cardiac surgery. *Brazilian Journal of Cardiovascular Surgery*, June 2013. doi: 10.5935/1678-9741.20130039.

[16] D. Drascic and P. Milgram. *Perceptual issues in augmented reality.* International Society for Optics and Photonics, 1996.

[17] J. Ender, J. Končar-Zeh, C. Mukherjee, S. Jacobs, M. A. Borger, C. Viola, M. Gessat, J. Fassl, F. W. Mohr, and V. Falk. Value of augmented reality-enhanced transesophageal echocardiography (TEE) for determining optimal annuloplasty ring size during mitral valve repair. *Annals of Thoracic Surgery*, November 2008. ISSN 00034975. doi: 10.1016/j.athoracsur.2008.07.073.

[18] L. D. Gennaro. Sviluppo e valutazione di applicazioni di realtà mista in ambito medico, 2019.

[19] H. H. Glas, J. Kraeima, P. M. van Ooijen, F. K. Spijkervet, L. Yu, and M. J. Witjes. Augmented reality visualization for image-guided surgery: A validation study using a three-dimensional printed phantom. *Journal of Oral and Maxillofacial Surgery*, September 2021. ISSN 15315053. doi: 10.1016/j.joms.2021.04.001.

[20] M. P. Haw, G. Baliulis, and N. D. Hillman. 3D printing and interactive 3D visualization for surgical planning in complex congenital heart disease. October 2019.

[21] J. D. Kasprzak, J. Pawlowski, J. Z. Peruga, J. Kaminski, and P. Lipiec. First-in-man experience with real-time holographic mixed reality display of three-dimensional echocardiography during structural intervention: balloon mitral commissurotomy. February 2020. doi: 10.1093/eurheartj/ehz127.

[22] J. C. Lu, G. J. Ensing, R. G. Ohye, J. C. Romano, S. T. O. Peter Sassalos, T. Thorsson, S. Yu, R. Lowery, and M.-S. Si. Stereoscopic three-dimensional visualization for

congenital heart surgery planning: Surgeons' perspectives. *Journal of the American Society of Echocardiography*, June 2020. doi: 10.1016/j.echo.2020.02.003.

[23] Microsoft. Hololens 2, 2022. URL `https://www.microsoft.com/en-us/hololens/hardware`.

[24] S. Moosburner, C. Remde, P. Tang, M. Queisner, N. Haep, J. Pratschke, and I. M. Sauer. Real world usability analysis of two augmented reality headsets in visceral surgery. *Artificial Organs*, 43, July 2019. ISSN 15251594. doi: 10.1111/aor.13396.

[25] J. L. Mosso-Vázquez, K. Gao, B. K. Wiederhold, and M. D. Wiederhold. Virtual reality for pain management in cardiac surgery. *Cyberpsychology, Behaviour and Social Networking*, June 2014. doi: 10.1089/cyber.2014.0198.

[26] C. S. Ong, A. Krishnan, C. Y. Huang, P. Spevak, L. Vricella, N. Hibino, J. R. Garcia, and L. Gaur. Role of virtual reality in congenital heart disease. *Congenital Heart Disease*, 2018. ISSN 17470803. doi: 10.1111/chd.12587.

[27] L. Qian, A. Barthel, A. Johnson, G. Osgood, P. Kazanzides, N. Navab, and B. Fuerst. Comparison of optical see-through head-mounted displays for surgical interventions with object-anchored 2D-display. *International Journal of Computer Assisted Radiology and Surgery*, June 2017. ISSN 18616429. doi: 10.1007/s11548-017-1564-y.

[28] A. A. Rad, R. Vardanyan, A. Lopuszko, C. Alt, I. Stoffels, B. Schmack, A. Ruhparwar, K. Zhigalov, A. Zubarevich, and A. Weymann. Virtual and augmented reality in cardiac surgery. *Brazilian Journal of Cardiovascular Surgery*, 2022. ISSN 16789741. doi: 10.21470/1678-9741-2020-0511.

[29] J. Sauro. 5 second usability tests, November 2010. URL `https://measuringu.com/five-second-tests/`.

[30] J. Sauro. Predicting Net Promoter Scores from System Usability Scale scores, January 2012. URL `https://measuringu.com/nps-sus/`.

[31] J. Sauro. 5 ways to interpret a SUS score, September 2018. URL `https://measuringu.com/interpret-sus-score/`.

[32] M. Schrepp. User Experience Questionnaire handbook, 2019. URL `www.ueq-online.org`.

[33] J. N. A. Silva, M. Southworth, A. Dalal, G. F. V. Hare, and J. R. Silva. Improving visualization and interaction during transcatheter ablation using a mixed reality system: First-in-human experience. URL `http://silvalab.bme.wustl.edu`.

[34] J. N. A. Silva, M. B. Privitera, M. K. Southworth, and J. R. Silva. Development and human factors considerations for extended reality applications in medicine: The enhanced electrophysiology visualization and interaction system (ĒLVIS). Springer, 2020. ISBN 9783030496975. doi: 10.1007/978-3-030-49698-2_23.

[35] M. K. Southworth, J. N. Silva, W. M. Blume, G. F. V. Hare, A. S. Dalal, and J. R. Silva. Performance evaluation of mixed reality display for guidance during transcatheter cardiac mapping and ablation. *IEEE Journal of Translational Engineering in Health and Medicine*, 2020. ISSN 21682372. doi: 10.1109/JTEHM.2020.3007031.

[36] F. Sternini, G. Isu, G. Iannizzi, D. Manfrin, N. Stuppia, F. Rusinà, and A. Ravizza. Usability assessment of an intraoperative planning software. SciTePress, 2021. ISBN 9789897584909. doi: 10.5220/0010252904830492.

[37] T. S. Tullis and J. N. Stetson. A comparison of questionnaires for assessing website usability. 2006.

[38] M. R. Wilson, J. M. Poolton, N. Malhotra, K. Ngo, E. Bright, and R. S. W. Masters. Development and validation of a surgical workload measure: the surgery task load index (SURG-TLX). *World J Surg.*, 2011. doi: 10.1007/s00268-011-1141-4.

List of Figures

List of Tables