Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Scienze dell'Ingegneria
Corso di Dottorato in Ingegneria Industriale
----------------------------------------------------------------------------------------

# Effectiveness analysis of traditional and mixed reality simulations in medical training: a methodological approach for the assessment of stress, cognitive load and performance

Ph.D. Dissertation of:

**Agnese Brunzini**

Supervisor:

**Prof. Michele Germani**

Assistant Supervisor:

**Eng. Alessandra Papetti**

Ph.D. Course coordinator:

**Prof. Giovanni Di Nicola**

XIX edition - new series

Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Scienze dell'Ingegneria
Corso di Dottorato in Ingegneria Industriale

---

# Analisi dell'efficacia di simulazioni tradizionali e in Realtà Mista nella formazione in medicina: un approccio metodologico per la valutazione di stress, carico cognitivo e performance

Tesi di Dottorato di:

**Agnese Brunzini**

Supervisor:

**Prof. Michele Germani**

Assistant Supervisor:

**Ing. Alessandra Papetti**

Coordinatore del corso di Dottorato:

**Prof. Giovanni Di Nicola**

XIX edizione – nuova serie

Università Politecnica delle Marche
*Dipartimento di Ingegneria Industriale e Scienze Matematiche*
Via Brecce Bianche — 60131 - Ancona, Italy

# Abstract

Simulation in medical education is considered a training method capable of improving clinical competence and practitioners' behaviour, and, consequently quality of care and patient's outcome. Moreover, the use of new technologies, such as augmented reality, offers to the learners the opportunity to engage themselves in an immersive environment. The opportunity to experiment with this innovative instructional method is effective not only in reducing the risk of errors and wrong approaches but also in experiencing anxiety and stress as in the real practice. The challenge is to find the right stress balance: learners have to feel as if they were practicing in the real stressful clinical case, and, at the same time, post-traumatic stress disorders, verifiable especially in the emergency field, must be controlled and avoided. Moreover, it is fundamental also to obtain high performance and learning, thus avoiding cognitive overloads. However, extensive researches about the impact of medical simulations on students' stress, frustration, cognitive load, and learning are still lacking.

For this reason, the main objective of this study is to assess simulation training effectiveness by analysing performance, anxiety, stress, and cognitive load during traditional (with manikin) and advanced (with augmented reality) clinical simulations.

A structured and comprehensive methodological approach to assess performance, emotional and cognitive conditions of students has been developed. It includes the acquisition and analysis of psychological parameters (subjective assessment), biometric signals (objective assessment), and task performance. This investigation allows to point out simulations'

weaknesses and offers the opportunity to define useful optimisation guidelines.

The methodology has been applied on three case studies: the first one refers to high-fidelity simulations, for the patient management in the emergency room, the second one refers to low-fidelity simulation for rachicentesis. For the third case study, a prototype of mixed reality simulator for the rachicentesis practice has been designed and developed aiming at improving the sense of realism and immersion of the low-fidelity simulation.

While 148 students have been enrolled in the first two case studies, only 36 students have taken part in the pilot study about mixed reality simulation.

Descriptive analysis about performance, cognitive and emotional states have been done in all the case studies. For the high-fidelity and low-fidelity simulations, the statistical regression analysis has pointed out which variables affect students' performance, stress, and cognitive load. For the pilot study about mixed reality, the user experience analysis highlighted the technical limitations of the new technology.

# Contents

# List of Abbreviations

| | |
|---|---|
| ANS | Autonomous Nervous System |
| AR | Augmented Reality |
| ARA | Augmented Reality Application |
| AUCi | Area Under the Curve with respect to Increase |
| AV | Augmented Virtuality |
| BPM | Beats Per Minute |
| BR | Breathing Rate |
| CE | Cognitive Ergonomics |
| CFS | Cerebrospinal Fluid |
| CL | Cognitive Load |
| CLT | Cognitive Load Theory |
| CPR | CardioPulmonary Resuscitation |
| ECG | Electrocardiography |
| ECL | Extraneous Cognitive Load |
| EDA | Electrodermal Activity |
| EEG | Electroencephalography |
| EMG | Electromyography |
| EOG | Electrooculography |
| FDA | Food and Drug Administration |
| GCL | German Cognitive Load |
| GSR | Galvanic Skin Response |
| HF | High Fidelity |
| HFMS | High Fidelity Manikin Simulator |

| | |
|---|---|
| HFS | High Fidelity Simulation |
| HMD | Head-mounted Display |
| HR | Heart Rate |
| HRV | Heart Rate Variability |
| HW | Hardware |
| IBI | Inter-Beat Intervals |
| ICL | Intrinsic Cognitive Load |
| LF | Low Fidelity |
| LFS | Low Fidelity Simulation |
| MR | Mixed Reality |
| MRS | Mixed Reality Simulation |
| MWL | Mental WorkLoad |
| NAS | Numeric Analog Scale |
| NASA-TLX | NASA-Task Load Index |
| pNN50 | Percentage of successive RR intervals that differ by more than 50ms |
| RMSSD | Root Mean Square of Successive RR interval differences |
| RR | Consecutive peaks of the R-wave of the ECG |
| SCL | Skin Conductance Level |
| SDK | Software Development Kit |
| SDRR | Standard Deviation of RR intervals |
| STAI | State Trait Anxiety Inventory |
| SW | Software |
| UX | User Experience |
| VR | Virtual Reality |

# List of Figures

# List of Tables

# 1. Introduction

European data from the World Health Organization consistently evidence that medical errors and healthcare-related adverse events occur in 8% to 12% of hospitalizations. In this context, 23% of European Union citizens claim to have been affected by medical errors, 18% of them have experienced a serious medical error in the hospital setting, and 11% declare to have been prescribed the wrong medication. Evidence on medical errors shows that 50% to 70.2% of such harm can be prevented through comprehensive systematic approaches to patient safety. Statistics show that strategies to reduce the rate of adverse events in the European Union alone would lead to the prevention of more than 750.000 harm-inflicting medical errors per year, leading in turn to over 3.2 million fewer days of hospitalization, 260.000 fewer incidents of permanent disability, and 95.000 fewer deaths per year (WHO).

Despite the technical, legal, and organizational improvements, cultural change, and adequate medical education are the key drivers of quality improvement. In this context, the safety and innovation challenges have raised the need to step forward from a traditional 'apprentice' learning model to a simulation-based learning model; the old – 'see one, do one, teach one' has become 'see one, practice many, do one'. The increased awareness about clinical risk management and ethical issues has imposed the shift to a more skilled/hands-on training of healthcare professionals.

Technological innovations, such as advanced physical medical simulators, and extended reality applications (i.e. virtual, augmented, and mixed reality simulations), have led to consistent improvement in learning outcomes. Students may learn better and retain information more effectively by engaging themselves in an immersive and realistic experience. Today's highly sophisticated simulators make it possible for medical students and practitioners to examine rare conditions, witness complex procedures, and keep themselves up to date by acquiring new knowledge and attitudes. Thanks to the simulators, it is possible to reproduce particular conditions not always verifiable in clinical practice during the training sessions and to linger on critical issues in emergencies and not. In this way, junior and senior healthcare professionals have the chance to repeat, in a safe environment, specific situations particularly stressful in terms of both therapeutic and psychological approaches. Indeed, the stress' role in the simulation context is crucially important. On one side, the feeling of stress should be similar to that one felt in real practice, so that students can experience how to work in stressful conditions; on the other side, it should be maintained within certain limits, avoiding post-traumatic stress disorders. During simulations, the participant is also simultaneously exposed to the realism of the event, and to the demand to execute the correct intervention, that arise the ever-present stress of the time limit (Sahu, et al., 2010).

Therefore, the opportunity to experiment with this innovative educational method should be effective in reducing not only the risk of errors and wrong approaches for the patients but also anxiety and stress for the medical staff. In fact, in the case of wrong approaches, the clinician can feel discomfort and acute stress that can compromise her/his performance.

For this reason, simulations are not limited to the practice itself, but they usually consist of three different parts: briefing, simulated scenario, and debriefing. The debriefing is considered one of the most important elements in providing effective learning. In fact, while during the briefing, the teacher reviews essential skills, gives an overview of what to do, and explains the learning environment and tools, during the debriefing, the teacher asks the learners how they felt during and after the simulated scenario and discusses the students' actions and case management. This allows focusing on students' performance, not only discussing errors and clinical decisions but also reducing students' stress and frustration due to possible wrong approaches and actions taken during the simulation.

For all these reasons, simulation in clinical education has been viewed as a mean to improve the clinical competence of health practitioners. By providing opportunities for students to refine skills away from real patients and in controlled environments, the improvement of the quality of care and patient safety are guaranteed. To allow the exposure to typical as well as rare patient presentations, simulations can be engineered so that all students can encounter these situations during their academic path, providing the opportunity to experience realistic training, in terms of both clinical practice and stress management. A better, effective, and realistic medical training is fundamental to reach improved clinical performance and reduced risk for the patient.

Nevertheless, extensive researches about the impact of medical simulations on students' stress, perceptions, workload, and cognitive load are still lacking. For this reason, the main objective of this study is to analyse the effectiveness of medical simulation training, in terms of students' stress, cognitive load, and performance.

A structured and comprehensive methodology, relevant for every kind of medical simulation, has been defined to analyse the cognitive ergonomics (CE) of participants. A mixed-reality simulator prototype has been designed and developed to improve the realism of the simulation and to do preliminary analysis about the impact of augmented reality on students' performance, emotional, and cognitive states.

This analysis is extremely useful as a starting point for the optimisation and re-design of simulators and simulations. Indeed, the final aim of this work is to understand how to improve learners' performance, avoiding cognitive overloads, and controlling anxiety and post-traumatic stress disorders, especially verifiable in the context of emergencies.

The state of the art about simulations in healthcare, applications, and limitations of augmented reality in this field, and human factors and ergonomics related to the training of hospital personnel are described in Chapter 2. The proposed methodology to assess learners' stress, frustration, anxiety, and mental workload is described in Chapter 3 with the method to design effective medical simulations and to develop the mixed reality rachicentesis prototype. The proposed methodology has been tested on two different observational studies and a pilot study. The high-fidelity simulation has been selected as the first case study because, being involved in a stressful scenario to train technical and soft skills, the learner can feel mental discomfort, anxiety, and stress that must be evaluated. The second observational case study is about the low-fidelity simulation of rachicentesis practice, for the training of technical skills, while the third pilot study is related to the mixed reality simulation of rachicentesis. These case studies are analysed in Chapter 4. Chapter 5 reports preliminary guidelines for the optimisation and re-design of the assessed simulations, with some future perspectives.

# 2. Research and Background

## 2.1. Simulation in Healthcare

Simulation is defined as an imitation of a real situation or system and has been extensively used for educational purposes in several fields for decades. It is considered an excellent tool to reduce errors in high-risk industries such as aviation, defence, the nuclear energy field, and even healthcare. The healthcare industry has begun to use simulation to reduce the chance of bad outcomes, especially where events do not occur with regularity and need practice to be learnt (Ruddy, et al., 2008). Another important and compelling advantage of simulation is the ability to simulate complications without putting patients at risk. This is a core ability that drives the development of medical simulation (Dawson, et al., 1998).

Moreover, simulation-based learning is an innovative teaching method that provides healthcare students and professionals with more opportunities to acquire knowledge, skills, and attitudes for developing clinical abilities (AL Sabei, et al., 2016). By presenting the trainee with a variety of procedures, exposure to a specialty will be more consistent and uniform and learning can occur more rapidly, without the necessity of waiting for a patient with a specific disease (Dawson, et al., 1998). From the simulations, healthcare staff can learn a lot of technical skills and practical procedures and also several non-technical skills such as clinical decision-making, situation awareness, team working, communication skills. For this reason, simulation training satisfies almost all the medical and nursing fields, being the optimal option for preparing and assessing human responses to real-life problems. In fact, the value of simulation for clinical teaching, learning, and assessment is the object of growing acknowledgement (Vincent-Lambert, et al., 2017).

### *2.1.1. Simulation and Fidelity: Definitions and Classifications*

#### 2.1.1.1. Simulation Classification

Nowadays, simulation can be of different types and a standard classification does not exist. It can be classified into human or non-human simulation: in the first case it is carried out as a role-play among students, in the other case it is accomplished using a manikin or computer. It may be also classified according to the type (as compiler-driven and event-driven) or the fidelity (as low, medium, and high-fidelity) (Elshama, 2020).

It can also be divided into four areas by the educational tool (Sahu, et al., 2010):

- Standardized patients: the learner trains his/her skills on actors who give specific responses to certain medical conditions.
- Screen-based computer: the learner practices patient care and receives feedback from interactive software.
- Skill trainer (or partial task or procedural simulator): the learner tries specific procedures (such as making an injection or placing a chest tube) on skill trainers that usually represent the part of human anatomy relevant to the range of skills that has to be learned (Vincent-Lambert, et al., 2017).

- High-fidelity manikin simulators (HFMS): especially useful to train psychomotor and cognitive domains of learning, HFMS provide some of the most realistic and high-yield environments for trainees, being able to reproduce almost any disease entity. They are full-sized, computer-controlled, and can recreate heart/lung sounds, pupil movements, changes in heart rate, blood pressure, ECG, and breathing rate. Some of them may also physiologically respond to the medication selected by the trainees.

This work focuses on simulations based on high-fidelity manikin simulators and skill trainers (low fidelity simulations). An example of both is shown in *Figure **1***.

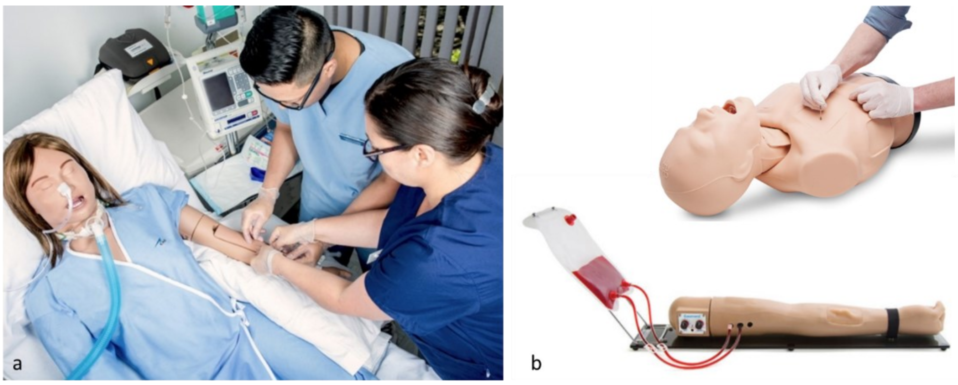


*Figure 1: Examples of HFMS (a) (CAEHealthcare) and skill trainers (b) (Gaumard) (Nasco)*

While HFMS automatically generates physiological outputs using mathematical algorithms and electronics, skill trainers are less expensive and easier to use but they provide limited feedbacks and seem less real.

The fully immersive simulation with high-fidelity simulators has been shown to improve the quality of resuscitation efforts by medical trainees (Mercer, 2017) and is considered an effective teaching and learning method not only for emergency management but also for the clinical training of medical students and residents in general (Issenberg, et al., 2005) (Fraser, et al., 2012).

Norman et al. contrasted high fidelity simulations (HFS) and low fidelity simulations (LFS) gathering 24 studies (about auscultation skills, surgical techniques, and complex management skills as cardiac resuscitation) and including some performance measurements. They found that both HFS and LFS learning resulted in consistent improvements in performance in comparisons with no-intervention control groups, but the advantage of HFS over LFS ranges only from 1% to 2% in almost all the studies. Therefore, the relationship between simulation fidelity and learning could be not unidimensional and linear (Norman, et al., 2012).

Also Hanshaw et al., in their review, verified that simulation is superior to no intervention and non-simulation instruction (Hanshaw, et al., 2020).

#### 2.1.1.2. The Concept of Fidelity

As for the simulation classification, there is not a standard definition of *fidelity*. However, researchers agree in considering it as a multidimensional construct, composed of various dimensions.

Fidelity is the perception of how real or lifelike a simulator/simulation is for the user. Norman et al. identified two levels of fidelity: ‘engineering fidelity' and 'psychological fidelity' referring to the realistic look in the first case, and to the demand of specific behaviours to complete the task in the second case (Norman, et al., 2012). Vincent-Lambert et al. defined five dimensions of fidelity: "physical (environment, equipment, tools), psychological (emotions, beliefs, self-awareness of participants), social (motivations, goals), culture of the group and degree of openness or trust" (Vincent-Lambert, et al., 2017). Curtis et al., based on human factors literature, drawn three dimensions of fidelity (Curtis, et al., 2012):

- Physical Fidelity: it refers to the environment, equipment, and the degree to which the sensory characteristics (visual, motion, auditory) are recreated;
- Functional Fidelity: it involves actions, responses, and instrument accuracy;
- Psychological Fidelity: it refers to temporal, perceptual, and experiential dimensions to accomplish the tasks in the simulated scenario.

From Norman et al. review, psychological fidelity is very critical and important for learning and transfer (more than the other fidelity dimensions) (Norman, et al., 2012). In fact, accurately reproducing stressful conditions is one of the most fundamental and challenging aspects of simulation design. It is very difficult to replicate the same pressure that a practitioner can feel in front of a patient who is fighting against death. However, as Curtis et al. said "… higher degrees of psychological fidelity can help to achieve a level of stress closer to what would actually be experienced in the operational environment. Achieving psychological fidelity can be accomplished through a cognitive task analysis, which informs the thorough development of realistic scenarios, providing trainees with realistic time restrictions, and using immersive environments. …One psychological goal of the simulation is to help individuals suspend belief that they are operating in a replicated environment, in favour of feeling like they are engaged in the real-world task" (Curtis, et al., 2012). For this reason, during the design of simulations and simulators, great attention must be placed on the realism and feeling of immersion.

### *2.1.2. Design of Simulators and Simulations*

#### 2.1.2.1. Simulators Design

As asserted by Dawson et al., while using a medical simulator, the physician needs to interact with the anatomy in a clinically realistic manner and this is possible only with a close collaboration between physicians and engineers who design the simulator (Dawson, et al., 1998).

A medical simulator is purpose-designed as a system or device with interactive features that actively engage students in a real-world clinical process. The simulator mimics human patients and allows us to recreate specific clinical interactions such as clinician-patient or clinician-clinicians in case of group simulations.

Therefore, patient simulators can be designed with different purposes and can be composed of different levels of technology (physical, virtual, or a blend of both).

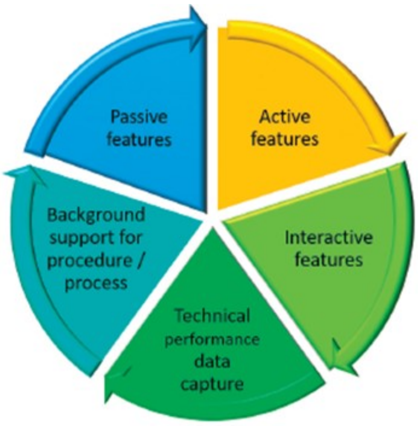


*Figure 2: Schematic overview of patient simulator features (Vincent-Lambert, et al., 2017)*

The features of physical/virtual medical simulators can be divided as shown in *Figure 2*:

- Passive features: they consist of the simulator appearance and structure and include size, shape, weight, colour, and texture of the simulator;
- Active features: they allow to mimic the real patient behaviour, such as movements, pulses, body sounds, and verbal sounds, and they can be changed during the simulation;
- Interactive features: they consist of the clinical progression and change in patient state as a response to students' medications, actions, and performance. Active and interactive features are necessary to promote psychological fidelity;
- Technical performance data capture: technology for capturing simulation activities, changes in the patient condition, and technical performance data that are not possible to collect and evaluate through observation;

- Background support for procedures/processes: support that may include a demonstration of the procedure, an explanation of possible risks and complications, a recap of the background anatomy, and so on. (Vincent-Lambert, et al., 2017)

The great level of realism of these kinds of simulators guarantees a full immersion in the simulated scenario, resulting in more efficient and fruitful learning. Nevertheless, the successful use of these sophisticated simulators needs additional guidance from medical staff and previous knowledge from the learners. Therefore, the inclusion of pedagogical and psychological expertise into the design and development of educational devices is essential (Holzinger, et al., 2009).

#### 2.1.2.2. Simulation Phases

To have a successful simulation, the scenario is not the only part to consider during its design: also briefing and debriefing must be carefully considered. Simulation-based training consists of three phases (AL Sabei, et al., 2016), (Vincent-Lambert, et al., 2017):

- Briefing or pre-briefing: before the simulation, the teacher explains the objectives of the simulation and how it will be conducted. He/she may also assign students' roles (but it is not mandatory);
- Scenario: it is the simulation itself, where trainees perform or observe the simulated clinical case based on a real-life situation. Several issues must be considered for the design of the scenario: training objectives, scenario

difficulty, setting, storyline, teacher and participants roles, expected outcomes, and different rationales depending on student performance;

- Debriefing: after the simulation, the teacher discusses with the trainee their performance and their feelings. It is an active retrospective assessment in which students appraise their technical and cognitive performance, highlighting the importance of human factors (Mercer, 2017).

Briefing and debriefing are generally conducted by those who designed and implemented the simulation. Debriefing has been considered the most important phase of simulation because it is a student-centred activity and a link between theory and practice that offers a meaningful time for reflection. As emerged from the literature analysis of AL Sabei et al., psychological debriefing is developed as a therapeutic practice for people experiencing traumatic and stressful events. In fact, discussing shared experiences in terms of practical performance, potential wrong approaches, and related feelings is a strategy to minimise the chance of post-traumatic stress disorders (AL Sabei, et al., 2016).

### *2.1.3. Limitations and Future Research*

Simulation-based training is characterised by a pedagogical framework that allows the students to experiment with the same workflow and workload that they would experience in real clinical cases (Elshama, 2020). The design of medical simulations passes through a careful analysis of learning objectives, technology to be used, instructor role, performance assessment, and so on. However, a series of issues need further research. Some research directions have been suggested by different authors. Bond et al. proposed to deepen the analysis of the simulation training impact on team function and the assessment of different debriefing techniques (Bond, et al., 2007). Hanshaw et al. recommended focusing on the design of fully integrated simulation curricula based on the learners' level (Hanshaw, et al., 2020). However, the evaluation of the simulation impact on actual healthcare outcomes is the main challenge still open. This core question can be split into several points:

- The effects of simulation learning on patient safety (Bond, et al., 2007), (Sawyer, et al., 2016), (Hanshaw, et al., 2020);
- The transfer of learned skills into real clinical practice (Bond, et al., 2007), (Hanshaw, et al., 2020);
- The evaluation of simulation learning retention over time (Hanshaw, et al., 2020);
- The development and use of assessment tools for the evaluation of procedural and behavioural competencies (Bond, et al., 2007), (Sawyer, et al., 2016), (Hanshaw, et al., 2020).

Among these suggested topics to be further analysed, this study mostly focuses on the last point. Indeed, one of the main goals of this work is to design and develop a structured and versatile methodology for the comprehensive simulation effectiveness assessment. Subjective user ratings can be misleading metrics of simulation effectiveness (Curtis, et al., 2012) and the need for new tools for the evaluation of technical and behavioural performances have emerged from literature (Sawyer, et al., 2016). Thus, the necessity of an innovative methodology for the assessment of performance, emotional, and cognitive conditions of learners, during and after the simulation training, has become always more compelling.

## 2.2. Augmented and Mixed Reality in Medical Education

### *2.2.1. Definitions, Tools, and Key Concepts*

Mixed Reality (MR) as defined by Milgram and Kishino (Milgram, et al., 1994), combines real and virtual worlds along the reality-virtuality continuum comprising of Augmented Reality (AR) and Augmented Virtuality (AV) technology, as shown in *Figure 3*.

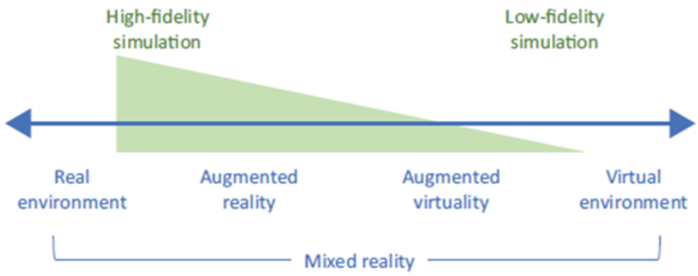


*Figure 3: Adapted from Milgram's reality-virtuality continuum (Stretton, et al., 2018)*

Augmented reality supplements the real world with virtual objects, that appear to coexist in the same space as the users' physical reality. This is the main difference with Virtual Reality (VR) which implies a complete immersion experience that excludes the real physical world. AR is interactive, having virtual content as part of the real world in real-time.

#### 2.2.1.1. Devices and tools

According to (Schmalstieg, et al., 2016), AR devices can be classified according to the distance between eye and display, as shown in *Figure 4*.

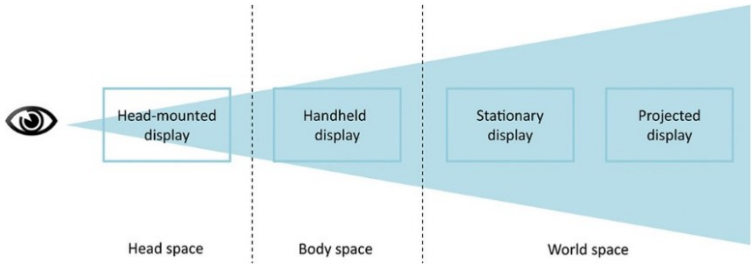


*Figure 4: Visual AR display classification (Schmalstieg, et al., 2016)*

Head-mounted displays (HMD) are the most advanced and common AR displays currently in the market. They are headsets in the form of glasses with optical see-through technology and they can be both monocular and binocular. Their function is to overlay virtual contents directly over the user's view. Some examples of HMD commercially available are the Epson Moverio®, the Microsoft Hololens®, the Meta® AR headset, etc. Certain HMDs allow users to control the virtual objects with voice commands, gestures, gaze, and even to "touch and grab" them with hands.

While handheld displays are small computing devices, such as smartphones and tablets, that the user can hold in their hands, stationary displays are non-moveable computers (such as mirror, desktop, kiosk displays) with built-in or tethered camera or webcam to track target objects.

Projected displays consist in the use of projectors to directly render surface details or visual effects on the surfaces of real objects.

#### 2.2.1.2. Key Concepts

A key issue in AR is the registration or correct alignment of the virtual world with the real one.

In order to display the virtual content in the real environment, the AR device needs to have real-time information about the position and orientation of the user, the surrounding environment, and any object populating the AR scene. This means that the AR device, through its sensors and cameras, has to continuously track such entities.

Tracking can be done based on two different working principles: marker-based or markerless. In the first case, the AR device has to scan a marker (e.g. a featured image with encoded information) and recognize the pattern to overlay the virtual object on the real environment. In the second case, the device can track and recognise the target object and overlay the virtual content into the scene, without the need for markers.

#### 2.2.1.3. Uses

Augmented Reality can be used for co-located (when users are simultaneously in the same place) or remote collaboration (when users are connected from different locations). This is a critical advantage in collaborative tasks that involve the physical world and experts from different locations (for example in case of equipment maintenance and repair, or healthcare emergencies).

Among other important AR use cases, mention should be made for its usage in design and planning (e.g. to prototype objects, devices, and buildings in 3D and understand their relationship with the real world), in training (e.g. to provide interactive additional information and feedback), in the retail industry (e.g. to enhance the customer experience).

According to the study of Garzon et al. about the AR use in the educational field, 52.5% of applications correspond to the field of Natural Sciences, Mathematics and Statistics, 15% to Social Sciences, journalism and information, 15% to Arts and Humanities and 15% to Engineering, manufacturing, and construction. Another important mentioned field of AR application is the healthcare sector (Garzón, et al., 2017). As an emerging technology, Augmented Reality has the potential to enhance medical training and education and, consequently, to improve the quality of care (Munzer, et al., 2019).

### *2.2.2. Applications in Healthcare and Medical Training*

Being the AR industry still in its infancy, publications in the healthcare field have become considerable starting from 2008 (even if the earliest study on AR in medical education was published in 2002) (Zhu, et al., 2014). Augmented Reality has been implemented in several healthcare areas, aiming at and involving all levels of learners (from novices to expert practitioners). During the last years, several review studies emerged in scientific literature trying to collect and categorize the AR applications in healthcare, especially in the area of medical training and education.

Most of the education and training research focused on high-risk, invasive skills such as endoscopy and surgery. Already in 2014, Zhu et al. found out that 64% of the reviewed papers were within surgery, primarily laparoscopic surgery (44%). However, a myriad of implementations and case studies exists.

Herron et al. reported that AR, among surgery, covers areas such as anatomy and forensic medicine and supports endotracheal intubation, joint injections, and local anaesthesia administration (Herron, 2016).

Barsom et al., in their systematic review, found twenty-seven relevant studies, describing seven augmented reality applications. They assigned the applications to three different categories: laparoscopic surgical training, echocardiography training, and MR training of neurosurgical procedures (Barsom, et al., 2016). Augmented reality neuro-navigation provides a real-time updated 3D virtual model of anatomical details, overlaid on the real surgical field. A systematic review of eighteen studies from 1996 to 2015 confirms that AR is a reliable and versatile tool when performing minimally invasive approaches in a wide range of neurosurgical diseases (Meola, et al., 2017). One reason to use augmented reality is to help the user having a visual of the patient's internal body state, interactively (Sherstyuk, et al., 2011). In the surgical field, this technology is used also in orthopaedic surgery, head and neck surgery, thoracic pedicle screw placement (Mehta, et al., 2018), (Zhu, et al., 2014). In neurosurgical oncology, AR has been used as an aid for the resection of cranial tumours. This technology is useful for both the presurgical planning and the intraoperative localization of lesions, in particular when combined with intraoperative imaging for real-time visualization (Mikhail, et al., 2019).

From the review of Munzer et al. concerning AR applied to emergency medicine, out of twenty-four articles, 50% focused on education and training and in particular on procedural training and clinical decision-making in a simulated environment (Munzer, et al., 2019).

Other medical areas of AR application include ventriculostomy, inguinal canal anatomy, diathermy, tissue engineering, alimentary canal physiology and anatomy, disease outbreak, clinical breast examination, life support training, echocardiography, ultrasound and anatomical education (Sherstyuk, et al., 2011), (Zhu, et al., 2014), (Mehta, et al., 2018), (Munzer, et al., 2019). The most recent and up-to-date review about AR and MR applications in healthcare education beyond surgery, considered twenty-six studies, from January 2013 to September 2018. Authors claimed that the most frequently studied subjects were within anatomy and anaesthesia (especially on central vein catheterization). Other subject areas are radiology, ophthalmology, cardiology, dermatology, family medicine, forensic medicine, gastroenterology, neurology, orthopaedics, paediatrics. Moreover, they highlighted that these studies involved established applications in 27% of all cases, while 73% regarded prototypes (Gerup, et al., 2020).

However, even if most research papers are about augmented and mixed reality prototypes, some AR/MR applications for healthcare are currently commercially available. The most famous and sophisticated examples were born from the collaboration between Microsoft® and CAE Healthcare®, Abiomed®, Pearson®, and 3D4Medical®. VimedixAR and LucinaAR by CAE Healthcare®, Abiomed Impella by Abiomed®, and HoloHuman by Pearson® and 3D4Medical® are all based on HoloLens employment (*Figure 5*). VimedixAR is an ultrasound simulator that displays human anatomy into its manikin body, synchronously with the ultrasound beam. LucinaAR is a childbirth simulator that shows real-time, interactive 3D holograms of anatomy and visual cues and feedback. Abiomed Impella allows training on ultrasound-guided placement of the world's smallest heart pump. Through the HoloLens AR framework, anatomy, physiological responses, and potential complications are shown and overlaid the ultrasound manikin. HoloHuman is a medical learning application for immersive 3D exploration of anatomy.


*Figure 5: (a) VimedixAR, (b) LucinaAR, (c) Abiomed Impella, (d) HoloHuman*

Therefore, these solutions are oriented to the study of human anatomy and the practice of specific procedures. Several AR and MR applications and prototypes involving procedural training can be found also in the scientific literature. As explained by Koziol et al., procedural learning "refers to the acquisition of motor skills and habits, and certain types of cognitive skills. It usually requires repetition of an activity, and associated learning is demonstrated through improved task performance" (Koziol, et al., 2012). Several examples of research papers about AR and MR application in procedural training are described in *Table 1*, with a specific focus on hardware and software used, assessment tools, and the number of participants in the experimentation.

*Table 1: Research papers about AR/MR applications in procedural training*

| PAPER | APPLICATION | HW AND SW | ASSESSMENT | NUMBER OF PARTICIPANTS |
|---|---|---|---|---|
| **(Magee, et al., 2007)** | AR for the simulation of ultrasound-guided needle insertion procedures for interventional radiology education and training | Manikin, mock ultrasound probe, needle, a pair of magnetic 3D position sensors | Experience evaluation (questionnaire 5-points Likert scale) | 60 users: 34 consultant interventional radiologists and 26 specialist registrars in radiology |
| **(Coles, et al., 2011)** | AR with haptics in femoral palpation and needle insertion | Two force feedback devices, custom-built hydraulic interface, a modified Phantom Omni end effector, LCD monitor | Face and content validation. Objective feedback in a 29-point questionnaire (7-point Likert scale) | 7 experts |
| **(Kotranza, et al., 2012)** | MR humans for clinical breast examination | Physical breast model, a webcam, | Study I: Efficacy study. Real-time, quantitative measures of | 3 groups: 12 novice medical students, 32 novices and 25 |

| | | | | |
|---|---|---|---|---|
| | | force sensors, passive manikin, a Head-Mounted Display | correct pressure and pattern of search are computed from the real-time sensor data. Study II: Performance study. Coverage and pressure recorded | experienced medical students, residents, and clinicians |
| **(Gutierrez-Puerto, et al., 2015)** | AR central venous access training simulator | Unity3D, Vuforia SDK, Oculus VR, Webcam, Surgical tools (real and virtual) | Technical test for the recognition of the objective and its location | 1 user |
| **(Wang, et al., 2017)** | AR for remote procedural training as a point of care ultrasound | Microsoft HoloLens, Leap Motion sensor, Unity3D | Comparative study. Performance with Global Rating Scale (GRS); Utility, Simplicity and Perceived Usefulness with a short Likert survey; Cognitive Load with time to perform the task, mental effort, and task difficulty rating | 25 users: 24 students and 1 mentor |
| **(Rochlen, et al., 2017)** | AR for central line insertion training | Epson Moverio, BT-200® Smart Glasses, Unity3D, | Performance checklist, Total time, Time to needle insertion, | 40 users (medical students and anaesthesiology residents) |

| | | | | |
|---|---|---|---|---|
| | | Vuforia SDK, Skill trainer | Survey about: level of training, previous experience, satisfaction with AR, perceptions, likes, and dislikes, potential barriers, suggestions | |
| **(Kobayashi, et al., 2017)** | AR/MR devices for acute care procedure training | Hololens, Maya software, 3DViewer Beta, Task trainer, a wirelessly networked laptop, LCD projector | Pilot application. Future work: Educational utility and effectiveness of AR/MR-based training on live-patient clinical outcomes. Metrics: operational quality markers (holoimage registration accuracy, stability, and usability), checklist-based procedural performance assessments | 40 learners |
| **(Lee, et al., 2018)** | AR to localize individual organ in surgical procedure | Unity3D, Vuforia SDK, Smartphone | NONE | |
| **(Bottino, et al., 2018)** | AR self-directed learning and evaluation system for | Microsoft HoloLens, Skill Trainer | Qualitative evaluation: assessment of the prototypal system. | 23 users (4 doctors and 19 residents) |

| | | | | |
|---|---|---|---|---|
| | effective basic life support defibrillation training | | Questionnaire 5-point Likert scale: appreciation, ease of use, cognitive load | |
| **(Tai, et al., 2019)** | AR-driven medical simulation platform for percutaneous renal access | Personal computer, 2 Phantom Omni with one stylus | Validation of: Face and content, skills improvement, construct, and criterion. Objective metrics and Global Rating Scale questionnaire | 54 professors (36 medical students and 18 urologists) |
| **(Margarido Mendes, et al., 2020)** | AR training assistance to pinpoint insertion of intravenous needles | Syringe, Seldinger needle, Physical Simulator of upper torso and neck, Aryzon SDK, HMD device with a slot for smartphone, Vuforia SDK, Unity3D | Comparative study. Performance through task completion time and number of needle insertion errors; Demographics questionnaire, Satisfaction questionnaire with a 6-level Likert scale; General questionnaire for face and content validity, satisfaction, and perceived workload (NASA-TLX); Semi-structured interview | 18 participants (attending specialists and medical residents) |

| | | | |
|---|---|---|---|
| **(Nausheen, et al., 2020)** | AR during early medical training of point-of-care ultrasound | Hololens | NONE |

Among the examples summarised in *Table **1***, the study by Kotranza et al. (2012) is noteworthy.

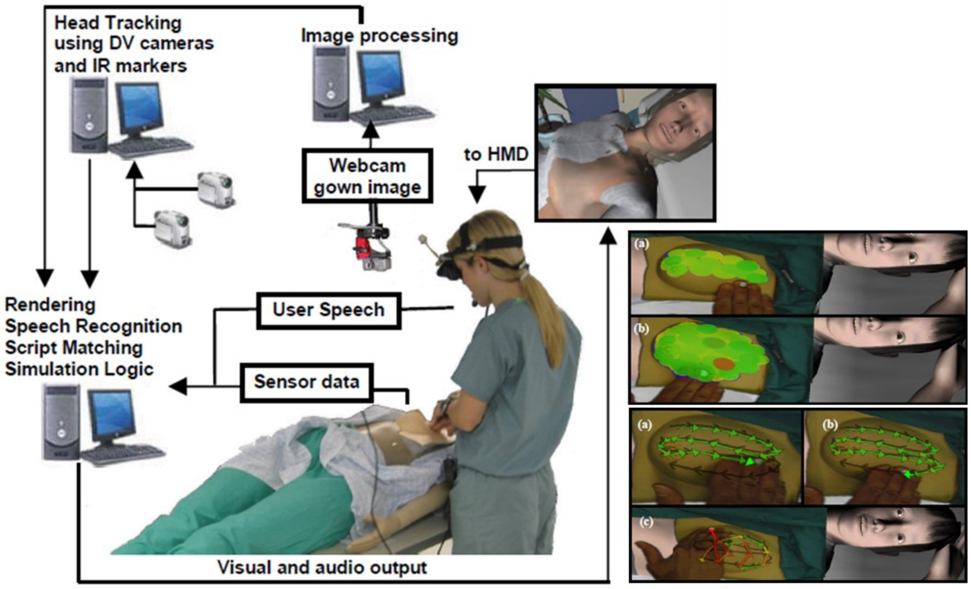


*Figure 6: System design by Kotranza et al.* (Kotranza, et al., 2009), (Kotranza, et al., 2009)

Even if in this case Mixed Reality is interpreted as the combination of physical and virtual reality, this extensive work has an interesting scope. Indeed, the authors developed an MR environment for teaching clinical breast examination, focusing on enhancing learner communication skills. During the examination, the system processes student's gestures and motions applied to the task trainer (equipped with force sensors), and the virtual patient provides appropriate responses, showing anxiety and distress. Moreover, real-time feedback is provided to the learner based on quantitative measures of correct palpation pressure and pattern of mass search (*Figure 6*). The system appeared to have significant educational benefits applied to cognitive and psychomotor tasks. Authors expected even that, through repeated use of the system, learners would be able to decrease their anxiety during real practice with human patients (Kotranza, et al., 2012), (Kotranza, et al., 2009).

However, generally, AR and MR are used with other objectives. As result in *Table **1***, seven studies (over a total of twelve) apply augmented reality for helping learners in needle insertion. Even if the medical procedures are different, five of them use AR to overlay on the skill trainer the internal human anatomy, useful for that kind of procedure (mainly the circulatory system (*Figure 7*)). The other two papers concern the use of haptic devices to give sensorial feedback.

Kobayashi et al. (2017) developed, for acute care procedure training, an AR/MR application that, using Hololens®, overlaid task-relevant anatomy images over the skill trainer. Wirelessly connecting a laptop to Hololens®, they showed the learner's view with superimposed holoimages to the viewer cohort *(Figure 7.a)*. The authors claimed that the manual overlay and registration process was the main technical limitation of this work. They manually scaled and registered the images onto the task trainer using surface anatomy landmarks but the spatial alignment between the image and the task trainer was intermittently lost. Concerning the assessment of the AR/MR application on trainees' performance, they did not make a comparative study with the traditional training method, but they mentioned the examination of the educational utility in their future work.


*Figure 7: AR for internal human anatomy superimposition (useful for needle insertion). (a) (Kobayashi, et al., 2017), (b) (Rochlen, et al., 2017), (c) (Margarido Mendes, et al., 2020)*

Rochlen et al. (2017) developed an AR application for the training of needle insertion for the specific procedure of central venous catheter placement. Using the smart glasses Epson Moverio BT-200®, they projected the relevant internal anatomical landmarks over the skill trainer (*Figure 7.b)*. The system was tested with forty subjects (anaesthesiology residents and medical students). Even if it resulted usable and feasible, some constraints were found: first, the alignment of images and the glasses fit. Moreover, this pilot study focused only on usability and feasibility and involved a small sample of learners. Future works should include the comparison of the AR technology with standard simulation training models (Rochlen, et al., 2017).

Also, Margarido Mendes et al. (2020) developed a system to pinpoint the insertion of an intravenous needle for central venous catheterization. This system supports not only the projection of internal anatomy over the skill trainer but also geometrical information about the position and orientation of the needle (*Figure 7.c*). The authors performed a comparative study with eighteen participants (attending specialists and medical residents) to assess the benefits of the AR system over the traditional training practice. Results showed that the new AR system is suitable to complement conventional training. However, some drawbacks emerged, related especially to technology. The main limitations are related to the tracking system that is affected by external factors (such as light), the virtual elements that are not stable and sometimes change position or disappear, the HMD that cause discomfort and has a limited field of view (Margarido Mendes, et al., 2020).

The core of these presented studies is always the use of AR as an aid in procedural training, giving information about what learners cannot actually see. This means that the technology is used to provide additional visual material, which is not available in real practice. In this way, the simulation moves away from reality.

Another example of application is the one presented by Bottino et al. (2018) in their ongoing work. They developed an interactive MR system to provide real-time feedback and cognitive aid to the learners during the Basic Life Support and Defibrillation (BLSD) procedure (*Figure 8*).

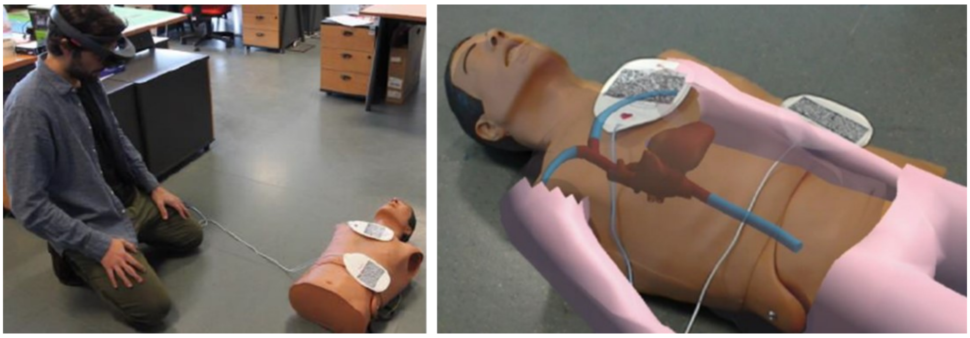

*Figure 8: MR System for Basic Life Support Defibrillation Training (Bottino, et al., 2018)*

The physical simulator was “augmented” by virtual realistic scenarios, where virtual characters (such as medical team members or patient relatives) deliver information and assist the trainee during the procedure. In this way, the reduction of teacher intervention and cost are fostered. Through the qualitative system’s evaluation with twenty-three users, it emerged that the main limitation is related to hand recognition that reduces the feeling of having control over the system (Bottino, et al., 2018).

Therefore, as resulting from the above-mentioned scientific literature, the principal drawbacks in using AR systems in training activities are related to the technology itself. The issues of registration and overlay, tracking and alignment of virtual content, the stability of virtual elements, hand recognition, discomfort, and limited field of view of HMDs should be solved with an advancement of the technology.

Another aspect that deserves more investigation is the educational utility of this kind of system, which moves the learners away from the actual procedure, instead of enhancing the realism of the simulation.

### *2.2.3. Assessments and Perspectives*

In the last years, several studies have tried to integrate augmented reality and manikins obtaining simulators in mixed reality. The main use of AR for learning is to offer immersion in a scenario and provide feedback or additional information. Several kinds of AR/MR system assessment emerge from the scientific literature. In the following paragraphs, a summary of systems' analysis and evaluation is reported, collecting benefits and drawbacks in the use of mixed reality simulation in medical training, and defining some possible perspectives.

#### 2.2.3.1. AR Assessment and Validation

All methods and tools developed for healthcare education and training should be assessed for their validity according to several consensus criteria. The full validation process of augmented reality applications (ARA) consists of five steps, as described in *Figure 9*. *Face validity* refers to the degree an ARA resembles the real working situation and is evaluated by novices and experts through a questionnaire after the use of the system. *Content validity* relates to the correctness and accuracy of the educational content and it is evaluated by experts through a questionnaire. *Construct validity* is defined by the difference in outcome between experts and novices and is evaluated using comparative studies and statistical analysis of the results. *Concurrent validity* refers to the performance improvement related to the AR application compared to an established training method (gold standard). Finally, *predictive validity* must be assessed through randomized controlled trials, to ensure that the skills acquired with the ARA can be translated in the real practice (Barsom, et al., 2016).

| Stages of validity | Description | Criteria for achievement | Appropriate method of examination |
|---|---|---|---|
| 1. Face validity | The degree of resemblance between an ARA and the educational construct as assessed by medical experts (referents) and novices (trainees) | Uniform and positive evaluation of the resemblance between the ARA with the educational construct among novice and expert medical professionals | Questionnaire after use of the ARA |
| 2. Content validity | The degree to which the ARA content adequately covers the dimensions of the medical content it aims to educate (or is associated with) ('the truth whole truth and nothing but the truth') | Uniform and positive evaluation of the ARA content and associated testing parameters by panel considered to be experts in the field | Questionnaire considering the content of the ARA |
| 3. Construct validity | Inherent difference in outcome between experts and novices on outcome parameters relevant to the educational construct | Outcome differences considered to be of statistical significance between subjects considered to be of different levels of skill | Comparative study measuring the relevant outcome parameters on the ARA for subjects with presumed different levels of expertise in the educational construct. |
| 4. Concurrent validity | Concordance of subject outcome parameters using tie ARA compared to outcome parameters on an established instrument or method, believed to measure the same educational construct (preferably the golden standard) training method) | Study results show correlation considered to be significant between ARA and the alternative, established training method | Comparative study comparing the outcome parameters of two different training methods in the same study participants |
| 5. Predictive validity | The degree of concordance of ARA outcome parameters and subjects' performance on the educational construct it aims to resemble in reality | Metrics show correlation considered to be significant between relevant outcome parameters on ARA and performance on educational construct it aims to resemble in reality | Randomized controlled trial comparing performance on educational construct in reality before/after training on ARA and control group using another training method |

*Figure 9: Matrix of validity type for augmented reality systems used for the training and education of healthcare professionals (Barsom, et al., 2016)*

The training AR instrument should be implemented only if all stages of the validation process have been positively assessed. However, to the best of our knowledge, there are no significant studies, in scientific literature, about predictive validity. Face, content, construct, and concurrent validities were recorded by Tai et al. (2019) using objective metrics and the Global Rating Scale questionnaire (Tai, et al., 2019). The same kinds of validity were assessed also by Margarido Mendes et al. (2020) to analyse the impact and reliability of their AR training application for needle insertion.

The assessment of AR/MR applications in medical training has had a notable increase in the very last few years (2016-2020), simultaneously and after the beginning of the present study (2017). The following list represents an overview of the metrics and parameters used to assess systems' usability, learners' performance, emotional state, and satisfaction in most of the state-of-the-art studies collected in recent reviews.

- Demographics Information: gender, level of expertise, and previous AR experience [by questionnaires] (Margarido Mendes, et al., 2020);
- Technical Assessment: voice and gesture interaction, virtual contents realism and field of view [Likert-scale questionnaire] (Chaballout, et al., 2016), (Bottino, et al., 2018);
- Performance: task completion time, number of errors (Margarido Mendes, et al., 2020), complication rates (Gerup, et al., 2020) [by author-developed Likert scale questionnaires or procedural-based checklists];
- Learning Experience and User Acceptance: easiness to identify anatomical landmarks, ease of use, learnability, usefulness, reliability, easiness for debriefing [by Likert scales + System Usability Scale (SUS) + semi-structured interview] (Bottino, et al., 2018), (Margarido Mendes, et al., 2020), (Gerup, et al., 2020);
- Emotional State: perceived workload [by NASA-TLX questionnaire + semi-structured interview] (Margarido Mendes, et al., 2020), perceived cognitive load, stress response, adverse health effects, and ergonomics [by questionnaire] (Bottino, et al., 2018), (Gerup, et al., 2020).

However, none of these authors includes all the parameters in their simulation effectiveness assessment studies, resulting only in partial evaluations.

#### 2.2.3.2. AR Pros and Cons

An integrative review of more than 2.500 papers found that 96% claimed AR to be useful for improving healthcare education (Zhu, et al., 2014). Additionally, authors were able to determine that AR increased the speed at which students learned and made the learning process easier. Several positive aspects and benefits were elicited in different studies. The following list supplies an overview of the advantages, provided by learners and teachers, applying augmented and mixed reality simulators in healthcare education:

- Decreased amount of practice needed (Zhu, et al., 2014)
- Decreased amount of errors (Munzer, et al., 2019)
- Reduced failure rate (Zhu, et al., 2014)
- Reduced simulation time (Munzer, et al., 2019) (Gerup, et al., 2020)
- Improved performance accuracy (Zhu, et al., 2014), (Garzón, et al., 2017)
- Shortened learning curve (Zhu, et al., 2014)
- Easier to capture learner's attention (Zhu, et al., 2014), (Munzer, et al., 2019), (Gerup, et al., 2020)
- Increased motivation (Zhu, et al., 2014), (Garzón, et al., 2017), (Gerup, et al., 2020)
- Improved assessment of trainees (Zhu, et al., 2014)
- Better understanding of spatial relationships (Zhu, et al., 2014)

- Enhanced learning retention and performance on cognitive-psychomotor tasks (Zhu, et al., 2014), (Herron, 2016), (Garzón, et al., 2017), (Munzer, et al., 2019), (Gerup, et al., 2020)
- Decreased cognitive load (Munzer, et al., 2019)

Therefore, augmented reality allows for more authentic learning, making the simulations more realistic and immersive (Zhu, et al., 2014), (Herron, 2016) and providing students a more personalized and explorative learning experience (Zhu, et al., 2014). It seems to be useful also in achieving core competencies, such as decision making and teamwork (Zhu, et al., 2014).

Although most users are very much in favour of these new technologies (Zhu, et al., 2014), (Herron, 2016) because of their easy and enjoyable use (Munzer, et al., 2019), the AR/MR applications are not as widely accepted as they perhaps should be (Herron, 2016), (Garzón, et al., 2017). The main reasons are reported in the following list:

- Technical and Usability issues: difficulty in looking at both the real and virtual environments (Herron, 2016), delay between real and virtual environments (Chaballout, et al., 2016), cumbersome hand swipes and gestures, distraction through multiple images, overheating of hardware, difficulties with establishing an Internet connection (Munzer, et al., 2019)
- The time needed to train students on how to use AR/MR applications (Chaballout, et al., 2016)
- The teacher resistance (Garzón, et al., 2017)
- The pedagogical issues (Garzón, et al., 2017)

The great amount of research in AR spans across different medical areas, learners' levels, and outcome focus, with growing evidence for improving learning. However, the MR advanced interactive training in healthcare education still presents some weaknesses that can be summarised as follows:

- Lack of explicit pedagogical theoretical framework to guide the design (80% of papers did not describe which kind of learning theory was used to guide design and application) (Zhu, et al., 2014);
- Traditional learning strategies applied (in 64% of papers, advanced technology was used as a guidance system or as a feedback tool) (Zhu, et al., 2014);
- Mostly applications prototypes reported (56% papers presented prototypes without studying their impact) (Zhu, et al., 2014), (Gerup, et al., 2020);
- Technological limitations and poor ergonomics (limited computing power, occlusion of the user's field of view, poor ergonomics, possible cognitive overload) (Garzón, et al., 2017), (Gerup, et al., 2020);
- Lack of strong evidence for improving learning (no statistical analysis of significance) and shortcomings of the study designs for transferability in real practice (Gerup, et al., 2020).

#### 2.2.3.3. Perspectives and Future Directions

At the beginning of this study, in 2017, the perspectives of AR applications in medical education were drawn in various research papers. Uniform and standard assessment strategies, and complete validation tests were needed to implement ARA, reliably and validly, in educational curricula (Barsom, et al., 2016). Defined metrics for the assessment of operational quality (e.g. image registration accuracy, stability, and usability) and procedural performance (e.g. checklist-based assessments) in simulated settings had to be established and validated (Kobayashi, et al., 2017). It was important to verify that AR systems satisfied the real purposes of education, complementing the learning process (Garzón, et al., 2017).

More recent studies (2019-2020) still confirm that Augmented Reality needs to be investigated more robustly (Munzer, et al., 2019), since rigorous, objective measurements of clinical procedural skills, and human performance metrics continue to be very limited or absent (Linde, et al., 2019). A throughout investigation of the educational context, learner types, and learning objectives (e.g. cognitive, technical, or non-technical such as measuring situational awareness, communication, or stress coping) must be implemented (Gerup, et al., 2020). The evaluation of learning curve improvement still has to be examined in depth (Tai, et al., 2019). Moreover, most studies about AR effectiveness were conducted only on small cohorts of participants (Munzer, et al., 2019) and statistical analyses reported incomplete or misinterpreted results (Gerup, et al., 2020). There is also little information about AR usability in the healthcare setting (Munzer, et al., 2019). Finally, a considerable shortcoming is the wide heterogeneity among research designs and outcome measurements. Establishing guidelines and standard methodologies to analyse and assess outcomes in medical education due to the use of AR technology, would lead to higher-quality studies (Gerup, et al., 2020).

## 2.3. Human Factors and Cognitive Ergonomics

The design of medical simulators and simulations must consider numerous requirements, merging both technical and social aspects, in order to work properly and satisfy the users' needs.

Overall, User Experience (UX) is based on the personal perceptions and responses that result from the use of a product, system, or service, including users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours, and accomplishments that occur before, during and after use (ISO 9241-210, 2010).

Human Factors and Ergonomics have been introduced in engineering in order to consider the physical, psychological, social, and cultural needs of human beings, during the product/system design, development, and assessment processes (ISO 9241-210, 2010). Human factors specifically refer to research "regarding human psychological, social, physical, and biological characteristics and working to apply that information with respect to the design, operation, or use of products or systems for optimizing human performance, health and safety" (Stramler, 1993). Therefore, human factors are fundamental not only in the design of a system but also in the evaluation of the human-machine interaction.

The United States Food and Drug Administration (FDA) defines human factors as the study of how people use technology and affirms that human factors engineering helps improve human performance and reduce the risks associated with user errors. In particular, the FDA has defined methods for the promotion of "patient safety and safety in medical device use" through the examination of the interface between healthcare practitioners and technology. Indeed Medicine, as an industry in which human lives depend on the skill and performance of operators, must create and maintain a culture of safety, in addition to designing systems to mitigate errors. Medical simulation and human factors engineering can be used to examine and to enhance the interface between healthcare practitioners and medical technology, with the potential to make a significant contribution to patient safety (Hunt, et al., 2006).

The main aim of Human Factors and Ergonomics is to guarantee human comfort and safety, and consequently to improve user performance. Indeed, the physical and cognitive factors that can affect the users' performance and the quality of human-machine interaction are several: from physical and mental workload to task complexity, the overload of information, or time pressure. Moreover, the response to the same stimuli differs among users, being different the capabilities of everyone. Therefore, the optimization of physical and mental workload, comfort, and perceived effort is necessary to prevent disorders and stressful conditions, assuring the best human performances (Pheasant, 1999). For this reason, human factors are often applied also to the issue of effective training, which includes the effective use of simulators, as well as techniques for assessing performance and improving learning. However, even if education in healthcare focuses on high-stakes environments and the acquisition of complex manual and cognitive skills, human factors are not well integrated and adopted into medical training (Seagull, 2012). Indeed, the number of successful implementation and development of medical simulations is relatively small compared with the manufacturing industry. If the system and interface design is not designed with human capabilities and by considering the limitations of the cognitive, perception, and physical human factors, physician, operators, and healthcare providers are placed in situations where the imposed demands are unrealistic from a psychological point of view, resulting in inevitable errors (AlRomi, 2015).

Moreover, with the application of augmented reality, the digital information that augments the experience of the user in the real world must be presented without distracting or overloading the user or making the task more difficult (Webb, et al., 2016).

In this context, the adoption of a human-centred and ergonomic approach is compulsory for the creation of successful training paths and simulators. The practice of complex medical procedures (both for the training of technical and non-technical skills) is a stressful activity, physically and mentally. The eventual use of advanced technological devices, such as HMD for augmented reality, requires an additional mental and physical burden, demanding different skills and experience. For this reason, in this scenario, it is evident the need to assess and integrate ergonomics in simulation design.

As defined by Wilson, "ergonomics is the theoretical and fundamental understanding of human behaviour and performance in purposeful interacting sociotechnical systems, and the application of that understanding to design of interactions in the context of real settings" (Wilson, 2000). In other words, ergonomics is a multidisciplinary science aimed at studying the functions and the interaction between the following three elements that constitute a working system:

- Human: user assigned to carry out a specific task, both in physical (anthropometric characteristics, biomechanical aspects, etc.) and cognitive (induced mental load, social interactions, psychological factors, mental processes) senses;
- Machine: equipment, devices, and tools used to perform the assigned task or to manage information flows;
- Environment: the set of characteristics of the place where the task is performed (layout of the space, workstation, or room used for the performance of the activity).

To improve system performance and user satisfaction, well-being, and safety, ergonomics embraces four main domains: physical, cognitive, environmental, and organizational. This work focuses on cognitive ergonomics.

Cognitive ergonomics originates from the concept of cognitive engineering. As claimed by Norman (1987) "Cognitive engineering is meant to combine with the applied disciplines not to replace them […] A new approach ... more than just psychology... more than psychology coupled with engineering. We need all the disciplines of cognitive science, plus engineering" (Norman, 1987). Cognitive ergonomics involves psychological processes such as awareness, understanding, human information elaboration, reasoning, and use of knowledge, as it concerns human interacting with other system components. Some significant topics include workload, decision-making, perception, attention, motor response, skill, memory, and learning. It is oriented towards the optimization of human-machine interaction, according to three main criteria: characteristics of human cognitive processes, software science knowledge, and knowledge in diverse work domain technologies. As a logical consequence, the training topic is included in such perspective, since it can contribute to the enhancement of human performances and work conditions (Green, et al., 1991). Indeed, sometimes, even based on a lot of experience, human beings can misinterpret information, make mistakes or make wrong choices, which may have fatal consequences for people's health and safety. The increase in professional activities that have a "mental dimension" has therefore encouraged the development of cognitive ergonomics, which thus results fundamental in the design and assessment of medical training. Indeed, its objective is to improve the performance of cognitive tasks in dynamic and technologically advanced environments, through the design of effective support, understanding the fundamental principles of human activities associated with the principles of engineering design and development.

The first domains studied by cognitive ergonomics have been nuclear power plants and air traffic control systems because of their nature of complex environments with potentially life-threatening situations. In the following years, many studies have been conducted on "softer" domains, such as banks, office work, and recreational activities, to which its principles proved to be transferable. It has been applied even in the healthcare training domain, but a standard, overall, and complete cognitive ergonomics assessment is still missing in this field. Therefore, it is evident the need to design and develop a standard methodology and procedure to analyse the mental state of learners during medical simulations, in order to avoid mental overloading and stressful situations, ensuring productive learning.

### *2.3.1. Cognitive Load in Medical Simulations*

The analysis of Cognitive Load (CL) is one of the most widely studied topics in cognitive ergonomics. However, although performance measures are strongly used in the field of medical education to evaluate the skills of trainees and medical students, the assessment of their cognitive state is relatively "uncommon". This fact is disadvantageous if we consider the introduction of technologies as advanced simulators and augmented/virtual reality devices, which may represent an improvement in the students' immersion in the simulated scenario or, conversely, potential risk of information overload (Atalay, et al., 2016). Indeed, even if it has been demonstrated that AR can support learning and teaching, the comparison among reviewed research studies shows some conflicting conclusions. While some studies reported that AR decreases cognitive load, others stated that it causes cognitive overload (Akçayır, et al., 2017). Therefore, the study of cognitive load, related to AR technology applications, merits further in-depth analysis. In this context, a precise assessment of cognitive conditions results in an essential element in the design process of medical training sessions.

For medical students, the training phase is a decisive moment to practice and learn indispensable technical and non-technical skills. Training on advanced simulators, even without extended reality applications, has emerged as an effective way of complementing the clinical training, offering conditions that are optimal for skills acquisition (Fraser, et al., 2012). However, both technical and non-technical skills can be adversely affected by the high demand on the cognitive resources of the learners, caused by these complex learning environments (Charles, et al., 2019). From this perspective, the assessment of students' cognitive load becomes a key aspect, and its proper management becomes an essential component of medical education.

Despite over 40 years of research, there is still no clear and universally accepted definition of human cognitive load. Indeed, CL is often described by terms like 'mental workload' (MWL), 'cognitive effort', 'mental strain', or 'mental effort'. Mental workload "emerges from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills, behaviours, and perceptions of the operator" (Hart, 2006).

According to the cognitive load theory (CLT), the learning process consists of developing cognitive patterns and storing them in unlimited long-term memory (Sweller, et al., 1998). Indeed, the concept of mental workload assumes that each person has a relatively limited cognitive capacity (called working memory) that deals with auditory, verbal, and visual material. This capacity is likened to a pool from which resources can be drawn to meet the demands of ongoing tasks (Wickens, et al., 2004). Given this assumption, the mental workload for a given task is the relationship between the required mental resources and the total resources available, moment by moment. In other words, the mental workload is inversely related to spare capacity when performing the task of interest (Carswell, et al., 2005). Thus, during the learning process, the limits of human working memory must be considered and overcome by creating a comfortable environment which encourages the schemas construction, and by lowering the number of elements not strictly connected with the content goals (Kalakoski, et al., 2019). Indeed, CL can be divided in:

- Intrinsic Cognitive Load (ICL): that depends on the task itself and the learner's prior knowledge or experience on the task;
- Extraneous Cognitive Load (ECL): that is related to external useless factors and can damage the learning process overloading the working memory;
- German Cognitive Load (GCL): that is not intrinsic to the task itself but is directly connected to schemas construction, with benefits on learning (Sweller, et al., 1998), (Fraser, et al., 2012).

Cognitive load theory assumes that ICL and ECL are additive.

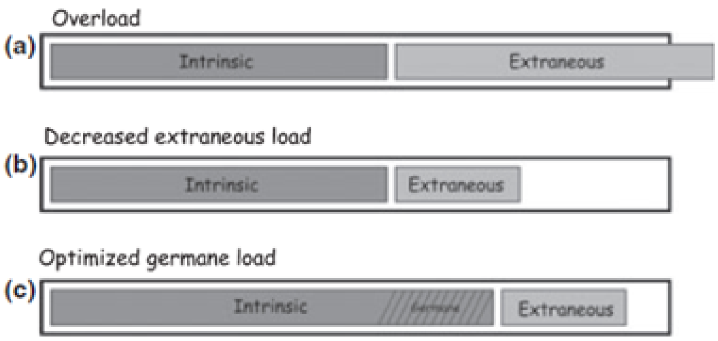


*Figure 10: The additive nature of cognitive load (Van Merrienboer, et al., 2010)*

In the case of complex tasks training, the sum of ICL and ECL may easily surpass working memory capacity and yield overload (*Figure 10(a)*). Extraneous cognitive load must be lowered as much as possible in order to induce a greater germane cognitive load and optimise the learning process (*Figure 10(c)*) (Van Merrienboer, et al., 2010).

Regarding medical simulation training, Fraser et al. (2012) found out that learners with limited clinical experience had a high cognitive load and were at risk for cognitive overload, with a resultant decline in learning. For this reason, they suggested to pursue further studies to understand how to best manage the different types of cognitive load, to maximise the potential gains in performance. Also, Tremblay et al. (2019) asserted that simulated-clinical-immersion learning environment contributes to both intrinsic and extraneous cognitive load and found their sources in the lack of knowledge, unfamiliar resources, and time limitations (Tremblay, et al., 2019).

Since CL can positively or negatively affect human performances, the principal reason for measuring it is to quantify the mental cost of performing a task to predict the performances (Cain, 2007).

Concerning the MWL assessment methods, researchers agree in classifying them into three main broad categories (also according to the (ISO 10075-3, 2004)): performance assessment method, self-assessment (or subjective scaling) method, and physiological measurements method. [For completeness, it should be stated that in the (ISO 10075-3, 2004) there is a fourth technique for the assessment of MWL in the workplace: the job and task analysis. It consists of assessing task elements, physical and psychosocial work conditions, environmental conditions, and the organization of the work process].

#### 2.3.1.1. Performance Assessment Method

The class of task performance measures assumes that MWL is relevant only if it affects performance. For example, lowered and/or irregular performance may indicate that the user is reaching unacceptable levels of

MWL. These measures are usually divided into primary and secondary task measurements. In the primary task method, the performance is monitored and analysed according to changes in mental demand for the execution of the task. This method tries to deduct the mental load from the performance on the activity of interest. It uses techniques to directly quantify the ability to perform the primary task at an acceptable level. Examples of common measurement parameters are response, reaction time, accuracy, error rate, estimation time, objective speed, and signal detection (Karwowski, 2006). However, it is demonstrated that performance errors are not necessarily related to a high mental load imposed by the main activity. This is the reason why the secondary task method is more used. In this type of analysis, the user is required to perform a secondary activity concurrently with the main activity. In particular, the learner is asked to perform two tasks at the same time, and the secondary task is used to calculate the mental load associated with the primary task (Blanco, et al., 2006). To the best of my knowledge, the secondary task method in the medical education field is always associated with another CL assessment technique. Haji et al. presented three studies on this topic. The first one was about the sensitivity of secondary task performance and subjective ratings during simulation-based psychomotor skills training (Haji, et al., 2015). The other two studies were about the mental effort assessment, using the same two techniques (secondary task + subjective ratings), investigating the effects of variations in task complexity during simulation-based surgical skills training (Haji, et al., 2015), and simulation training of lumbar puncture (Haji, et al., 2016).

2.3.1.2. Self-Assessment Method

The category of self-assessment/subjective measures is based on the personal perceived experience about the interaction with the system and is obtained from the direct estimation of task difficulty. This kind of measure has always attracted many researchers because of the belief that only the individual concerned in the task can provide an accurate judgment for the experienced MWL (Woods, et al., 2018). The self-assessment provides information on how humans subjectively evaluate various aspects of workload for accomplishing a task, using questionnaires or psychometric scales. NASA Task Load Index (NASA-TLX) is a multidimensional assessment questionnaire that rates perceived workload under six different dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration levels. This is the most used tool for workload subjective assessment, and it is applied in several research studies about medical training and simulation (Muresan, et al., 2008), (Tomasko, et al., 2012), (Bosse, et al., 2015), (Bhandary, et al., 2016). The Surgery Task Load Index (SURG-TLX) is a modified version of NASA-TLX applied in the specific field of surgery. This surgery-specific, multi-dimensional workload measure enables subjective assessments of load relevant to a specific task, and it is based on five items: mental demand, physical demand, temporal demands, complexity, and situational stress (Wilson, et al., 2011). For example, Wucherer et al. used SURG-TLX to evaluate the students' mental workload using an innovative simulator for vertebroplasty training. Through this analysis, they evidenced the necessity to develop realistic simulation

environments that prepare young medical to respond to emergent events in the operating room (Wucherer, et al., 2015). While methods such as NASA-TLX are more suitable for the analysis of the mental load of a single task, the Multiple Resources Questionnaire (MRQ) is used to measure the cognitive load in more complex training processes (Boles, et al., 2007). Other studies assessed mental effort, in medical simulation training, with a 9-point subjective rating scale, from very, very small to very, very high (Fraser, et al., 2012), (Fraser, et al., 2014), (Sewell, et al., 2019).
Naismith et al. evaluated three different self-assessment methods with two groups of medical residents after participating in simulation-based procedural skills training sessions. The three questionnaires were NASA-TLX, Paas Cognitive Load Scale, and a cognitive load component (CLC) questionnaire that they developed to assess total cognitive load as the sum of ICL, ECL, and GCL. They found out that NASA-TLX, Paas Scale, and CLC questionnaire were interchangeable for measuring ICL, but not for total CL (Naismith, et al., 2015). Another questionnaire for the assessment of the different parts of CL was evaluated by Cook et al. (Cook, et al., 2017).

#### 2.3.1.3. Physiological Assessment Method

The category of physiological measures considers physiological responses of the body that are believed to be correlated with MWL. Indeed, changes in psychophysiological parameters, such as heart rate (HR), heart rate variability (HRV), breathing rate (BR), galvanic skin response or electrodermal activity (GSR or EDA), brain activity (EEG), muscular activity

(EMG), eye activity (EOG, pupil diameter, gaze entropy, and velocity), can be indirect indicators of mental workload. These physiological parameters can be collected using wearable devices such as smart bands or bracelets for HR, HRV, BR, EDA monitoring, and smart glasses for eye-tracking.

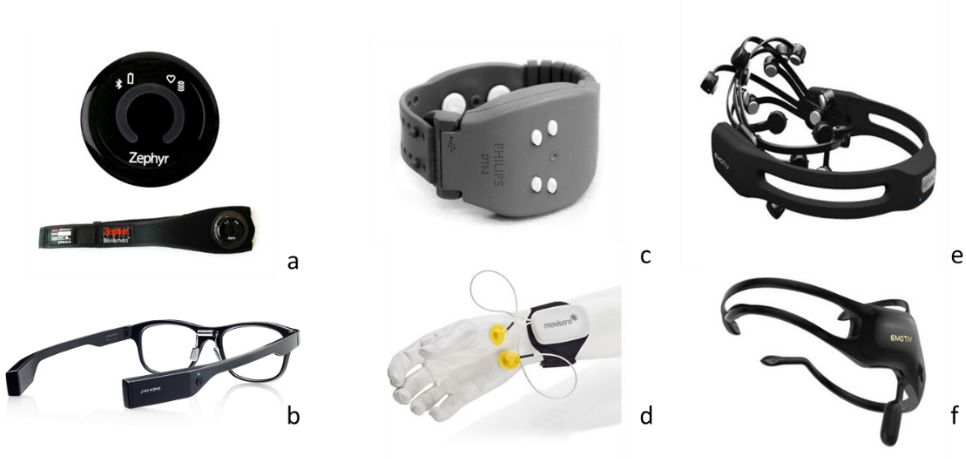


*Figure 11: Examples of smart wearables for biometric monitoring: (a) Zephyr BioHarness, (b) Jins Meme, (c) Philips DTI-2, (d) edaMove, (e) Emotiv EPOC+, (f) EMOTIV Insight 5 Channel Mobile EEG*

*Figure 11* shows few examples of commercially available wearable devices for the biometric monitoring. For example, the chest band Zephyr BioHarness allows for the recording of HR, HRV, BR, temperature, acceleration, posture. The smart glasses J!ns Meme is an electrooculogram embedded with a gyroscope, used for the analysis of eye movements (saccades, fixations, and blinks). The edaMove is a psycho-physiologic ambulatory measurement system that can detect and measure EDA, as well as the Philips DTI-2 (Discrete Tension Indicator) wristband sensor. Moreover, the DTI-2, combining multiple sensors, can measure also 3D acceleration, band temperature, skin temperature and ambient light. The Emotiv EPOC+

Neuroheadset is a neuro-signal acquisition personal interface for human and computer interaction. It uses sensors to detect electric signals produced by the brain to detect subject's thoughts, feelings and expressions. The Emotiv Insight is a wearable EEG headset that offers 5 EEG sensors and 2 reference sensors providing in-depth information on brain activity. The measurements are based on six key cognitive and emotional metrics: focus, stress, excitement, relaxation, interest and engagement. These measurements allow an individual to monitor their cognitive health and wellbeing.

Unfortunately, these devices are considered invasive in some circumstances and, for this reason, their application in the medical field is very limited. However, some examples are present in the scientific literature. For example, Di Stasi et al. analysed gaze entropy and velocity of surgical trainees and attending surgeons during two surgical procedures through a wearable eye-tracker device (Di Stasi, et al., 2017). An interesting pilot study investigated the potential of gaze-tracking technology to study decision-making and leadership behaviours in simulated medical emergencies. However, the authors stated that pupil dilation and micro-eye movement frequency are influenced by both cognitive load and emotional load, thus it could be difficult to distinguish CL from stress (Szulewski, et al., 2014). Lee et al. tested whether prior knowledge affected performance and mental effort in a medical simulation game for resuscitation skills training. In addition to using the NASA-TLX, they used eye-tracking measurements as an objective indicator of cognitive load. Statistical analysis revealed a significant difference between medical professionals and novices (Lee, et al., 2019). Another example that goes beyond training, is the one of Dias et al. (2019)

who presented a novel approach to characterize dynamic changes in team CL by measuring synchronization and entropy of HRV parameters during real-life cardiac surgery. CL was assessed by measuring inter-beat intervals (IBI) using an unobtrusive wearable heart rate sensor (Dias, et al., 2019).

Naismith and Cavalcanti (2015) conducted a review study to assess the validity of different CL measures across simulation training contexts, dividing medical education from other domains (*Figure 12*). Concerning medical education until 2014, most studies used self-report measures, others included secondary task performance, but none used physiological indices.

| | Medical education | | Other domains | |
|---|---|---|---|---|
| Cognitive load measure | No. of studies | Validity scores (mean ± SD) | No. of studies | Validity scores (mean ± SD) |
| **Self-report** | 10 | 1.35±0.53 | 24[a,b] | 1.48±0.70 |
| NASA Task Load Index | 5 | 1.40±0.58 | 9 | 1.44±0.44 |
| Paas scale | 3 | 1.33±0.58 | 9 | 1.33±0.96 |
| Other self-report | 3 | 1.17±0.29 | 6 | 1.75±0.52 |
| **Secondary task** | 3 | 1.50±0.50 | 4[a] | 1.38±0.75 |
| Reaction time | 2 | 1.50±0.71 | 2 | 2.00±0.00 |
| Memory task | — | — | 2 | 1.50±0.71 |
| Other task | 1 | 1.50 | 1 | 0.50 |
| **Physiological index** | — | — | 7[b] | 1.71±0.70 |
| Heart activity | — | — | 4 | 1.50±0.41 |
| Eye activity | — | — | 4 | 1.88±0.85 |
| Brain activity | — | — | 1 | 3.00 |
| **Observer rating** | — | — | 3 | 2.38±0.75 |

*Figure 12: Cognitive Load Measures in Medical Education and Other Domains (Naismith, et al., 2015)*

Although the high levels of inter- and intra-individual variability of biometric indices, current findings suggest that physiological parameters are the most sensitive means for detecting variations in cognitive load levels during simulation training.

However, correlations between CL and learning among studies varied from positive to negative with greater evidence that high cognitive load impairs learning (Naismith, et al., 2015).

The findings of this review suggest using multiple concurrent measurements to increase the validity of cognitive load assessment and help improve the rigor of studies in medical education and simulation training (Naismith, et al., 2015).

A more recent systematic review analysed the different methods used to assess cognitive load in the specific field of surgery. Most studies (70%) were carried out in simulated settings. 73% of them used self-report methods, of which NASA-TLX was the most applied tool (52%), while heart rate variability analysis was the most used objective method (13%) (Dias, et al., 2018). Self-report instruments are suitable for the evaluation of mental workload over long periods of time and are not very sensitive for rapid and short-lasting CL changes. When the aim is to assess cognitive load related to specific task phases, real-time tools should be used, since they allow capture of cognitive load fluctuations. Moreover, even if subjective measures present low application costs and lack of interference with on-going tasks, they present some limitations, mostly linked to the difficulty in trying to quantify the mental effort perceived in a task. Therefore, a combination of both subjective and objective methods might provide an enhanced measuring of mental conditions (Dias, et al., 2018).

From a last more specific review about the assessment of students' cognitive state in medical simulation training, 78% of the analysed papers applied only one method to measure the cognitive load, while the remaining 22% used two methods. Among the 3 different CL assessment techniques, the self-assessment method was the most applied (72%), with the NASA-TLX as the most used questionnaire (56%). The method of task performance measure was applied in 39% of paper and the tools used by all of them is the secondary task measure. The method of psychophysiological measures was applied only in 11% of studies (Scafà, et al., 2019).

Therefore, being the subjective measurements the most used MWL assessment method, it is clear the need to develop a more comprehensive and systematic methodology to analyse the cognitive load more objectively. The three assessment methods should be combined in a standard protocol designed to be less invasive as possible, without interfering with learners' actions and performance.

### *2.3.2. Stress in Medical Simulations*

Another important and underestimated topic is the evaluation of stress in medical simulations. Indeed, medical training can generate excessive levels of acute stress or emotional states of anxiety and worry, that can compromise students' performance and health (Dias, et al., 2018). Moreover, excessive levels of stress in real practice results in a greater amount of risk for the patient. For this reason, it is important to simulate stressful events in medical training, and learn how to manage them, avoiding episodes of acute stress.

Therefore, it is worth considering the influence that elements such as high fidelity, immersion, realism and even the use of AR/MR applications have on the emotions of students, assessing their effect on performance and learning. Indeed, MR training systems enable learners to experience also rare extreme situations, and, especially in these contexts, learners' stress levels should be measured and taken into account to optimise the learning process (Szu-Chi Chen, et al., 2018).

During the simulation sessions (even without AR/MR applications), trainees can feel discomfort, fatigue, and stress. Commonly recognized stressors include technical complications, time pressure, distractions, interruptions, errors, and increased workload (Arora, et al., 2009). For this reason, it is important to monitor stress levels during medical training and simulations.

According to the (ISO 10075-1, 2017), psychological stress is the effect of all conditions with a mental impact on a subject, either cognitive or emotional. It emerges when the perceived demands of the environment exceed a person's ability to cope with these demands (Lazarus, et al., 1984). Stress is also defined as a "state of high general arousal and negatively tuned but unspecific emotion, which appears as a consequence of stressors acting upon individuals" (Boucsein, 2012).

Arousal, in its most general sense, refers to the subject's readiness for performing a task. It has been demonstrated that an optimal level of arousal helps in reaching better performance, while non-optimal levels may create fatigue, sleepiness, or, at the opposite, strong stress and anxiety (*Figure 13*).

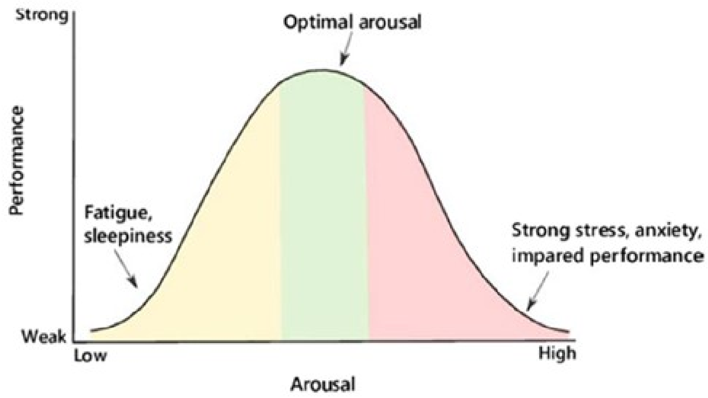


*Figure 13: Yerkes-Dodson law of arousal: Inverted U-Model* (Yerkes, et al., 1908)

The relationship between the level of arousal and the performance of the individual is traditionally represented by the inverted U-shaped function (*Figure 13*). However, this relationship varies from person to person and is thus challenging to measure.

Another traditional model for the relationship between arousal and emotion is the affective circumflex emotional plane (or Russell's model) (Feldman Barrett, et al., 1998). A particular emotion perceived by the subject can be positioned on this plane based on the values of the variables used for the analysis: EDA signal can be used as an index of the excitement of the perceived emotion (arousal) and HR signal as an index of the pleasantness of the emotion perceived by the subject (valence).

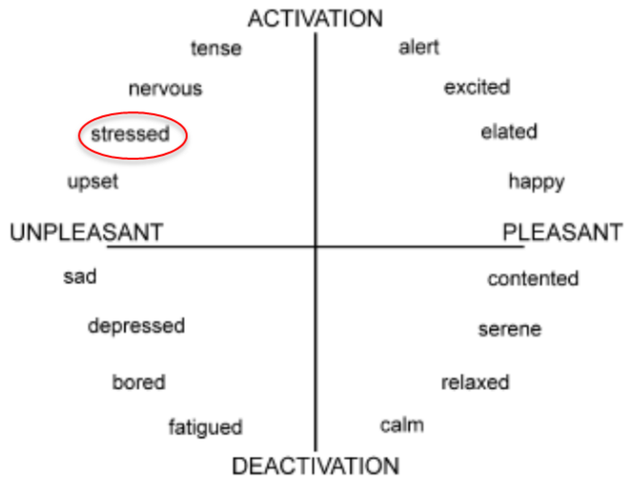


*Figure 14: The affective circumplex depicting each emotion along continuous dimensions of arousal (Feldman Barrett, et al., 1998)*

Therefore, according to Russell's circumplex plan, the condition of the subject can be classified as "stressed" when EDA is above the mean value (arousal) and HR is simultaneously below the minimum value (valence).

Stress is a concept that includes a wide spectrum of variables and cognitive processes and, for this reason, can be misinterpreted or confused with other kinds of negative emotions. In the medical context, it is usually described as two general types of response: anxiety or frustration, and the physiological response of the sympathetic nervous system which emerges after a challenge or threat. The first category is explained by Bjørshol et al. (2011), who exposed students to socioemotional stress in a simulated emergency resuscitation situation and found out that their subjective workload and feelings of frustration increased (Bjørshol, et al., 2011). Concerning the second category, it has been demonstrated that stress causes reactions such as changes in skin conductance (sweating), heart rate (tachycardia), blood pressure (increase), and in the stress hormone cortisol (increase) that spreads to saliva within minutes, during and immediately after performing a stressful task (LeBlanc, et al., 2008), (Sandroni, et al., 2005).

The multimodal dimension of stress makes the research field very broad; however, according to (ISO 10075-3, 2004), four main criteria can be distinguished in detecting stress: psychological, physiological, behavioural, and biochemical. The most common analyses typically include the subjective assessment based on self-report and physiological assessment based on ECG (for heart rate monitoring) and skin conductivity (to measure sweat activity).

#### 2.3.2.1. Self-Assessment Method

One of the most used scales to measure anxiety in student populations is the State-Trait Anxiety Inventory (STAI). This test consists of two separate, self-report scales for measuring the distinct concepts of state and trait anxiety (Takaia, et al., 2004). While the trait anxiety reflects a predisposition to anxiety as determined by the personality pattern, the state anxiety reflects the feelings that the subject is experiencing in that precise moment.

The Short Stress State Questionnaire (SSSQ) has been tested by Helton (2004) through two studies providing initial psychometric and validation evidence of a short multidimensional self-report measure of the stress state. He aimed to prove the validity of information on the SSSQ regarding its sensitivity to task stressors (Helton, 2004).

The Borg score of the Perceived Exertion scale (BORG) is a tool for measuring perceived mental fatigue during physical work. In one study, participants rated their exertion during activity, combining all sensations and feelings of stress and fatigue. They were told to disregard any one factor such as leg pain or shortness of breath and to try to focus on the whole feeling of exertion (Williams, 2017).

Singh et al. (2016) administered the Perceived Stress Scale (a widely used 10-item psychological tool) to one hundred students, in their final year of medical training, to measure the degree to which situations in an individual's life are perceived as stressful. They administered also two other questionnaires related to burnout and coping behaviour. They found out that a higher score on perceived stress was associated with higher scores on general psychopathology and burnout (Singh, et al., 2016).

#### 2.3.2.2. Physiological Assessment Method

Human stress can be measured even through heart rate (HR) (DeMaria, et al., 2010), (Hunziker, et al., 2012), (Sandroni, et al., 2005), (Waller, et al., 2017), electrodermal activity (EDA) (Boucsein, 2012), and salivary cortisol (Hunziker, et al., 2012), (Müller, et al., 2009).

The electrodermal activity reflects the surface changes in skin conductance due to the sympathetic nervous system and it is considered "one of the most sensitive psychophysiological indicators of stress" in medical settings (Boucsein, 2012).

Cortisol concentration has been used as a biological indicator of stress in most psychobiological research. However, it presents several limitations, and the interpretation of the results is challenging.

The cardiac pattern was used by Gouin et al. to assess if repeated high-fidelity simulations decrease stress level and increase performance, in anaesthesiology registrars. Physiological stress was evaluated via the maximal heart rate measured by a Holter system, and perceived stress was estimated by self-assessment (numerical scale from 0 to 10). While physiological stress remains unchanged with repeating simulation sessions, they observed a reduction in perceived stress levels (Gouin, et al., 2017). Marjanovic et al. (2018) assessed emotional excitation during simulated endotracheal intubation, using several physiological (HRV, EDA, and eye-tracking) and psycho-cognitive patterns (10 points Likert-type scale, and self-assessment stress level evaluation) (Marjanovic, et al., 2018). Theodoraki et al. (2015) presented a case in which students were connected to a biofeedback device that measures heart rate, heart rate variability, respiratory rate, and EMG. Stress situations have been identified by an increase in heart rate and a decrease in heart rate variability (Theodoraki, et al., 2015). Nevertheless, attention should be paid to the task performed. For example, Hunziker et al. warned of the limiting value of HR in cardiopulmonary resuscitation settings, due to the influences of physical activity, such as giving compressions (Hunziker, et al., 2012).

Moreover, physiological reactions emerge while experiencing both distress (negative stress), and eustress (positive stress) (Boucsein, 2012), and, through the monitoring of physiological signals, only the intensity can be assessed, not the valence. For this reason, studies found that the objectively measured arousal using HR, EDA, or cortisol is not always in line with perceived feelings of stress (Hunziker, et al., 2012), (Waller, et al., 2017).

Moreover, the intensity of physiological reactions differs per individual. For this reason, it is suggested to record a baseline measurement for each subject (Boucsein, 2012). The reliability of stress measurement can be improved by using both physiological and psychological (subjective) measures.

Finally, research results about the relationship between stress and team performance are contradictory. However, on an individual level, several researchers found a positive effect of stress on performance (LeBlanc, et al., 2008), (DeMaria, et al., 2010), (Pottier, et al., 2015). Stress management education may thus enhance technical performance during the simulation (Goldberg, et al., 2018).

To the best of my knowledge, the effect of the use of AR/MR systems in medical simulation, on stress and performance, must be still analysed.

# 3. Methods

Simulation in medical education is a high-demanding activity in terms of human capacity, attitude and behaviour, mental workload, and required performance. All of these aspects are influenced by students' previous experience and competencies, task complexity, use of new devices and technologies (such as high-fidelity manikins and applications in extended reality), as well as subjective response to stress.

The study of stress and mental load originated during and after low- and high-fidelity medical simulations is crucial to detect stressful events and eventual cognitive overloads which may damage students' learning and performance. This is important because inadequate learning and a high error rate may result in a greater possibility of risk for the patients, who would see their safety compromised.

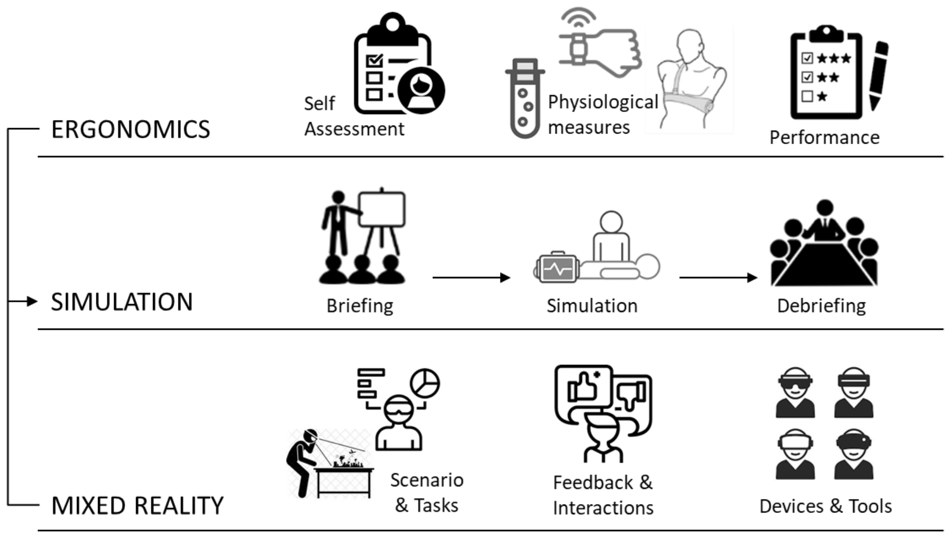


*Figure 15: Framework of the study for the assessment of medical simulations*

The main objective of this work is to assess the effectiveness of simulations in medical training in terms of performance, human emotional and cognitive states, and the effects that the use of AR applications may have in this context. This main goal results in a twofold activity, that can be summarised as follow:

- Design and development of a standard and structured methodology for the study of stress, anxiety, frustration, effort, cognitive load, and resultant performance of learners in the context of medical simulations;
- Design and development of a mixed reality simulator that allows enhancing the realism and immersion of the simulation, keeping the stress level and cognitive load close to those perceived in the real practice.

*Figure 15* shows the framework of the study. Medical simulations are composed of three main phases: briefing, simulated scenario, and debriefing. It is important to analyse not only the levels of stress and CL during the simulation but also how they vary before and after the simulation. In this work, a formal and rigorous protocol for the assessment of stress and cognitive load has been defined. It involves subjective (self-assessment questionnaires), objective (physiological signals acquired through wearable sensors and cortisol concentration through saliva sampling) measures, and performance assessment. Meanwhile, a mixed reality skill trainer has been designed and developed considering scenario features, tasks to be trained, kind of feedback and interactions to be supplied, and hardware and software to be used. The methodology for stress and cognitive load assessment has been applied to three different kinds of simulations: high-fidelity, low-fidelity, and mixed reality simulations.

## 3.1. Methodology for Cognitive Load and Stress Assessment

### *3.1.1. Data Acquisition Workflow*

As already explained, participation in immersive and engaging simulations (such as high-fidelity simulations or performing tasks in mixed reality environments) has a considerable psychological impact on learners. In particular, it can cause high stress (Geeraerts, et al., 2017) that, although rarely, can even be prolonged after the end of the simulation, especially in the most anxious subjects (Evain, et al., 2017). These aspects cannot be underestimated when planning the training activity. For this reason, an accurate evaluation of the cognitive and behavioural conditions of medical students is required.

Simulations consist of three different phases. The first one is the briefing in which the teacher explains the clinical case to be simulated, the tools and equipment available, and the procedures to be performed on the manikin. Subsequently, students simulate the clinical practice on the proposed scenario and, in the end, during the debriefing phase, the teacher discusses the performance with the students. He/she tries to understand the motivation behind their clinical choices and decision-making, explaining and solving the committed errors. In this phase, the teacher also asks them how they felt during and after the simulation, to manage and cope with eventual excessive stress.

As shown in *Figure 16*, the proposed methodology defines the formal protocol according to which the psychometric and biometric measures, for the analysis of stress and cognitive load, have to be collected.

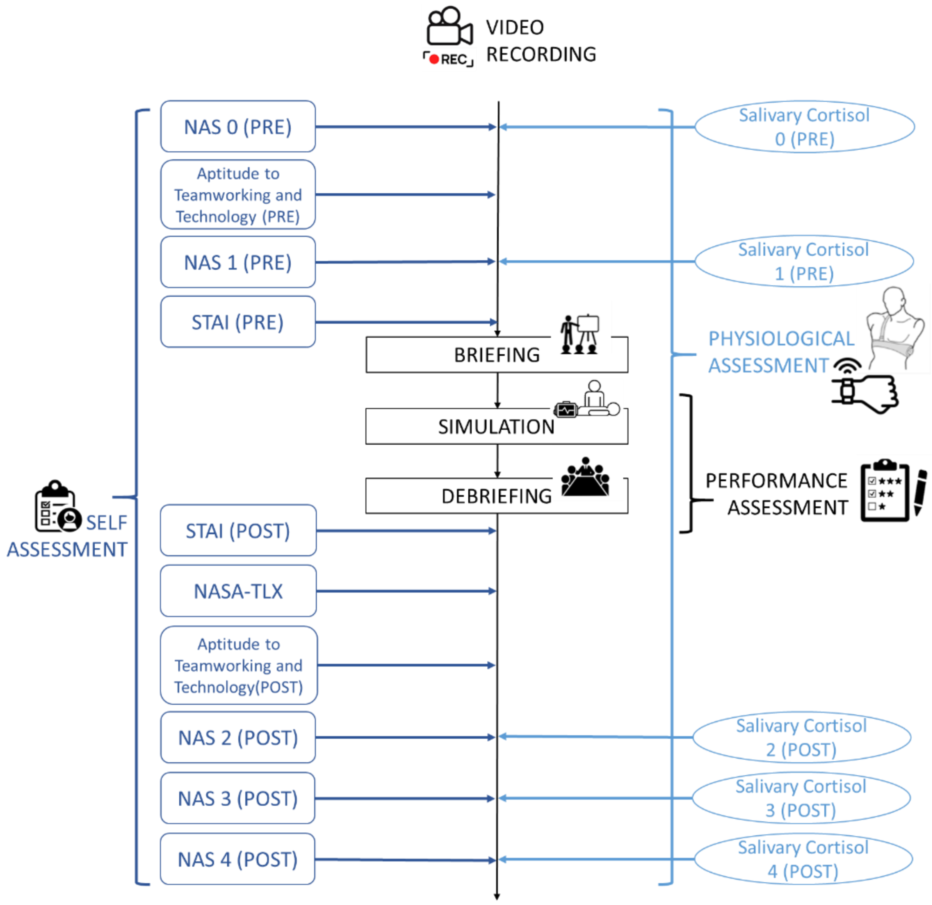


*Figure 16: Structured procedure for the overall assessment of stress and cognitive load in medical simulations*

Following the defined methodology, proposed in *Figure 16*, at the arrival of the learners in the classroom, a sample of salivary cortisol is collected with the compilation of the Numerical Analog Scale (NAS) to register the 'basal' level of stress. The same procedure is repeated after 10 minutes of rest and, in the meantime, a questionnaire about the aptitude for teamwork and use of technology is administered to the students. This questionnaire has been drafted with the healthcare professionals responsible for the simulations in Università Politecnica delle Marche, specifically for this study. Before the briefing, the standard questionnaire State Trait Anxiety Inventory (STAI) is distributed to the students to assess the current state of anxiety and the subjects' tendency to anxiety. In this way, it is possible to define the anxiety and stress level at rest, before undergoing stressful activities during the simulation (STAI PRE). Successively, the briefing starts, and then the simulation takes place. Then, the STAI questionnaire is administered again, to understand the differences in learners' perceived anxiety after performing the simulation (STAI POST). In this phase, also the NASA-TLX questionnaire (NASA Task Load IndeX) has to be fulfilled by the learners, to assess their perceived workload and, in particular, the perceived mental demand, effort, and frustration. The questionnaire on the aptitude for teamwork and use of technology is administered again, to evaluate any changes of opinion and perception after having experimented with the simulation and/or used the AR/MR training application. In the meantime, other three samples of salivary cortisol are taken, at 10, 20, and 30 minutes after the end of the simulation, always associated with the NAS scale. This makes it possible to study the variations in cortisol concentration trend due to the stressful activity.

Meanwhile, students wear smart devices for the collection of biometric indicators, from their arrival in the classroom to the end of the debriefing. Thus, even the physiological signals are recorded both during the periods of rest and during the stressful simulated scenario for clinical practice.

The three phases of briefing, simulation, and debriefing have to be recorded through a video camera. This is useful to track events in relation to time, consider peaks or variations of specific physiological signals in relation to events, and watch back the recording to analyse students' actions and errors.
Regarding the performance assessment, a proper checklist must be prepared, to discriminate between several scenario's phases and different tasks to be executed (considering correct/incorrect/not performed tasks). Based on the simulation type, also success, times, errors, consultations, and quality of performance could be evaluated for each task.
Moreover, specific questionnaires should also be used to assess students' skills, before and after the training session.
This methodology for the assessment of stress and cognitive load applies to any type of medical simulation, i.e. low fidelity simulations with skill trainers, high-fidelity simulations with advanced manikins for the training of team-working and decision-making, and training simulations combined with the use of devices for augmented and mixed reality applications.
In paragraph 3.1.2 methods and tools needed to accomplish this procedure are described in detail.

### *3.1.2. Tools and Methods*

As previously stated, the purposes of this study include the analysis of the simulation's activities and events that induce greater stress and/or mental load, and how simulation training affects the emotional and cognitive involvement of the learners. In literature, a potential dissociation between objective and subjective measures of mental workload has been highlighted

(Luque-Casado, et al., 2016). This founding has important implications in the analysis of cognitive states in applied settings. For this reason, the methodology illustrated in paragraph 3.1.1. comprehends three methods for the overall analysis of cognitive ergonomics during medical simulation: psychometric analysis (subjective, self-assessment method), biometric analysis (objective, physiological measurements), and performance analysis. While the psychometric and biometric analysis are thought to be structured in the same manner for every kind of simulation (i.e. high-fidelity simulations, low-fidelity simulations, and simulations in augmented or mixed reality), the performance analysis distinguishes between general elements to be evaluated in each simulation type, and particular items based on the specific simulation.

#### 3.1.2.1. Psychometric Analysis

The psychometric analysis encompasses the use of five self-assessment questionnaires: two of them properly developed for this work and three standard ones.

Survey about Aptitude to Teamworking and Use of Technology

The survey about aptitude for team-working and use of technology addresses team dynamics, personal feelings and attitudes, technology's opportunities and barriers, obtaining an overall evaluation of benefits and limits of simulation-based training, from the students' point of view. This survey is divided into two sessions: the first one is about the team simulation and the second one is about the role of technology. For each session, several questions have been prepared to allow the users to quantitatively evaluate different aspects of simulations.

The questionnaire development passed through different phases. First, during the conceptualization, the state of the art has been analysed to identify and select the indicators related to the predisposition toward the multidisciplinary team-working (Sigalet, et al., 2012) and the use of technology in simulation settings (Chi, et al., 2014). During the structuring phase, a team of experts, through the focus group method, drafted the survey, organising, and developing specific items. This draft was administered to a small group of subjects for the assessment of face and content validity, and reliability. The final, validated version consists of 35 statements divided into 2 sections for each session. Each statement is measurable on a 5-points Likert scale (with 1 corresponding to “strongly disagree”, 2 to “disagree”, 3 to “undecided/neutral”, 4 to ‘agree’ and 5 to “strongly agree”).
An introductive session is included to collect demographic and personal data such as sex, year of birth, degree course, previous experiences in clinical case management, group simulations, invasive procedures (such as blood sampling or catheter placing), and eventual participation in health volunteer activities.

First session: Team Simulation Training. In the survey, the simulation is explicitly defined as a process based on the reproduction, through models, of a system or environment in which the participants act to acquire or implement the necessary skills to face the simulated context. The questionnaire session relative to the group simulation training consists of two sections, described in detail in *Table 2*:

*Table 2: Items for the assessment of the aptitude to simulation and teamworking*

| | |
|---|---|
| The relevance of the simulation | Simulation is an effective teaching method |
| | I would not recommend simulation-based training to my colleagues |
| | Simulation succeeds in transmitting a greater motivation in the learning phase |
| | Simulation can help to deal with situations that cause anxiety or fear |
| | Simulation allows developing greater empathy with the patient |
| | Simulation is an inadequate tool for training group decision-making skills |
| | Simulation training methods help to understand the importance of team working |
| | Simulation does not provide me with clear information for understanding the actions to be performed during clinical practice |
| | Simulation confuses me in the identification of the significant theoretical elements, among those acquired by studying, for the practical resolution of the simulated case |
| | I can keep more attention during simulation training |
| | Simulation allows me to face a good discussion on the simulated case |
| | Simulation helps me to formulate a workable solution for the problem |
| | Simulation can help me to develop critical thinking (the process of forming a judgment through the analysis and objective evaluation of information) |
| The relevance of inter-professional education | Learning together with other professional figures is important for the improvement of collaboration in the workplace |
| | Simulation carried out together with other professional figures is an effective context for learning |

| | |
|---|---|
| | The opportunity to learn together with other professionals should be a priority in my training |
| | Learning shared with other professionals will improve my ability to understand clinical problems |
| | Interprofessional learning opportunities will not have a positive impact on the outcome of my patients |
| | Communication within the group is as important as technical skills |
| | It is not necessary for members of the team, providing immediate assistance to the patient, to announce their actions aloud |
| | Team members should paraphrase or repeat the instructions received to clarify what they understood |
| | Safety in the delivery of care increases if all the members of the team share information regarding patient management |
| | Frequent summaries of patient test results are useful for keeping team members' attention to patient needs |
| | Within the team, establishing and knowing "who-does-what" is essential for improving the quality of the care provided |
| | Team members must ask for assistance if they need help completing a task |

Second session: Role of Technology. In the survey, technology is defined as the application of IT and telematic devices. The questionnaire session relative to the role of technology in the simulation-based training is composed of two sections, described in detail in *Table 3*:

*Table 3: Items for the assessment of the aptitude to technology usage*

| | |
|---|---|
| The personal attitude to the use of technological devices | I am familiar with the use of the following devices:<br>• Personal Computer<br>• Tablet<br>• Smartphone<br>• Virtual Reality glasses |

| | |
|---|---|
| | • Gloves with haptic feedback for gaming |
| | I am familiar with simulation videogames (e.g. "The Sims") and serious games (i.e. videogames with educational purposes) |
| | I use technological products that support health and lifestyle (e.g. smartwatch, activity tracker, etc.) |
| | I feel ready to work in a high-tech environment |
| | I find it stressful to work in a high-tech environment |
| The relevance of technology in the simulation context | Technological devices (i.e. virtual reality glasses, gloves with haptic feedback, …) are a valuable tool for learning during training |
| | Multisensory interaction (tactile, visual, and auditory) through technological devices such as gloves with sensors, glasses, and earphones, encourages learning in the simulation |
| | Providing feedback through technological devices promotes learning in the simulation |
| | A high degree of immersion during the simulation has a positive effect on:<br>• Learning<br>• Psychological component |
| | In the context of learning during simulation, I would like to be supported by high-tech devices for:<br>• Team working<br>• Decision process<br>• Practice with the physical simulator<br>• Understanding of human anatomy, and physiological and pathological processes |

Each statement has to be evaluated on a 5-points Likert scale, before and after having taken part in the simulation, in order to register any opinion change after having experienced team simulations and used high-tech devices.

Numerical Analogue Scale (NAS)

A Visual Analogue Scale (VAS) is an instrument used to measure an attitude or mind state that is believed to range across a continuum of values and cannot easily be directly measured. It is often used in epidemiologic and clinical research to measure the intensity of various symptoms or in clinical practice by occupational physicians for the assessment of workers' perceived stress. The Numerical Analog Scale (NAS) is the numbered version of the VAS, in which a bar or line is divided into 10 intervals, numbered from 0 to 10. The subject is asked to select the whole number (0–10 integers) that best reflects the intensity of his/her stress (0 = no stress, 10 = very strong stress). The VAS, and consequently the NAS, have proved to be valid, effective, and easy-to-implement tools for the rapid assessment of perceived stress (Lesage, et al., 2012), (Mitchell, et al., 2008). In the proposed procedure, the NAS rating method has been selected against the VAS method, in order to have integer values to analyse and to correlate with cortisol concentration and other parameters (from other questionnaires, physiological measurements, and performance ratings). The NAS scale has to be answered five times, at the arrival in the classroom and 10 minutes later for the measurement of "basal" perceived stress, and 10, 20, 30 minutes after the end of the simulation for the measurement of the simulation's influence on perceived stress. It has to be always administered at the same time as the salivary cortisol sampling, to correlate subjective and objective stress variations.

State Trait Anxiety Inventory (STAI)

The State Trait Anxiety Inventory (STAI) questionnaire is often used in research studies for the assessment of anxiety. It consists of two modules, each one composed of 20 questions:

- The STAI Y-1 form (Appendix A.1) allows for the assessment of the subject's current state of anxiety, considering feelings of apprehension, tension, worry, and nervousness. By submitting it before and after a particular activity, results in a sensitive indicator of transient changes in the level of anxiety.
- The STAI Y-2 form (Appendix A.2) allows for the evaluation of the anxious trait, i.e. the individual predisposition to anxiety. It gives an index of how a subject generally feels, and it can be used to identify people predisposed to develop anxiety in stressful situations. This form has to be submitted only once, before the activity begins.

For each statement, the subject must answer on a 4-points Likert scale. The values chosen for each item are added together, thus obtaining the total score of the form. Therefore, the total score of each module can range from a minimum of 20 to a maximum of 80 points. In literature, cut-offs of 37/38 points for males and 39/40 points for females have been identified for each module. Beyond these cut-offs, the anxiety level is considered significant (Julian, 2011). From a study about anxiety assessment in working adults subdivided into three age groups, mean values of 36.54 (male, Y1), 36.17 (female, Y1), 35.55 (male, Y2), and 36.15 (female, Y2) have been found in subjects with an age ranging from 19 to 39 years old. Reducing the investigation to college students (the ones considered in this work), the STAI mean total scores increase to 36.47 (male, Y1), 38.76 (female, Y1), 38.30 (male, Y2), and 40.40 (female, Y2) (Spielberger, et al., 1983).

The STAI questionnaire has been inserted in the methodological procedure for the assessment of cognitive and emotional conditions, to record, on one side, the tendency to the anxiety of the subjects (as a demographic characteristic), and, on the other side, the effect that the simulation has on their stress and anxiety perception. Indeed, while the STAI Y-2 form must be submitted only before the simulation, the STAI Y-1 form must be administered twice (before and after the simulation).

NASA Task Load Index (NASA-TLX)

The NASA-Task Load Index (NASA-TLX) is a subjective scale for the assessment of workload. The analysis of the workload can be quantitative, (i.e. the number of activities to be done), or qualitative (i.e. the difficulty and complexity of the task to be carried out).

The NASA-TLX was developed to minimize the assessment variability among subjects (Hart, et al., 1988). It consists of a questionnaire composed of six questions to be answered via bipolar scales. The subject is asked to self-evaluate his/her performance by considering six different items:

- Mental Demand: How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

- Physical Demand: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
- Temporal Demand: How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
- Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?
- Performance: How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
- Frustration: How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

NASA-TLX has to be filled in by participants at the end of each experimental session. It consists of two main parts, weights and ratings, that are combined to calculate a weighted average, giving as result an overall workload score, on a 100-points scale. Thus, the questionnaire is divided into two main parts:

- Weights (Sources of Load): the first requirement is that each subject assesses the contribution (weight) of each factor, concerning the workload, for the execution of a specific task. The weights are determined by the choice of the factors (subscale) that the subject believes most relevant for the workload, among a couple of choices. Therefore, the possible pairwise comparisons of the six subscales are 15. Each pair is presented to the subject on a card (Appendix B.1). Subjects choose the member of each couple that contributed the most to the activity's workload. The number of times each factor is selected is noted. Annotations can range from 0 (irrelevant) to 5 (more important than any other factor).

- Ratings (Magnitude of Load): the second requirement is to obtain numerical evaluations for each subscale, which reflect the magnitude of that factor in the task performed. The six subscales are shown on a card (Appendix B.2). The subjects answer by marking each scale in the desired position. Each scale is presented as a line divided into 20 equal intervals, anchored by bipolar descriptors (e.g. High / Low). The 21 division signs for each scale divide it from 0 to 100, in steps of 5.

The overall workload score, for each subject, is obtained by multiplying each evaluation by the weight attributed to that factor; the sum of the weighted ratings for each activity is then divided by 15.

This questionnaire is included in the proposed methodology because it allows for the assessment of not only the perceived cognitive load (mental demand) needed to perform the simulation, but also emotional states related to stress such as perceived effort and frustration. For obvious reasons, it has to be administered after the simulation.

#### 3.1.2.2. Biometric Analysis

Changes associated with different levels of stress and mental load have been reported mainly in the cardiovascular and respiratory systems, in the brain’s electrical activity, and gaze entropy, velocity, and pupil dilation.

However, for the assessment of cerebral and ocular activities, EEG and eye-tracking systems are needed, and this may limit the analysis and interpretation of such physiological parameters in the real-field. Indeed, these instruments are usually exploited in lab-based researches, since their usage outside the laboratory may be invasive and compromise the correct completion of the tasks. Moreover, on one side, the use of EEG in situations that require a lot of movements may result in a very noisy signal, questioning its significance; on the other side, the simultaneous use of eye-tracking systems and HMD for augmented reality may be cumbersome (unless the HMD includes itself the eye-tracker).

Therefore, to include in the methodology a continuous and non-invasive acquisition of biometric parameters, for the objective assessment of stress and cognitive load, physiological signals acquirable through the easy-to-use wrist and chest bands have been selected. Moreover, the collection of salivary cortisol samples through "salivette" has been chosen as a reference gold standard. While the physiological parameters are recorded during the overall duration of the lesson (from the briefing to the end of the debriefing), the salivary cortisol is collected five times (twice before the briefing and three times during the debriefing) in order to analyse its trend. The selected biometric measurements are described hereunder.

Heart Rate (HR)

Heart rate (HR) is the number of heartbeats per unit of time, typically expressed as beats per minute (BPM), and it is one of the physiological parameters used for the assessment of subjects' cognitive states. Some authors are critical about the use of this parameter for the analysis of mental load because it is easily influenced by external factors such as physical activity, environment, and psychological elements (such as emotional involvement) that are not related to the analysed activity. However, in literature, it has been proved that heart rate variations are directly related to the mental load, i.e. HR increases as CL increases (De Waard, 1996), (Miller, et al., 2001).

<u>Heart Rate Variability (HRV)</u>

Another measurement related to electrical heart activity is heart rate variability (HRV). HRV represents the variability of the Inter-Beat Interval (IBI), i.e. the time interval (in ms) between two successive heartbeats (Miller, et al., 2001). HRV is under the control of the Autonomous Nervous System (ANS), thus it is considered a reliable estimator of ANS statuses, also in real-life settings. It is demonstrated that the ANS, which controls our capability to react to external stimuli through the parasympathetic and the sympathetic branches, is influenced by mental stress (Castaldo, et al., 2015). For this reason, HRV is considered a reliable indirect means to monitor cognitive states.

HRV fluctuations can be analysed using time domain, frequency domain, and non-linear domain methods. Four measures in the time domain (RR, SDRR, RMSSD, and pNN50) and one measure in the non-linear domain (D2) result significantly reduced during stressful events. The ratio LF/HF in frequency domain results instead significantly increased, suggesting a sympathetic activation and a parasympathetic withdrawal during acute stress

(Castaldo, et al., 2015). However, this result could change if HRV measurement was taken after, and not during, the stressful event, and it could vary consistently with the phases of the stress (Castaldo, et al., 2015). For this reason, it is essential to monitor the physiological parameters for the entire duration of the simulation (rest phases included).

Moreover, it has been shown that the extent of inter-beat variability decreases with increasing mental demand (Luque-Casado, et al., 2016).

As for the HR, also HRV is sensitive to physical activity and strong emotional reactions and this must be taken into account during the simulation and signal analysis, tracking and correlating the events that occur during the acquisition.

Breathing Rate (BR)

As pointed out for heart activity, also respiration rate and respiration depth are not selectively optimal indices for measuring mental workload, since they are sensitive to physical activity, strong emotional reactions, and speech. However, the breathing rate and depth are used for the measurement of the mental workload because changes in BR reflect variations in the mental effort (De Waard, 1996).

The amount of oxygen that the body needs is determined by the level of activity in the various body tissues. When the task requires an increase in this activity, the brain activation rises to adequately manage these requests. This causes an increase in respiratory rate and a decrease in breathing depth (Roscoe, 1992).

Galvanic Skin Response (GSR) or Electrodermal Activity (EDA)

Galvanic Skin Response (GSR), also called Electrodermal Activity (EDA), is the measure of continuous changes in the electrical characteristics of the skin, such as skin conductance, as a result of the change in the human body sweating. The traditional theory regarding the GSR analysis is based on the hypothesis that skin resistance varies according to the state of the skin sweat glands. Human body sweating is regulated by the Autonomous Nervous System. In particular, if the ANS sympathetic branch is highly activated, the activity of the sweat glands accordingly increases, and so does the skin conductance. For this reason, skin conductance can be used as an index of the responses of the human Sympathetic Nervous System, which is directly involved in the regulation of emotional arousal.

GSR is therefore measured by recording the variation of a low voltage current applied between two electrodes, placed on the skin. As the sweat glands become more active (e.g. in the presence of physical or mental inputs), the skin begins to conduct a greater amount of electricity: at that time a change in GSR is registered (in μS). The GSR signal is divided into two components:

- Tonic component: responsible for the slow changes in the GSR signal. It is considered the background level of activity on top of which rapid GSR responses appear. It is measured as “skin conductance level” (SCL) and it can be related to the general level of emotional state and level of stress.
- Phasic component: responsible for relatively rapid changes in the GSR signal, in the form of rapid fluctuations or peaks. It is also called “skin conductance response” (SCR) and it can be generated by specific events.

Some studies have highlighted the relationship between the EDA signal and mental states such as stress, anxiety, fatigue, emotional involvement, mental load, and level of the excitement of the perceived emotion. However, like the previous methods, the GSR does not allow to measure the mental workload selectively. Indeed, it is influenced by temperature, age, time of the day, and physical effort.

Salivary cortisol

The alteration of salivary cortisol is considered an objective indicator of individuals' stress levels. Saliva can be easily sampled using an absorbent oral swab to be placed in a test tube. To characterise the variation of salivary cortisol concentration due to the stressful event, at least 5 samples are required (Clements, 2013): baseline at the arrival of the subject, 10 minutes later (so that the subject has found a condition of tranquillity), 10, 20, and 30 minutes after the end of the test, to observe the cortisol peak determined by the stressful event, and the subsequent decrement.

However, even this type of parameter is influenced by several factors that must be taken into account as they can cause a bias in the data analysis. For example, salivary cortisol in women varies according to the phases of the menstrual cycle. Other important factors that influence the cortisol level are the presence of stressogenic events in the last 6 months (e.g. severe bereavement, abortions, separation/divorce, transfers/relocations), the presence of chronic pathologies or celiac disease, a very high body mass index (BMI> 30), the intake of certain drugs (such as cortisone, beta-blockers, diuretics, glucocorticoids, interferon, statins or hormonal therapies with estrogen and progestogens), having a vegan diet (Clements, 2013). Subjects with these characteristics should be excluded from the cortisol measurement.

#### 3.1.2.3. Performance Analysis

In the performance analysis, general elements to be evaluated in every simulation type can be distinguished from particular items to be assessed only in specific applications. In general, simulations can be divided into different phases, and for each phase, some specific tasks can be evaluated. As regards team-based, high-fidelity simulations, discrimination of events by phases can be carried out. The same assessment per phase can be performed also in low-fidelity, practical simulations. In both cases, for each phase, several tasks are expected and required. A checklist should be used to assess the tasks as "correctly performed", "incorrectly performed", or "not performed".

Then, in addition to the assessment of task correctness, other parameters, based on the specific simulation (and especially for low-fidelity simulations), can be evaluated. Some examples are:

- Times: time for the preparation, time to achieve success, time for each subtask (when relevant), time to solve the case;
- Errors: type, quantity, severity;
- Consultations: number and type of aids given by the teacher and requested by the student;
- Other meaningful parameters that are specific for each simulation (e.g. number of attempts to succeed).

A questionnaire for the assessment of skills acquisition should be administered before and after each simulation.

Students' performance should always be recorded through video cameras to track the events, identify critical issues, and assess skills after the simulation. Watching the recorded video can help teachers and students to detect errors and bad clinical choices that can be reviewed, discussed, and improved. In this way, the skills for clinical case management are enhanced and eventual excessive stress due to the outcome of the simulation can be coped with.

Moreover, the tracking of the events is also useful for the study and discrimination of intrinsic and extraneous cognitive loads, because it allows distinguishing the distracting elements that should be avoided and removed.

### *3.1.3. Data Analysis*

The outcomes of the proposed methodology for the assessment of stress, cognitive conditions, and performance of students in medical simulation training, can be analysed and interpreted as follow:

- Subjective measures: NAS, STAI, and NASA-TLX describe the perceived stress, anxiety, cognitive load, frustration, and effort. The questionnaire concerning the aptitude for team-working and use of technology shows the predisposition toward simulation-based medical education and the use of new devices as HMD for augmented reality. By administering the self-assessment surveys before and after the simulation, it is possible to understand how the simulation influence the students' cognitive conditions and opinions, and how they perceive this variation.

- Objective measures: the non-invasive and continuous monitoring of physiological parameters allows us to analyse how the human body responds to mentally demanding and stressful activities, discriminating among different tasks for complexity and difficulty. The events tracking also allows distinguishing between intrinsic and extraneous cognitive load. The collection of 5 salivary cortisol samples permits the detection of the variation of the stress level due to the simulation. Moreover, the computation of the fold increase and the area under the curve (AUC) allows simplifying the statistical analysis, without forgoing the information contained in multiple cortisol measurements (Pruessner, et al., 2003).
- Performance measure: the assessment of the single task, together with the analysis of errors and times, allows relating performance to acute events of stress and cognitive load. The performance measure, together with the students' skills evaluation pre- and post-simulation, is useful for the assessment of the training session effectiveness.

This overall and comprehensive analysis gives the opportunity to discern between the perceived, subjective cognitive state and the physiological, objective variation in stress and cognitive load. This kind of analysis is also extremely useful in comparative studies to assess the differences between traditional simulations and those performed with AR/MR applications.

Standard central trend measures (such as mean and standard deviation) can be calculated for the description of the demographic characteristics, the performance, the biometric indices, as well as for the analysis of the responses to the self-assessment questionnaires.

All the collected variables can be statistically analysed to understand the variables that affect students' performances, stress, and cognitive load, during simulations. Several models of single and multiple linear regression can be accomplished to discover which factors are significantly related to each other, the ones useful for the study of performance, stress, and cognitive load, and the ones that could be ignored. Indeed, in statistical modeling, linear regression analysis is used to estimate the relationships between two or more variables:

- Dependent variable: the main factor to understand and predict;
- Independent variable(s): factor(s) which is(are) supposed to influence the dependent variable.

The case of one independent variable is called simple linear regression, while for more than one independent variable, the process is called multiple linear regression. Single and multiple linear regression analysis can be used in this context to find out variations in the dependent variable (e.g. performance, stress, CL) that can be attributed to variations in the independent variables. It allows also quantifying the strength of the relationship between the dependent and independent variables, or to understand if some independent variables had no linear relationship with the dependent variable at all. For example, in this context, the linear regression analysis consents to understand which are the factors that influence students' performance, stress, and cognitive load. Based on these relationships among variables, it is possible to define specific guidelines for the optimisation and re-design of the simulations, to improve learners' performance e balance the stress and CL levels.

## 3.2. Design and Development of a Mixed Reality Rachicentesis Simulator

The second activity of this work concerns the design and development of a mixed reality simulation system, which integrates augmented reality with a specific skill trainer for lumbar puncture.

Firstly, a framework for the design of any kind of advanced clinical simulation has been defined.

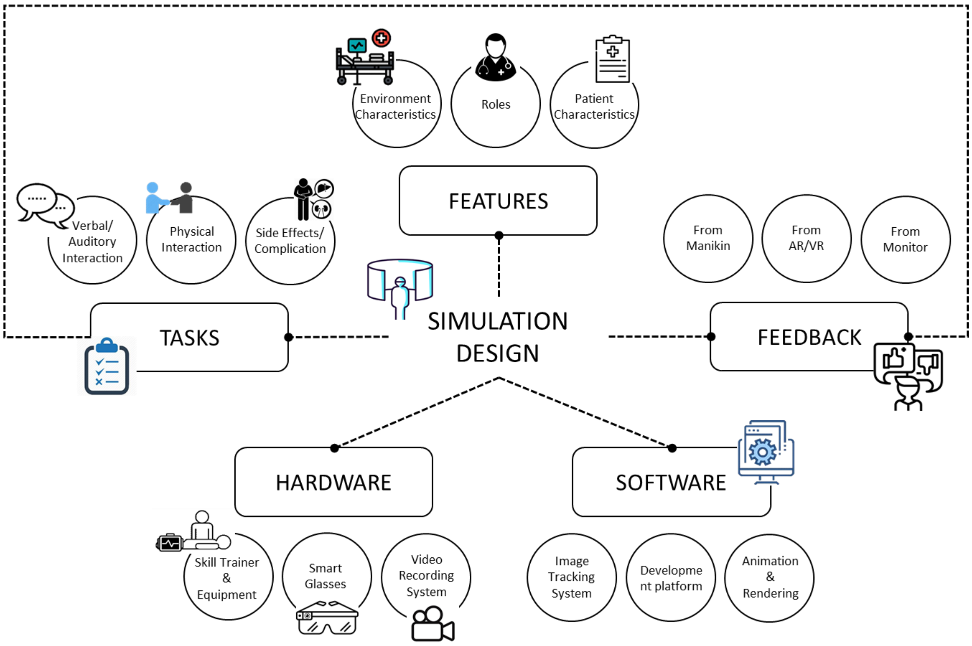


*Figure 17: Framework for the design of medical simulations*

The proposed architecture is suitable for the design of both transversal and specific practical simulations (e.g. emergency management, first aid, or catheter insertion, pericardiocentesis, blood sampling, etc.), in high-fidelity or low-fidelity, with or without the AR integration. Indeed, the framework is composed of different modules and sub-modules that have to be individually considered and implemented based on the simulation type and purpose. As shown in *Figure 17*, the architecture supports:

- Features Definition: the simulation features must be defined in terms of room layout, clinical actors, patient anamnesis (characterization of clinical scenarios for the diagnosis and management of different conditions);
- Tasks Definition: characterisation of each task to be performed by the students, considering verbal and physical interaction with the patient, potential side effects and complications, clinical risks associated with different treatment options, and patient management;
- Feedbacks Definition: characterisation of feedback to be given by physical and/or virtual systems;
- Hardware Selection: it consists in the selection of manikin and equipment suitable for the simulation purpose, the choice of the potential AR device and video-recording system;
- Software Selection: it includes the selection of AR development platform, tracking system, animation, and rendering software.

Simulation features, tasks, and feedbacks are described in *Table 4*:

*Table 4: Description of simulation features, tasks, and feedbacks*

| **Aspect** | **Description** |
|---|---|
| Environment Characteristics | Detailed definition of the simulation setting in terms of room characteristics and equipment layout. To be specified real and virtual components |

| | |
|---|---|
| Roles | Detailed definition of students' role (e.g. actors as nurse, doctor, patient relative, etc.) both for multidisciplinary team simulations and individual simulations |
| Patient Characteristics | Detailed characterization of patient specifying relevant demographic information (e.g. gender, age), physical characteristics (e.g. obesity), anamnesis, and signs/symptoms |
| Tasks | Identification of the sequence of actions to be performed by the student to contextualize the following items |
| Verbal Interaction | Specification for each task if a verbal interaction between students and patient is expected. In that case, possible patient answers should be planned. Verbal interaction is important to consider empathy skills and interprofessional collaboration |
| Physical Interaction | Specification for each task which physical interaction between students and patient is expected (e.g. palpation, needle insertion) |
| Side Effects | Identification of possible side effects (predictable events) that could occur during each task |
| Complications | Identification of possible complications (unpredictable events) that could occur during each task |
| Manikin Feedback | Identification of feedback to be provided by the manikin (e.g. fluid leak from a paracentesis puncture site) |
| AR/VR Feedback | Identification of feedback to be provided through virtual or augmented reality (e.g. patient movement, sounds) |
| Quantitative Feedback | Identification of quantitative feedback (e.g. heart rate, $SPO_2$) to be provided through a monitor connected to the manikin |

### *3.2.1. MR Design: Features, Tasks, Feedback, Hardware, and Software*

In the aviation context, simulators are used in flight training to provide auxiliary information and feedback (such as proper speed, heading, approach angle for final landing) that are not provided in the actual flight. Studies suggested that pilots trained with this additional information learn to fly more quickly but they become dependent on it and, in the real flight deck, they could not perform as well as those trained without such guidance (Seagull, 2012). This could be a reason why it is not convenient to add in AR the patient's interior organs since it would be information that students would not have in the real practice. Therefore, the decision to reconstruct in AR the rest of the patient body to be superimposed on the skill trainer, to make the simulation more realistic, has been taken. In this way, the student can receive some feedback from the simulated patient, as it would happen in reality. The purpose of the MR application is not to facilitate the operation, but to “immerse” the student in a more realistic environment, which puts him/her under pressure during the exercise. Indeed, by simulating the stimuli due to the patient’s reactions during the actual operation, the students should feel more involved and have the impression of being immersed in a real context (and not in a simulation).

For this study, a mixed reality application has been developed for the training of the rachicentesis procedure. It has been designed to carry out the following main functions, aiming at affecting student's cognitive and emotional reactions:

- To overlay the 3D model of a patient on the real manikin, positioned as during the actual procedure. Indeed, the skill trainer consists only in the patient's abdomen, while the chest, head, arms, and legs are not represented.

- To track the needle position during the operation, to provide visual and sound feedback based on the action performed. In particular, when the student inserts the needle into the manikin, the AR patient complains and has a spasm.

Simulation features, tasks, and needed feedback have been defined with the teachers of the Faculty of Medicine, of Università Politecnica delle Marche, responsible for the rachicentesis training.

Concerning the environmental characteristics, the simulation takes place in a classroom equipped with an abdomen skill-trainer lying on a desk. The manikin can be placed in left lateral decubitus or sitting positions. On the right of the manikin, on the same desk, all the equipment and instruments useful for the procedure (such as needle, latex gloves, sterile cloth, disinfection swab, test tube) are provided. The student is supposed to individually act as an expert anaesthetist, and the patient could have different physical characteristics (e.g. normal BMI, obesity, pregnancy) that may affect the outcome of the procedure. Indeed, if the patient is sitting, and pregnant or obese, the higher pressure can modify the spilling out of the cerebrospinal fluid (CSF or liquor). While the 3D virtual patient allows representing every kind of physical characteristic (obesity, pregnancy, etc.), the pressure inside the spinal canal can be directly modified on the skill trainer according to the patient's physical condition, thus assuring a more realistic simulation. The patient anamnesis is not relevant for this kind of simulation.

Afterward, rachicentesis simulation tasks have been defined. For each task, the verbal and physical interactions between student and skill trainer /

virtual patient, and possible side effects and complications have been specified. Also, the visual, and auditory feedback, useful to improve simulation realism, have been defined for each task. A summary is shown in *Table 5*, and the list of tasks is hereunder presented:

- Task 1: reception and technical explanation of the procedure to the patient
- Task 2: positioning of the patient (left lateral decubitus or sitting)
- Task 3: palpation of the anatomical landmarks on the skill trainer
- Task 4: disinfection
- Task 5: threading the needle
- Task 6: extracting the needle stylet
- Task 7: waiting for the liquor spilling out
- Task 8: taking the test tube and collect the liquor
- Task 9: putting the stylet into the needle
- Task 10: removing the needle
- Task 11: tamponing the skin

*Table 5: Interactions, side effects, and feedbacks expected for each task of rachicentesis simulation*

| | **Verbal Interaction (1)** | **Physical Interaction (2)** | **Side Effects or Complications** | **Manikin Feedback (3)** | **AR Feedback (4)** | **Monitor Feedback** |
|---|---|---|---|---|---|---|
| **Task 1** | No | No | No | No | No | No |
| **Task 2** | No | Yes | No | No | Yes | No |
| **Task 3** | No | Yes | No | Yes | No | No |
| **Task 4** | No | Yes | No | No | No | No |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Task 5** | Patient complains | Yes | Rare blood spilling out | Yes | Yes | No |
| **Task 6** | No | Yes | No | No | No | No |
| **Task 7** | No | Yes | No | No | No | No |
| **Task 8** | No | No | No | No | No | No |
| **Task 9** | No | Yes | No | No | No | No |
| **Task 10** | No | Yes | No | Yes | No | No |
| **Task 11** | No | Yes | No | No | No | No |

(1) If the patient speaks or not
(2) With the patient
(3) Haptic feedback
(4) Visual and auditory feedback

Furthermore, hardware and software have been selected for the development of the mixed reality rachicentesis application.

The manikin used for the mixed reality system was the Gaumard® Lumbar Puncture Trainer. Its anatomic features include iliac crests, lumbar vertebrae L2-L5, ligamentum flavum, epidural space, the skin layer, subcutaneous layer, and connective tissue. It can be placed in the left lateral decubitus or sitting positions. It provides realistic tactile feedback, and lifelike needle resistance combined with a fluid pressure system, that allows the liquor spilling out and collection (*Figure 18 (a)*). Thanks to these characteristics, it can be used for the simulation of the needle insertion between vertebrae, injection of local anaesthesia, lumbar puncture, epidural, and rachicentesis.

For the augmented reality application, after a wide benchmarking, two devices have been selected:

- Microsoft Hololens® (*Figure 18 (b)*)
- Headset for smartphone Vox Gear Plus® (*Figure 18 (c)*)

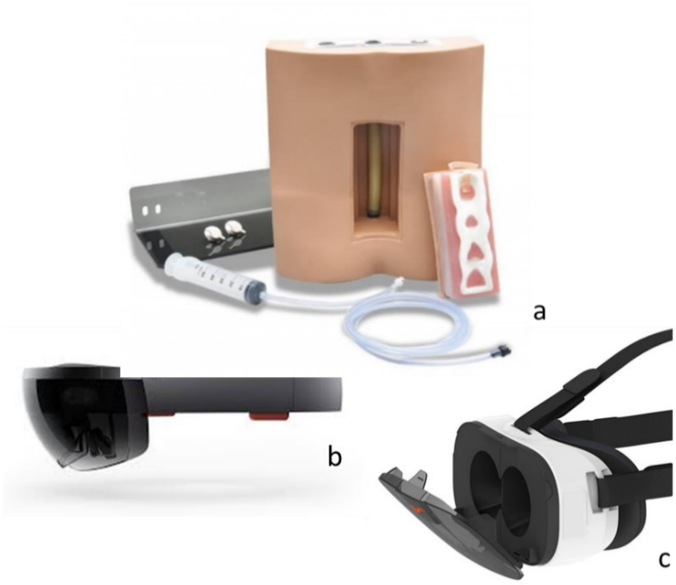


*Figure 18: Mixed Reality System Hardware: (a) Gaumard® Lumbar Puncture Trainer, (b) Microsoft Hololens®, (c) Vox Gear Plus®*

Even if both the devices allow using the augmented reality application, their characteristics place them at the antipodes in the current technological panorama (and for this reason they have been selected).

Vox Gear Plus® is a headset for the smartphone. It does not have an own computing power but can only be used in combination with a smartphone, on which the application is launched. For this reason, it has an extremely low cost (tens of euro). Therefore, Vox Gear Plus, holding the smartphone in front of the user's eyes, gives the feeling of visual three-dimensionality through the use of appropriate lenses. This type of device is generally used for virtual reality applications; however, it is possible to use it for augmented reality applications, using the camera of the smartphone in a "see-through" manner. Indeed, this allows the user to see the reality in front of him/her adding the virtual elements on the screen.

Microsoft Hololens® is a technologically very advanced device. Indeed, with Hololens the user does not look at a screen, but he/she can see the environment populated with holograms. Hololens is transparent glasses, able to generate small holograms in front of the eyes of the user, who thus has the sensation of seeing a life-size hologram in front of him/her. Concerning the sensors onboard, in addition to the classic accelerometers and magnetometers, Hololens has a special camera (derived from Microsoft Kinect®) able to reconstruct the 3D structure of the surrounding environment. However, the field of view for the holograms' visualisation is of only 35°. Hololens is also significantly more expensive than other AR devices, with a cost starting from 3000€ for the "Development Edition".

Regarding the software, the following ones have been used for the development of the AR application:

- Vuforia
- Blender
- Unity

Vuforia is a software development kit (SDK) for augmented reality. It uses computer vision technology to recognise and track the position of target

images and simple 3D objects in the space. These features allow developers to position and orient virtual objects, such as 3D models, in the real space, visualised through a camera or other device. The position and orientation of the 3D model are traced in real-time, and the user has the impression that the virtual object is part of the real scene.
In the developed application, Vuforia was used to track the position of two target images:

- Patient Target Image: target related to the patient, fixed in space, needed to visualise the virtual body of the patient overlying the manikin;
- Doctor Target Image: target attached to the learner's hand, and therefore mobile, needed to visualise the real needle (further details in the following paragraphs).

Blender is a 3D computer graphics software used to create animated films, visual effects, 3D models. It was mainly used for the creation of the 3D virtual model of the patient, and the related animations. Indeed, it allows using statistical body shape modelling, that permits to apply the same motion to a wide range of different body shapes (Scataglini, et al., 2019) (thus simulating several patients with different physical characteristics).

Unity is a multi-platform graphics engine, which can be used to create multimedia content, games, simulations, both in 3D and 2D. Its graphic interface facilitates the programming, but it is also possible to write scripts to insert automatisms in the created application. Unity is the principal software for the creation of augmented reality applications. Indeed, it allows using the

Vuforia SDK and then associating the 3D model, created in Blender, to each target. Therefore, it is used to create the AR scene.

### *3.2.2. Development of the MR Rachicentesis Prototype*

#### 3.2.2.1. Target images definition

The target images were created using an online free tool that generates images rich in features, that Vuforia can easily recognise. The symbols of a patient and a doctor were superimposed on the target images, to facilitate their recognition to the user (*Figure 19*).

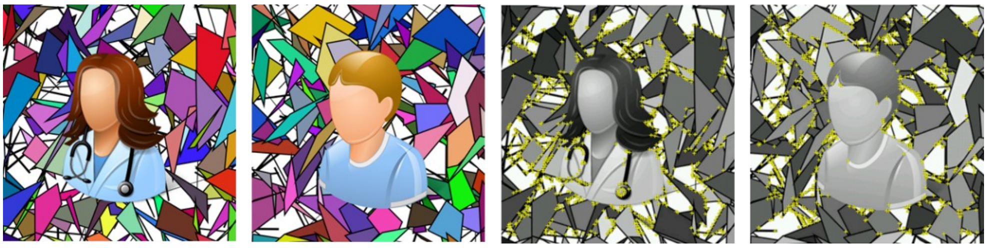

*Figure 19: Target images (left) and features (in yellow) recognisable by Vuforia (right)*

#### 3.2.2.2. Vuforia library generation

The images were then uploaded in Vuforia, to generate the libraries to be imported into Unity for tracking. During the upload of the target images, it is necessary to indicate the real dimensions of the physical targets to be tracked; this information is particularly important for the use of the AR application with Hololens. Indeed, this device is equipped with sensors able to estimate the size of the objects it detects in its field of view. If the physical targets have different dimensions from the indicated ones, they will not be recognised.

A .unitypackage file and a string for the license were then downloaded and imported into Unity, where the targets can be associated with the 3D models.

#### 3.2.2.3. 3D models development

For the development of the AR patient, a 3D model of a woman and the texture were respectively downloaded in .obj and .mtl file formats, from an anthropometric database (as suggested in (Paul, et al., 2019)). These models are generally supplied in the upright, open arms position (*Figure 20*), so that the user can modify the posture and create the desired animations, depending on his/her needs.

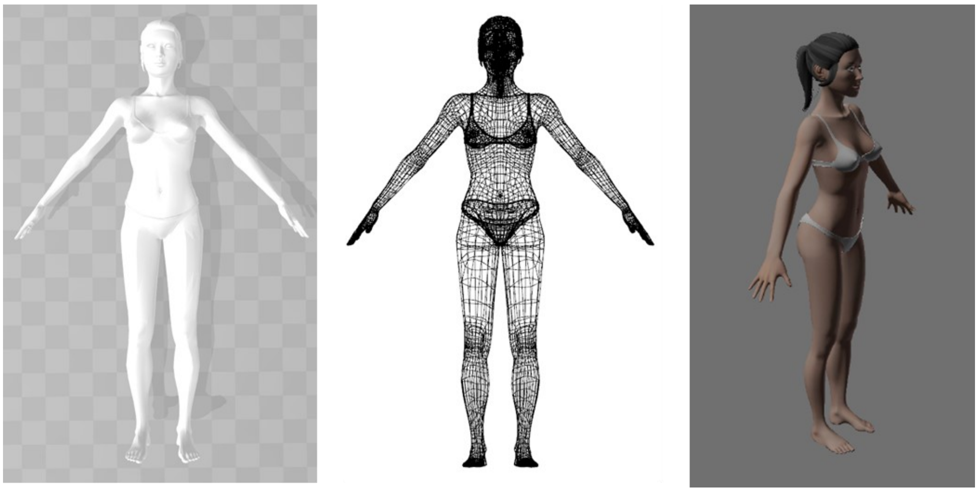

*Figure 20: From left to right: solid mesh, wireframe, and render of the patient 3D model used for the AR application*

Based on the rachicentesis procedure and the manikin's configurations, the 3D model of the patient was then repositioned in the left lateral decubitus and sitting positions. This repositioning was performed in

Blender, through a procedure called "rigging". During this procedure, a skeleton of rigid elements, connected by nodes, is generated. By deforming the skeleton, it is possible to modify the position of the 3D model (*Figure 21 left*).

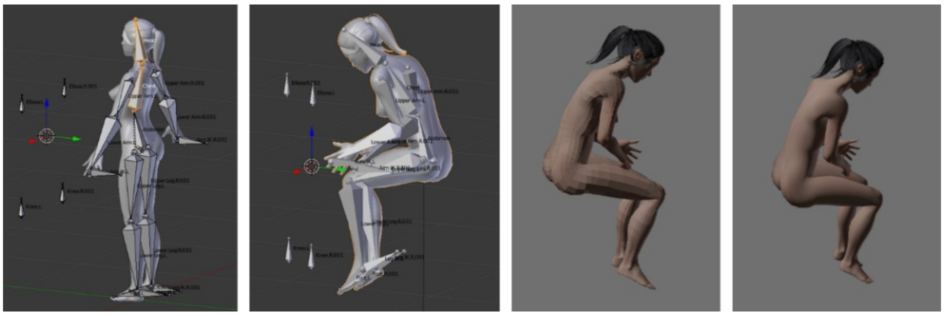

*Figure 21: From left to right: skeleton creation, model repositioning, the mesh of the model, smoothed model (after applying the filter)*

In the texturing phase, a "smooth" filter must be used on the 3D model to round the edges of the mesh (*Figure 21 right*). Indeed, this filter allows to change the normal of each surface as it gets closer to the edge; in this way, the 3D model surface is smoothed, and the user does not have the impression of watching a polygonal model. Then, Blender was used also to create model animation. Thanks to dedicated commands, it is possible to define the times and movements of the skeleton. In this case, the model was animated to simulate the spasm due to needle insertion. The process output is a .fbx file, which can be imported and in Unity.

3.2.2.4. 3D scanning of the Gaumard® Lumbar Puncture Trainer

While using the AR application, the virtual model must accurately overlap the real manikin, without appearing either too large or too small.

Indeed, in the first case, the virtual model would appear disproportionate, and therefore unrealistic. In the second case, some parts of the skill trainer could remain visible, thus worsening the user's sensation of immersion and realism. To ensure the perfect dimensioning of the virtual model over the skill trainer, the Gaumard® Lumbar Puncture Trainer was scanned, to create a virtual counterpart.

For the scanning process, a non-contact 3D laser scanner (Konika Minolta® Range 7) was used. To facilitate the scanning, registration was done through markers applied directly on the skill trainer (*Figure 22*). Given the opacity of the manikin material, it was not necessary to apply the opacifying spray on it.

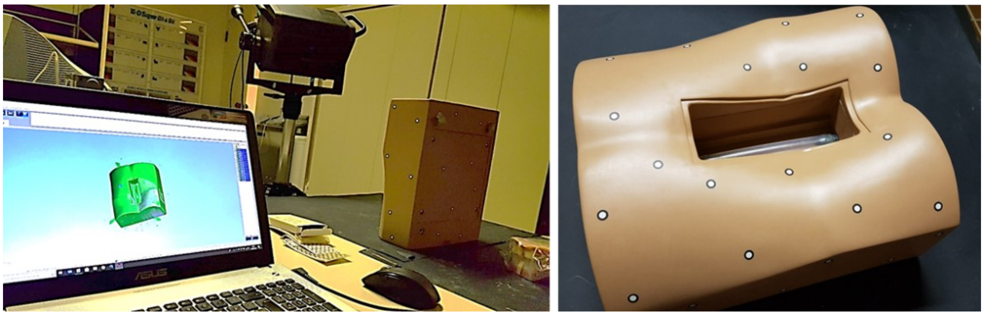

*Figure 22: Lumbar puncture trainer scanning (left) and markers application (right)*

### 3.2.2.5. Unity Project

All the steps described above have been put together in Unity, to create the main scene of the application.

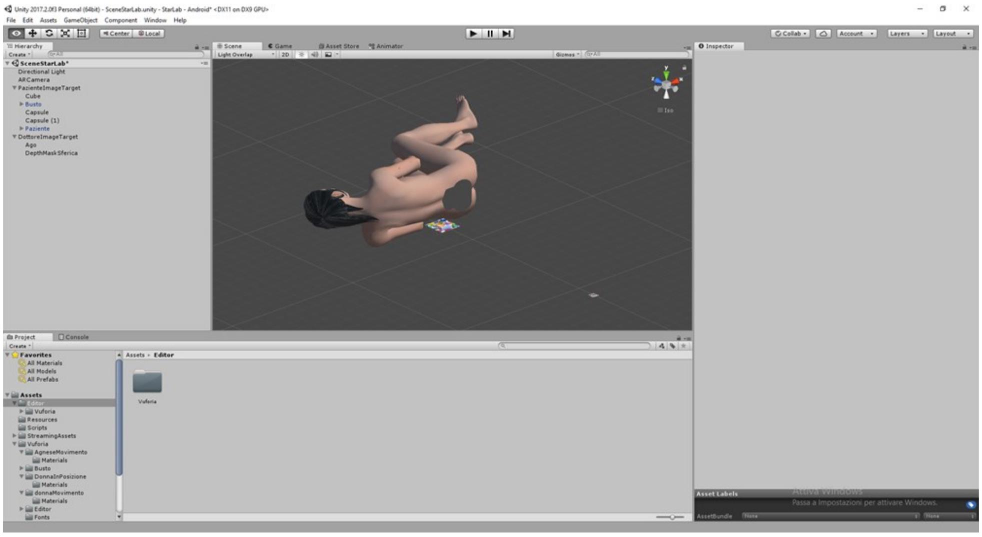

*Figure 23: Unity user interface – Scene*

<u>Scene</u>

The scene contains the following elements:

- Directional Light: light that realistically illuminates the virtual objects
- AR Camera: this element is linked to the use of Vuforia, and contains the main information relating to target tracking such as:
  - Vuforia App License Key
  - Digital eyewear and viewer type
  - Dataset used, i.e. list of targets to be located
- Patient Target Image: it is the 2D target image that represents the patient, 15x15cm in size. This target is integral with the manikin for rachicentesis, and therefore it is fixed on the desk. For this tracker, the "extended tracking" option is activated. This means that, even if the target image is no longer in the user's field of view, the system, based on the available sensors, hypothesises its position and continues to overlap the virtual

model to the manikin. This option is activated because it is possible that, during the simulation, the student focuses on the manikin and the target image goes outside his/her field of view.

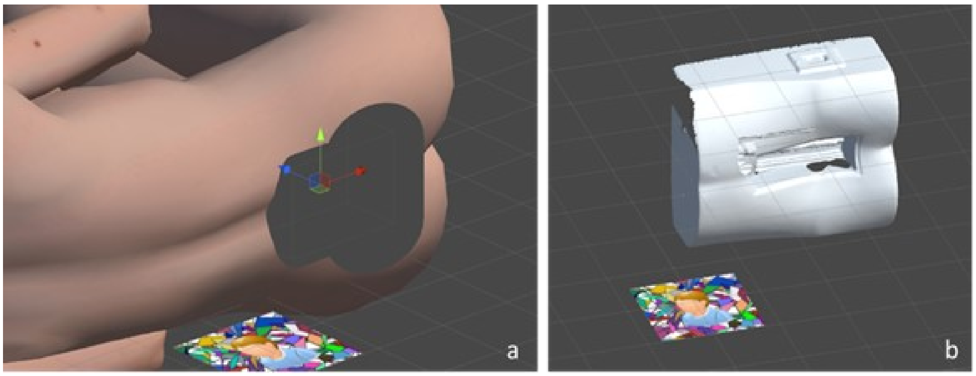


*Figure 24: Cube Box Collider and patient 3D model with depth masks (a); Virtual manikin/target positioning (b)*

The Patient Target Image has the following children:

- Cube: it is a parallelepiped hidden within the virtual model of the patient, immediately behind the operation area (*Figure 24(a)*). It is equipped with its own "Box Collider", whose function will be explained later;
- Bust: it represents the virtual model of the manikin and is used to correctly dimensioning and positioning the 3D patient model *(Figure 24(b));*
- Capsules: they are two capsule-shaped elements, positioned in correspondence with the operation area *(Figure 24(a))*. They are rendered as "Depth Mask" so that the user can see through them, while the application is running. Their function is to ensure that

the student can always see the operation area on the real skill trainer (the rest of the manikin is not visible, since it is covered by the virtual patient);

  - Patient: 3D model of the patient in the left lateral decubitus or sitting position *(Figure 24(a))*.

- Doctor Target Image: it is the 2D target image that represents the doctor, 5x5cm in size. During the rachicentesis procedure, this target is fixed to the student's hand through adhesive tape. For this reason, it is smaller than the other target.

  The Doctor Target Image has the following children:

  - Needle: it is a spherical collider, with a radius of 20cm, located in correspondence with the doctor target *(Figure 25).* A radius of 20cm was selected because it represents the average distance between Doctor Target and the tip of the needle held in the student's hand during the rachicentesis procedure. Its function will be clarified later.
  - Spherical Depth Mask: it is a depth mask centred on the doctor target, and therefore integral with the student's hand during the use of the application. Thanks to this element, the student can always see his/her hand during the operation. Otherwise, the hand could be covered by the virtual image of the patient.

    The "occlusion" issue is well known in the field of augmented reality, but neither advanced devices, such as Hololens, are yet able to completely solve it. Indeed, although Hololens is equipped with a depth sensor capable of mapping the surrounding environment, it is deactivated for distances under 0.5m from the device, thus it is not able to track the hands' position in real-time.

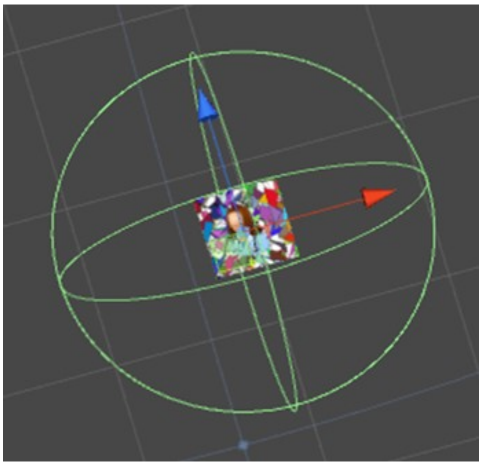

*Figure 25: Needle spherical collider*

<u>Movement and auditive feedback triggering</u>

To enhance the sense of realism, an animation of the scene has been provided. The animation (i.e. the spasm movement of the patient) is triggered by the contact between "Needle" and "Cube" colliders, namely when the student touches the manikin with the needle during the simulation. A C# script in Visual Studio has been developed for the triggering.

The animation is activated by the "move" Boolean parameter of the animator. The animator is a file in Unity, which represents the sequence of states present in the scene. The used animator consists of two main states: the "Rest" state, and the "Move" state. As soon as the application is started, the "Rest" state is activated. In this state, the 3D model is stationary. The transition to the "Move" state occurs when the "move" Boolean parameter is set to true; the inverse transition occurs when "move" is set to false. "Move" is set to false by the script, when the needle comes out of the cube (i.e. when it comes out from the manikin). This condition is necessary to prevent the patient from moving in a loop when the needle has been inserted into the manikin. Moreover, a 3-seconds cooldown prevents the animation from re-run before the previous one is finished. It may happen with uncertain movements in inserting the needle or with several repeated attempts.

In addition to the spasm of the virtual patient, audio files reproducing the patient's laments and complaints due to pain are randomly played.

#### 3.2.2.6. Calibration and Final Prototype

To ensure that the virtual patient is precisely superimposed over the real skill trainer, the Patient Target Image must be correctly positioned with respect to the manikin (as described in *Figure 24*). Therefore, a simple "calibration sheet" has been created starting from the Patient Target Image. This sheet allows to correctly position the manikin above the relevant profile (*Figure 26*), and thus to have the correct alignment between skill trainer and virtual patient.

*Figure 26: Calibration sheet*

As explained in paragraph 3.2.1, the AR application has been developed to be used with two different devices. In *Figure 27,* it is shown the screenshot from the smartphone used with Vox Gear Plus. In *Figure 28*, it is shown the screenshots of the application using Hololens, with different transparency levels of the virtual content.

Their usage, effectiveness, advantages, and limitations are discussed in paragraph 4.3.

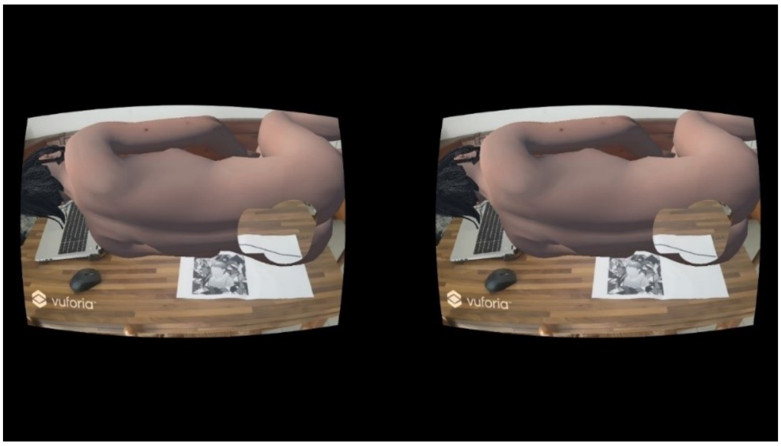


*Figure 27: Smartphone screenshot (used with Vox Gear Plus)*

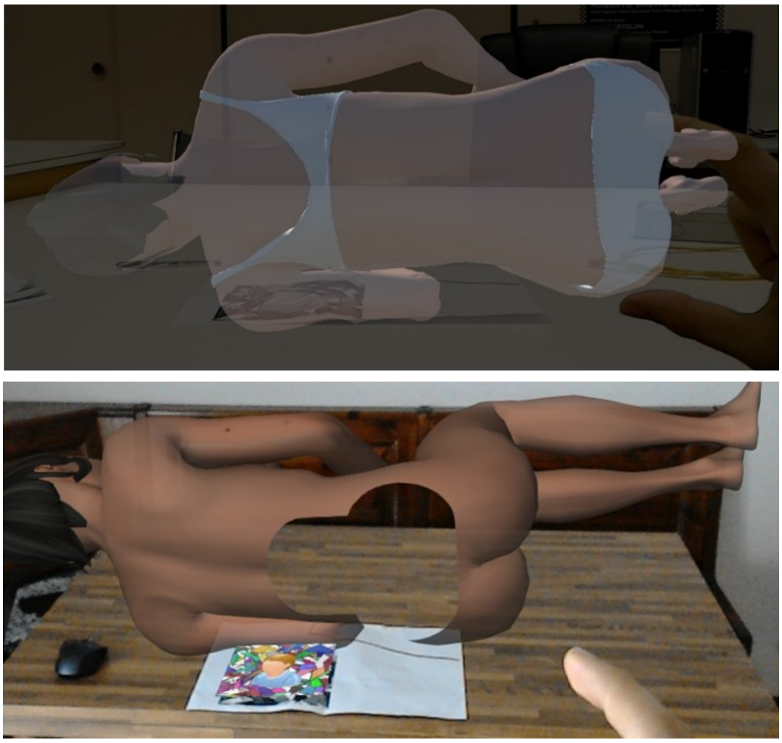

*Figure 28: Hololens screenshots (with different models and levels of transparency)*

# 4. Case Studies

The study has been divided into three main different case studies: the one related to the high-fidelity simulation for the emergency and trauma management, the one related to the low-fidelity simulation for the rachicentesis procedure, and the one related to the rachicentesis simulation in mixed reality. The first two cases consist in descriptive observational studies for the assessment of the simulation effectiveness in terms of performance, anxiety, frustration, stress, and cognitive load. The third case is a pilot study for the assessment of the developed MR rachicentesis simulation and comprehends the analysis of the user experience about the use of Hololens and Vox Gear Plus.

Students of the 6$^{th}$ year of Medicine and Surgery degree course, from Università Politecnica delle Marche (Italy), were enrolled in the study.

In each case study, students were submitted to the analysis of stress, cognitive load, and performance through psychometric and physiological measurements, following the structured procedure proposed in Chapter 3.1.

For the objective measure of stress and cognitive load, the same instrumentation for the biometric data collection was used for high-fidelity, low-fidelity, and MR simulations. In particular, the following smart chest and wrist bands have been used:

- Zephyr BioHarness™ BH3 chest belt (*Figure 29(a)*). The Zephyr ™ Psychophysical Health Monitoring System is the result of the integration between the BioHarness™ module technology and the OmniSense™ software that allows measuring the user's aerobic profile in real-time. The

OmniSense Analysis software module combines heart rate, heart rate variability, respiratory rate, acceleration, posture, and skin temperature measures, all integrated into a unique analysis tool.

- Empatica E4® Wristband Rev.2 (*Figure 29(b)*). The Empatica E4 bracelet is a wearable device that allows the acquisition of physiological data in real-time. Physiological data include electrodermal activity, blood volume pulse (BVP), acceleration, heart rate, and skin temperature.

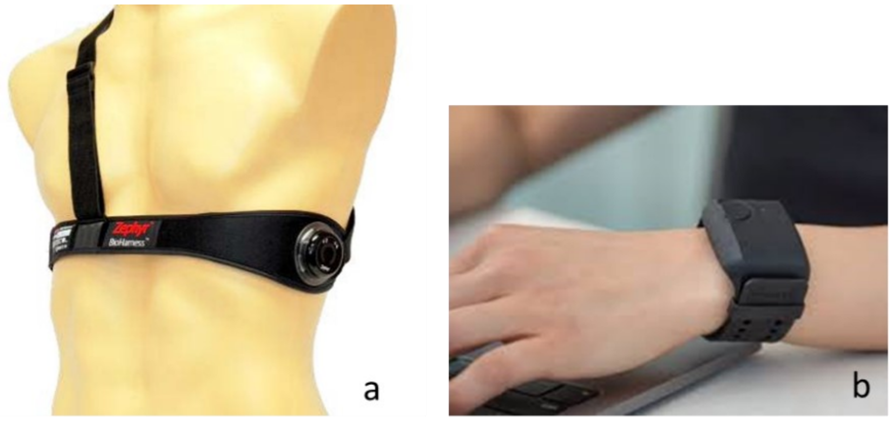


*Figure 29: Zephyr BioHarness™ BH3 chest belt (a), and Empatica E4® Wristband (b)*

At the beginning of the study, after reading the informative note and listening to the detailed explanation about the trial, all participants were asked to fill in the informed consent and the authorization for the processing of personal data. Subjects who refused to sign informed consent or the authorization to process personal data were excluded from the study.

Even the students who fell within the “contraindication cases” provided by the use of the wearable devices and the HMD for AR were excluded from the study. Based on the user manuals, the following relevant contraindications have been identified for each device:

- Zephyr BioHarness™ BH3 chest belt should not be used by:
    - Anyone with prior evidence of skin irritation at any point where the chest band will be in contact with the subject;
    - Users who have a cardiac pacemaker or automatic defibrillator.
- Empatica E4® Wristband Rev. 2 is forbidden to use:
    - In case of injured skin or other types of skin diseases in the area where the E4 is worn.
    - Among all E4 users, two cases of skin allergic reactions to light intensity were reported. These people were also allergic to other types of substances.
- Vox Gear Plus AR viewer should not be used:
    - In the case of strabismus, amblyopia, and anisometropia.
    - People prone to motion sickness in the real world, have a high risk of experiencing discomfort while using the device;
    - It is recommended to consult a doctor before using the device in case of pregnancy, psychiatric disorders, heart problems, pre-existing visual abnormalities, or other pathologies.
- HoloLens is not forbidden in any case. However, visual discomfort or fatigue, during the use, may be possible.

Finally, severely overweight or obese subjects (BMI> 30), subjects suffering from certain chronic diseases, celiac disease, who have had stressful events (e.g. severe bereavement, separation/divorce, abortions or transfers) in the last six months, who were taking certain categories of drugs (beta-blockers, diuretics, glucocorticoids or hormone therapies with oestrogen and progestogen) and who followed a vegan diet were excluded from the salivary cortisol collection. Indeed, these conditions can significantly alter salivary cortisol levels, making it unreliable for the evaluation of the stress level related to the simulation. It was investigated through the administration of a "participant recruitment form" about this personal information.

## 4.1. High-Fidelity Simulation

### *4.1.1. Emergency Management Simulation*

The high-fidelity simulation (HFS) concerns in critical care and trauma simulations for the patient management in the emergency room, with a high-fidelity manikin. HFS is a group simulation useful not only for technical practice but also, and above all, for the training of non-technical skills such as team-working, decision-making, and communicative skills.

In this case study, a standard simulation flowchart was followed. Indeed, HFS was preceded by the teacher briefing about essential skills, students' roles, learning environment, and equipment.

The simulation room was equipped with the instrumentation of an emergency room. The patient was the high-fidelity manikin SimMan®3G by Laerdal, connected to a monitor that showed the variations of physiological parameters according to the simulation program and students' actions. The stretcher was placed at the centre of the room near an infusion holder and a bench to practice cardiac massage. The monitor was placed over a cart that contained all the needed medical tools. The defibrillator was placed on one side of the room, on the right of the manikin. All around the room, near the walls, there were chairs for students that watched the simulation. The teacher who performed the phone calls and guided the simulations was placed behind a mobile wall in the same room. Therefore, the room layout was not optimal, since students performing the simulation, students watching the simulation, and the teacher should not be in the same room. Indeed, the simulation room should be divided into three separated parts: the one for the simulation itself, the control room for the teacher, and the part for the students assisting to the real-time video recording of the simulation.

In this study, each simulation session consisted of three consequential scenarios (S1, S2, and S3) of increasing difficulty. For each simulation session, one students' group was convened and divided into three sub-groups: each sub-group participated actively in one scenario and passively (observation) in the other two.

Each scenario started with an emergency phone call and all the students heard the speaker describing the health conditions of the patient who was arriving in the emergency room. Students had to define their roles among themselves, and identify a group leader, in order to accomplish the simulation's goal, under the external supervision of the teacher. The goal of the simulation was always the stabilization of the patient's health conditions. In some scenarios, students also have to perform cardiopulmonary resuscitation (CPR), combining chest compression and artificial ventilation. However, the evaluated skills for scenario S1, S2, and S3 are the same, and correspond to the six phases shown in *Table 6*:

*Table 6: Phases and skills assessment for simulations of emergency management*

| **Phase** | **Skills** |
|---|---|
| Preparatory Phase | Communication with the team |
| | Notification to doctor and services |
| | Distribution of tasks |
| | Environmental check |
| | Materials check |
| | Drugs and infusive solutions check |

| | |
|---|---|
| | Instruments check |
| | PPE for all operators check |
| | Collection of information from the extra-hospital team |
| | General impression |
| Phase A (Airway) | Manual stabilisation of cervical rachis |
| | Leave the head to the manoeuvre leader |
| | Removal of the helmet |
| | Application rigid cervical collar |
| | Oral cavity check |
| | Cleaning and aspiration of the oral cavity |
| | Trauma jaw trust |
| | Trauma chin lift |
| | Oropharyngeal canula positioning |
| | Nasopharyngeal canula positioning |
| Phase B (Breathing) | Chest observation |
| | Chest palpation and percussion |
| | Chest auscultation |
| | Respiratory rate detection |
| | Pulse oximetry verification and arterial blood gas analysis |
| | Gives oxygen at high flows |
| | Ventilation with two operators |
| | Collaboration in orotracheal intubation |
| | Explorative-decompression pneumothorax puncture |
| | Positioning the chest drainage |

| | |
|---|---|
| Phase C (Circulation) | Peripheral pulse assessment |
| | Heart rate detection |
| | Blood pressure detection |
| | Cardiac monitoring |
| | The positioning of 2 venous accesses and blood sampling |
| | Administration of hot blood transfusion liquids |
| | Administration of drugs (specify) |
| | Response to liquid infusion assessment |
| | Impressive haemorrhage tamponing |
| | Pelvic stability and abdomen assessment |
| Phase D (Disability) | Eyes opening assessment |
| | Better motor response assessment |
| | Better verbal response assessment |
| | Pupils check |
| Phase E (Exposure) | Undress the patient |
| | Log roll |
| | Thermal protection |
| | Fractures stabilisation |
| | Bladder catheter positioning |
| | Gastric probe positioning |

Each scenario lasted about twenty minutes and was followed by a debriefing phase in which the teacher asked the learners how they felt during and after the HFS and discussed the students' perceptions, actions, decisions, and errors.

One video camera was placed above the manikin to monitor and record all the learners' actions. Another video camera recorded the entire simulation, from the briefing until the end of the debriefing. *Figure 30* shows a couple of frames of the HFS video recording.

*Figure 30: Frames of high-fidelity simulations*

### *4.1.2. Data Acquisition Workflow*

182 students were enrolled in the 6th year of Medicine and Surgery degree course of Università Politecnica delle Marche. Their degree curriculum provided for the execution of high-fidelity and low-fidelity simulations for the training of technical and non-technical skills. Therefore, this study was accomplished not as a laboratory research but as an observational study on the real field.

Concerning the high-fidelity simulation, the 182 students were divided into 13 groups composed of 14 students each. A total of 148 students signed the informed consent to participate in the study, with the consent form for the processing of personal data. Students who did not sign these forms were excluded from the study but, obviously, not from the simulation training.

Each simulation session lasted about four hours, from 9:00 a.m. to 1:00 p.m. in the morning and from 2:30 p.m. to 6:30 p.m. in the afternoon. Only one group (fourteen students) for each simulation session was convened. After a preparation phase (*Figure 31*), the fourteen students were divided into three subgroups of four, five, and five participants to accomplish three different and consecutive simulation scenarios (S1, S2, and S3) of increasing difficulty. Each subgroup performed one simulation and observed the other two.

At the arrival of the fourteen students in the classroom, before dividing them into three subgroups, the workflow in *Figure 31* was followed:

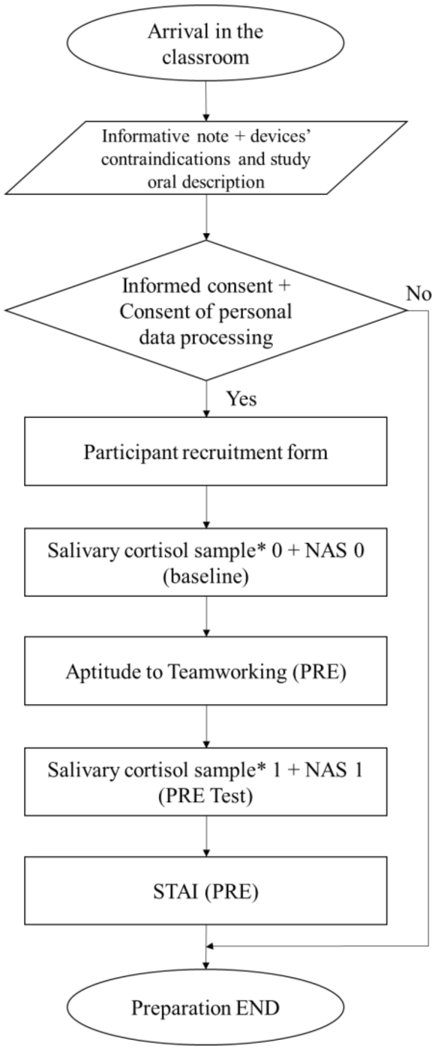


*Figure 31: Students preparation workflow for each simulation session (*suitable students)*

For each simulation session, the study was described in detail together with implications and possible contraindications in the use of the devices. An extensive informative note was administered to all the fourteen students. Then, the participant recruitment form, with demographic and personal

information, was presented to the students who gave informed consent and signed the form for personal data processing. According to the participant recruitment form (where they marked potential factors that may influence the cortisol analysis), a salivary cortisol sample and the NAS scale were collected from the suitable students to register the basal stress level. Then, the questionnaire session about aptitude to simulation and team-working was administered. After 10 minutes from the first sample, another salivette for cortisol analysis was acquired with the NAS scale. The STAI questionnaire was given to the students to assess their perceived level of anxiety in their daily life (Trait) and in that precise moment (State, before the simulation).
Students who did not sign the informed consent were not involved in this data acquisition.

At the end of this preparatory workflow, students were randomly divided into three subgroups. Four students were assigned to the first and easiest scenario (S1). The other two, more difficult scenarios (S2 and S3), were conducted by five students each. Due to the limited availability of wearable devices in our University Department, two students for each subgroup were randomly selected to wear the chest belt Zephyr BioHarness and the wrist band Empatica E4. Therefore, physiological signals were collected on a total of 78 students, while all the 148 participants answered the self-assessment psychometric surveys. Simulations were always video recorded, and performance was assessed for all the students.

For each scenario, while the participants in the study followed the workflow in *Figure 32*, the other students simply performed the simulation through the phases of the briefing, simulated scenario, and debriefing.

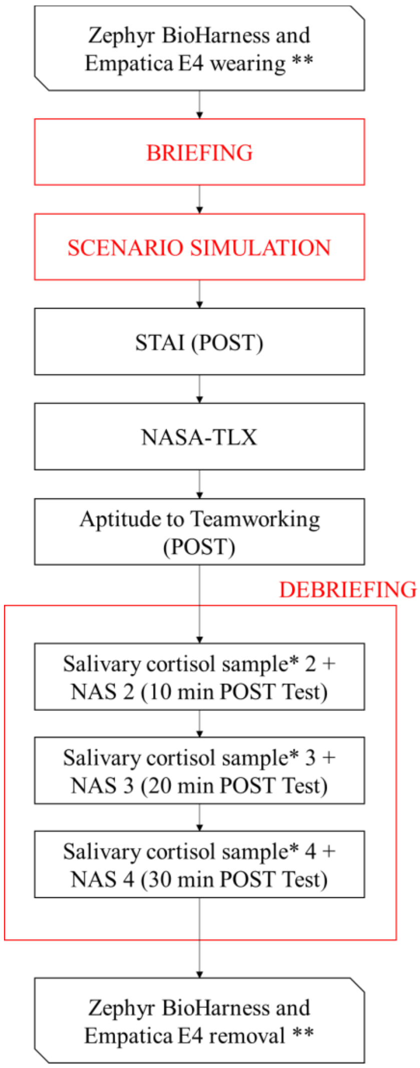


*Figure 32: Study's procedure for the assessment of students' stress and cognitive load during HFS (* suitable students; ** two randomly extracted participants)*

For each simulation scenario (S1, S2, S3), before the briefing, participants wore the wearable devices (Zephyr and E4) for the biometric monitoring. The physiological signals were recorded during the entire simulation, from the briefing to the scenario simulation, until the end of the debriefing. After having simulated the scenario, the following questionnaires were administered to the participants:

- STAI (POST): only the “state” section, to assess the impact of the simulation on perceived anxiety;
- NASA-TLX: to assess the perceived cognitive load (mental demand) and other emotional factors influencing the students’ cognitive state (such as perceived effort, frustration, temporal demand) and physical state (physical demand);
- APTITUDE TO TEAM-WORKING: to know the possible variation in students’ opinion about simulation and team-working, after having experienced a high-fidelity simulation.

Three samples of salivary cortisol were collected with the NAS, 10, 20, and 30 minutes after the end of the scenario, in order to analyse the variations in stress level (it is expected a peak ten minutes after the stressful event and a consecutive decrement).

### *4.1.3. Participants*

As described in the previous paragraph, 148 students of the 6$^{th}$ year of Medicine and Surgery degree course were enrolled in the observational study. Their characteristics are summarised in *Table 7*.

*Table 7: Participants characteristics*

| | % OF PARTICIPANTS |
|---|---|
| GENDER | |
| • Male<br>• Female | 42.57 %<br>57.43 % |
| SMOKER* | |
| • Yes<br>• No | 20.27 %<br>66.22 % |
| PERTINENT THESIS** | |
| • Yes<br>• No | 33.78 %<br>44.59 % |
| PREVIOUS EXPERIENCE IN | |
| • Group simulation<br>• Residency<br>• Professional training activity<br>• Training course (e.g. BLS, ATLS)<br>• Working experience<br>• Health volunteer activities<br>• Invasive procedure | 79.05 %<br>55.41 %<br>97.57 %<br>42.57 %<br>13.51 %<br>10.14 %<br>27.70 % |

(*) 13.51% did not answer
(**) 21.62% did not answer

Among participants, 57.43% were women and 42.57% were men, with an average age of 26 (± 1.02) years. Even weight and height were registered for each one of them. 20.27% of them was a smoker and 33.78% of them was developing a degree thesis pertinent to the investigated high-fidelity simulation. The courses considered pertinent to this HFS were the ones attended in the following wards: anaesthesia and intensive care, cardiology, every surgery, emergency, gastroenterology, obstetrics and gynaecology, orthopaedics, pneumology, radiology, and urology. Moreover, most of them had already performed a group simulation (79.05 %) and/or other kinds of practical training activities.

*Table 8* shows the subjective psychological state of participants before taking part in the high-fidelity simulation.

While STAI Trait describes the general level of anxiety in daily life, STAI State PRE shows the anxiety level perceived by the students, before performing the simulation. They are both scored on a 4-points Likert scale, thus the total score range between 20 and 80 for both Trait and State sections. It seems that, on average, the participants are more anxious in their life than in the classroom at that moment.

NAS 0 and NAS 1 are indexes of the stress level perceived by the students at the arrival in the classroom, and ten minutes later, before beginning the simulation briefing. They are scored on a 10-points scale (with 0 = not at all stressed, and 10 = extremely stressed).

*Table 8: Subjective psychological state of participants before HFS*

| | MEAN | STANDARD DEV. |
|---|---|---|
| STAI TRAIT | 48.06 | 9.95 |
| STAI STATE PRE | 46.14 | 10.50 |
| NAS 0 | 3.70 | 2.47 |
| NAS 1 | 4.00 | 2.42 |

*Figure 33* and *Figure 34* show the results of the survey about aptitude to simulation and team-working, described in *Table 2*, and administered before the HFS. It is divided into two sections.

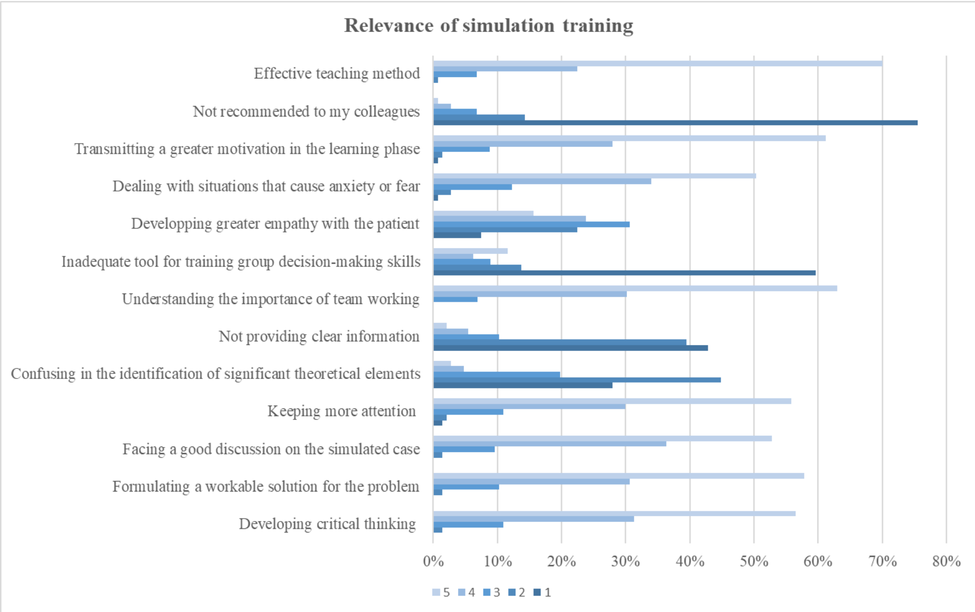


*Figure 33: Results of the survey about aptitude to simulation and team-working (Section: Relevance of simulation)*

By analysing students' opinions about the relevance of simulation (*Figure 33*), it emerges that it is considered an effective teaching method that transmits a greater motivation in the learning phase, helping to understand the importance of team-working. It is considered also useful to keep more attention, develop critical thinking and decision-making skills, face good discussions on the simulated case, and formulate a workable solution for the problem, clarifying the actions to perform. Conversely, the simulation does not seem to foster the development of empathy with the patient (Brunzini, et al., 2019). However, more than 84% of participants, declared that simulation can help to deal with situations that cause anxiety or fear.

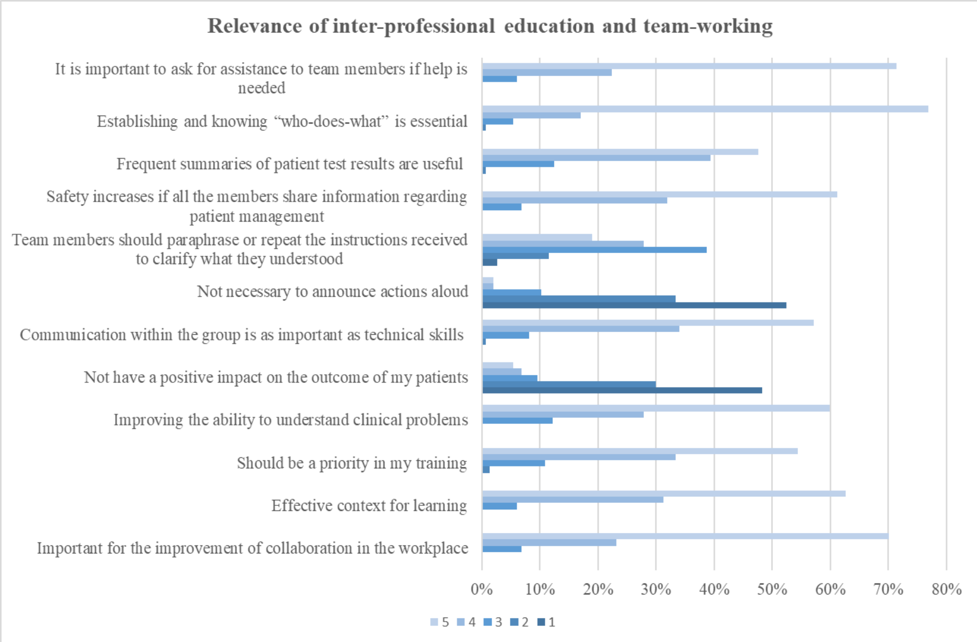


*Figure 34: Results of the survey about aptitude to simulation and team-working (Section: Relevance of inter-professional education and team-working)*

The analysis (*Figure 34*) shows that, on average, 90% of students think that learning together with other professional figures is an effective context for learning, and should be a training priority. Indeed, learning shared with other professionals is thought to improve the ability to understand clinical problems, and it is also supposed important for the improvement of collaboration in the workplace. However, not all students agree in thinking that interprofessional learning has a positive impact on patient outcomes.

Going into more detail of inter-professional education, 93.89 % stated that establishing and knowing "who-does-what", within the team, is essential for improving the quality of the provided care. Concerning the communication among team members, 91.15 % agree in considering communication as important as technical skills. The most of respondents consider it essential that team members ask for assistance if they need help completing a task and share information regarding patient management (on average 93.5% agree and strongly agree). On the other hand, even if 85.71% of students think that frequent summaries of patient test results are useful for keeping team members' attention to patient needs, the importance to paraphrase or repeat the instructions received to clarify what team members understood is perceived less (only 46.94 % agree and strongly agree).

### *4.1.4. Performance Analysis*

For each scenario of each simulation session, performance has been assessed per phase, in order to obtain a total performance score.

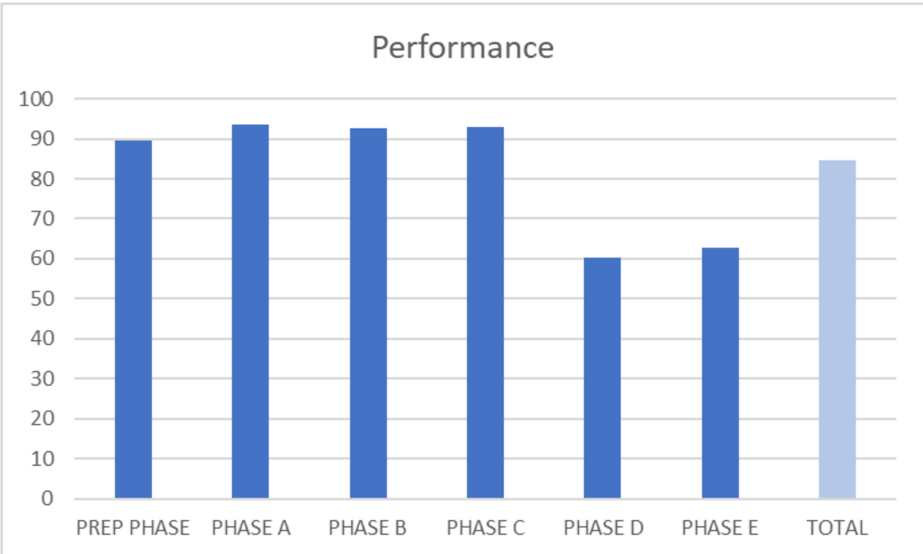


*Figure 35: HFS mean performance per phases and mean total score*

*Figure 35* shows the mean performance results, per phases, calculated over 148 students. On average, students achieved good results, with a mean total score of around 85/100 points. Performace is higher in the first four phases, while it decreases toward the end of the simulation, during the patient's consciousness assessment (Phase D) and exposure (Phase E).

Further analysis of performance in different scenarios is shown later in *Figure 51*.

### *4.1.5. Psychometric Analysis*

In this paragraph, results related to self-assessment questionnaires for the analysis of cognitive states are illustrated.

#### 4.1.5.1. Aptitude to simulation training and teamworking

*Figure 36* shows the comparison between students' points of view about the relevance of simulation training, before (PRE) and after (POST) having taken part in the HFS. Eleven participants did not answer the post-simulation survey; thus, their pre-simulation answers have been excluded from the comparison.

Results in *Figure 36* represent the mean values of each item (on the 5-points Likert scale), calculated on the answers of 92.57% of participants.

An overall slight improvement in the perception of simulation relevance can be observed in students' post-simulation responses. The greatest difference between pre- and post-simulation is about the importance of the simulation in developing empathy with the patient. An enhancement can be observed also in the perception of usefulness in dealing with situations that can cause anxiety and fear. This means that, during the HFS, students felt realistic emotional involvement, related both to the patient and to the emergency and critical situation. Conversely, after the HFS, students' opinion about simulation benefit toward the development of decision-making skills gets worse. It may be explained by the several disagreements among team members about patient management, happened during the scenarios' simulation. This issue should be solved during the debriefing.

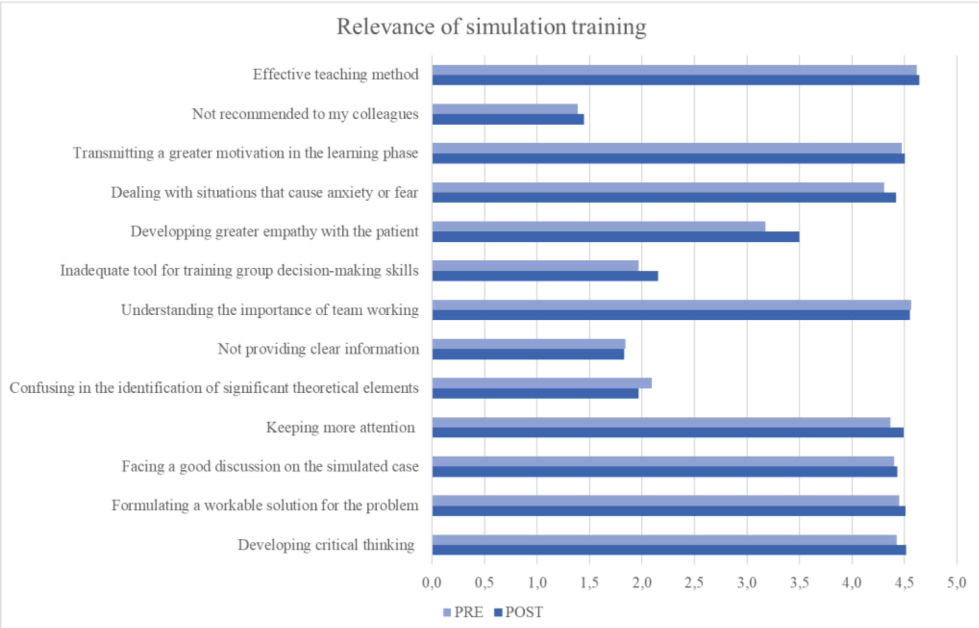


*Figure 36: Student's aptitude toward simulation training: comparison between opinions collected before and after the HFS*

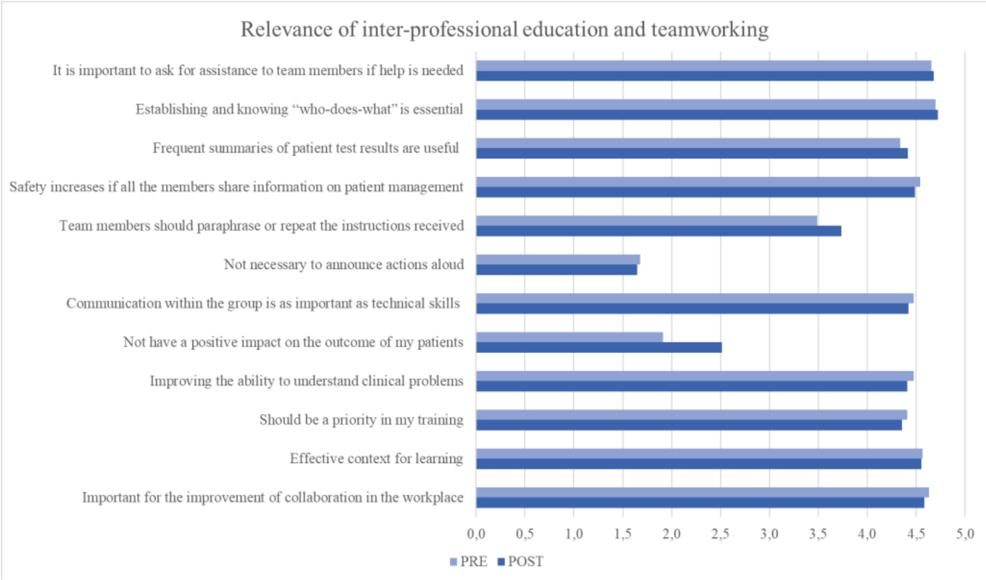


*Figure 37: Student's attitude toward interprofessional education and team-working: comparison between opinions collected before and after the HFS*

In the same way, the answers' mean values of 92.57% of participants about the relevance of interprofessional education and team-working, collected before (PRE) and after (POST) the HFS, are shown in *Figure 37*.
An overall slight improvement concerns the importance of communication among team members. Indeed, after the HFS, students strengthened their belief in the relevance of asking for help, summarising patient test results, establishing "who-does-what", and, above all, paraphrasing and repeating the instruction received.
On the other side, after the HFS, a slight worsening concerns the students' points of view related to the relevance of inter-professional education. This may be explained by the fact that the participants in the HF simulations were

students enrolled in the same degree course; thus the actual concept of inter-professionality was not perceived.

The greatest difference between pre- and post-simulation opinions is about the impact on patient outcome. Indeed, after the simulation, students were more inclined to believe that interprofessional learning does not have a positive impact on the outcome of the patients. As stated before, it can be because sometimes students argued about clinical choices rather than sharing their knowledge and skills to take a unique and proper decision. More attention to the importance of this topic should be placed by the instructor during the debriefing phase.

#### 4.1.5.2. State Trait Anxiety Inventory (STAI)

While three participants did not answer the STAI questionnaire before the simulation, other seven students did not answer after the simulation. Thus, a total of ten responses were excluded from the analysis.

Results in *Table 9* shows the mean values (20 questions on a 4-points Likert scale for each session) calculated on 138 responses (93.24% of participants).

*Table 9: STAI results before and after HFS*

| | MEAN | STANDARD DEV. |
|---|---|---|
| STAI TRAIT | 48.06 | 9.95 |
| STAI STATE PRE | 46.42 | 10.47 |
| STAI STATE POST | 42.51 | 12.00 |

Students seem generally more anxious in their life (STAI Trait) than during the lesson (STAI State). As expected, their feeling of anxiety decreases after the simulation (mean STAI State Post is lower than mean STAI State Pre). Moreover, while the STAI Trait mean value is ten points higher than the one expected for college students, the STAI State Post agrees with that one (Spielberger, et al., 1983). This means that the STAI State Pre shows the actual increment of anxiety perceived by the students before performing the simulation.

4.1.5.3. Numeric Analog Scale (NAS)

Concerning the analysis of the subjective stress, mean NAS values have been calculated for each sample: at the arrival in the classroom (NAS 0), ten minutes later (NAS 1), and ten (NAS 2), twenty (NAS 3), and thirty (NAS 4) minutes after the end of the HFS.

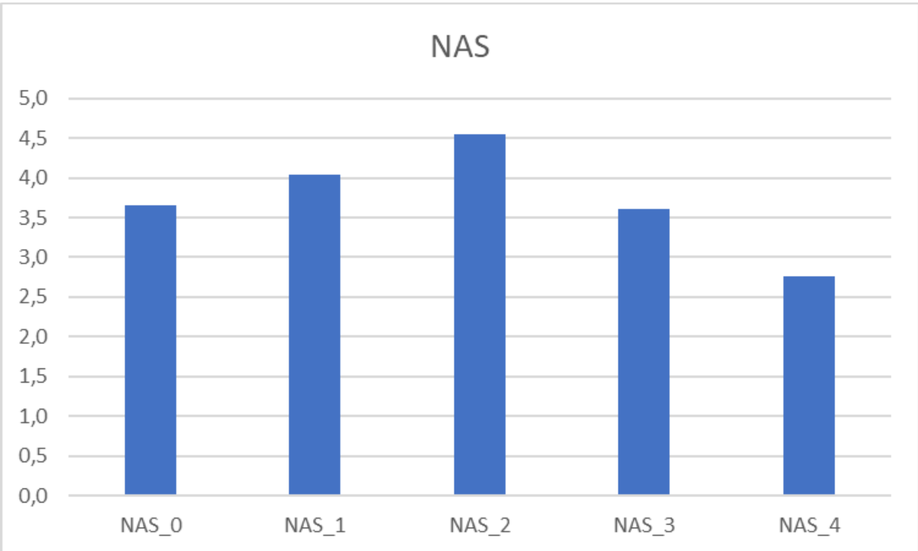


*Figure 38: NAS results (10-points Likert scale)*

Four students have been excluded from the analysis because they omitted some responses. Thus, the results in *Figure 38* are based on 144 students' answers (97.3%).
As expected, perceived stress follows a bell-shaped trend. Indeed, students' stress increments from NAS 0 to NAS 1 because the participants become more anxious and worried for the simulation to be executed. The peak in NAS 2 is obviously due to the just-finished stressful situation. Then, the feeling of stress gradually decreases during the debriefing (NAS 3). At the end of the debriefing, participants feel less stressed (NAS 4) than when they arrived in the classroom (NAS 0), confirming the effectiveness of the debriefing.

#### 4.1.5.4. NASA Task Load Index (NASA-TLX)

The answer rate of NASA-TLX is equal to 96.62%. *Table 10* shows the mean values of the total score for the perceived workload associated with the simulation, and the weight assigned to the six indexes of mental, physical, temporal demands, performance, effort, and frustration.

*Table 10: NASA-TLX results*

| | MEAN | STD. DEV. |
|---|---|---|
| NASA-TLX | 65.77 | 16.66 |
| Mental Demand | 305.38 | 120.89 |
| Physical Demand | 38.90 | 56.11 |
| Temporal Demand | 154.34 | 112.97 |
| Performance | 187.87 | 101.72 |

| | | |
|---|---|---|
| Effort | 155.31 | 97.28 |
| Frustration | 144.71 | 157.49 |

Sugarindra et al. considered workload low for scores in the range 0-9, medium for 10-29, rather high for 30-49, high for 50-79, and very high for score among 80 and 100 (Sugarindra, et al., 2017). According to their score interpretation, NASA-TLX after high-fidelity simulations is on average high (equal to 65.77). Therefore, students perceive the simulation as high demanding activity. The greatest weight is given by the mental demand that obtained a score considerably higher than the other indexes. Thus, it can be assumed that students perceived a noteworthy cognitive load in performing high fidelity simulations. Conversely, physical demand seems not to have an impact on the perceived workload. Concerning the other indexes (temporal demand, performance, effort, and frustration), approximately the same weight was assigned, on average, to each one.

#### *4.1.6. Biometric Analysis*

As provided for the proposed procedure for stress and CL assessment, several physiological parameters were collected during HF simulations: cardiac ones (HR, HRV, IBI), respiratory ones (BR), electrodermal activity (EDA), and so on. They were acquired from 78 participants using the chest band Zephyr BioHarness and the bracelet Empatica E4. Video analysis and events tracking were accomplished as well.

In this paragraph, stress and cognitive load, detected through the analysis of physiological signals, are discussed.

The proprietary algorithm and software module for cognitive states detection by *Phasya s.a. (Seraing, Belgium)* was used for signals analysis and identification of stress and CL levels. Information about the Phasya's algorithm for stress and CL detection, based on several physiological parameters, is protected by the non-disclosure agreement, and thus cannot be discussed in this work. However, some information about their core software module (for the drowsiness assessment) can be found in (François, et al., 2016) and (Stawarczy, et al., 2020).

The computed stress and CL levels range between 0 and 6.

By way of example, *Figure 39*, *Figure 40*, and *Figure 41* show the stress levels variation of one randomly selected participant, respectively related to heart rate (HR), breathing rate (BR), and electrodermal activity (EDA). Similarly, *Figure 42*, *Figure 43*, and *Figure 44* show the cognitive load levels of the same subject. These levels are given according to the simulation phase and the debriefing phase. Indeed, in every figure, the first black line indicates the beginning of the simulation, the second one indicates the end of the simulation, the third one shows the beginning of the debriefing and the fourth one shows the end of the debriefing.

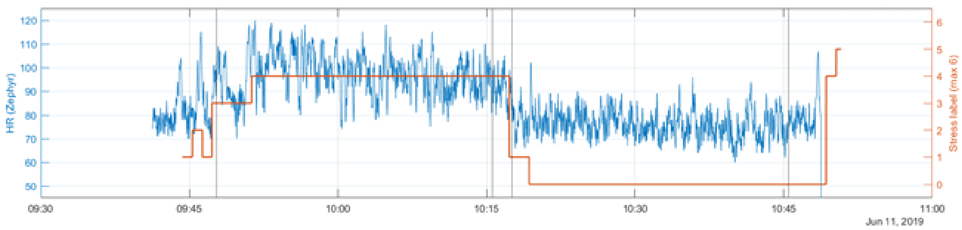


*Figure 39: Stress Level in relation to HR signal*

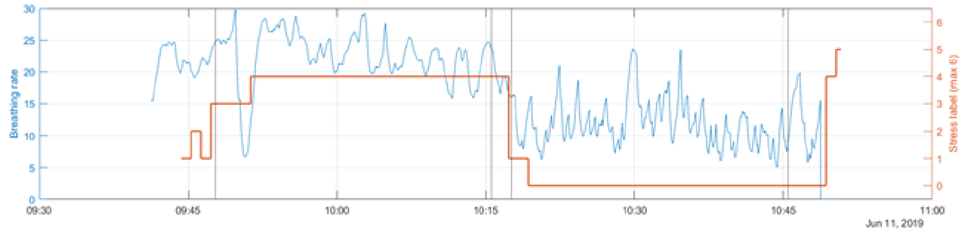


*Figure 40: Stress Level in relation to BR signal*

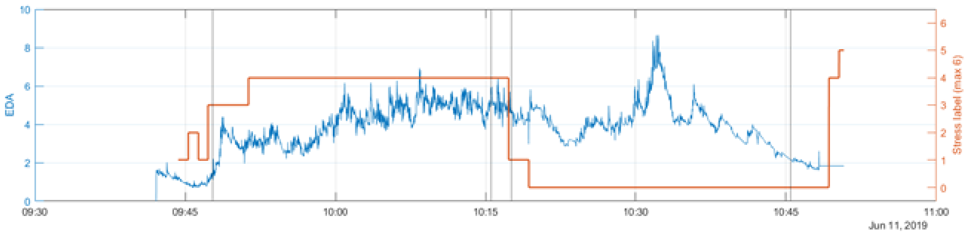


*Figure 41: Stress Level in relation to EDA signal*

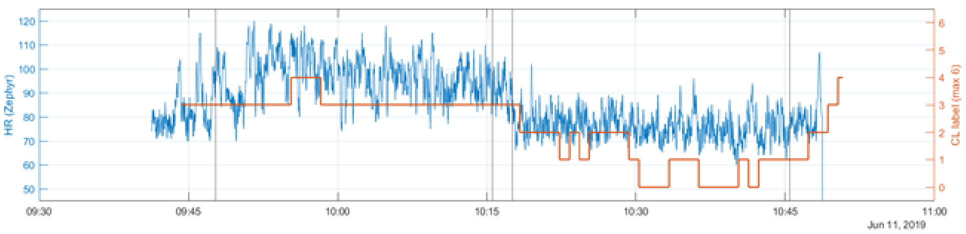


*Figure 42: Cognitive Load Level in relation to HR signal*

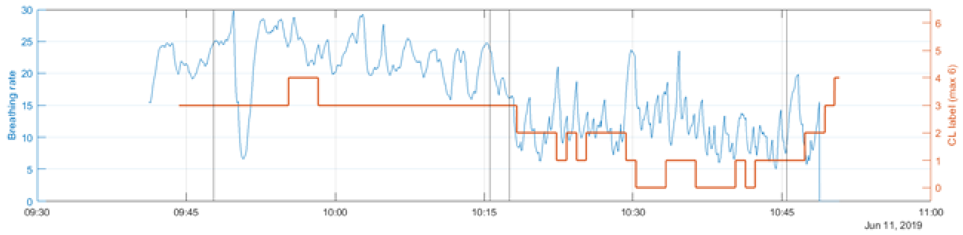


*Figure 43: Cognitive Load Level in relation to BR signal*

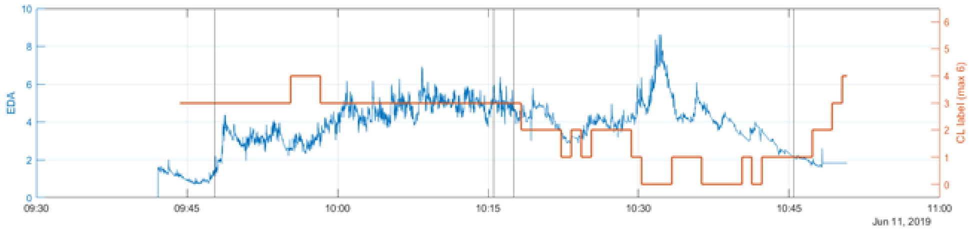

*Figure 44: Cognitive Load Level in relation to EDA signal*

*Figure 45* reports a comparison between stress and CL levels variations during simulation and debriefing phases, for the same subject.

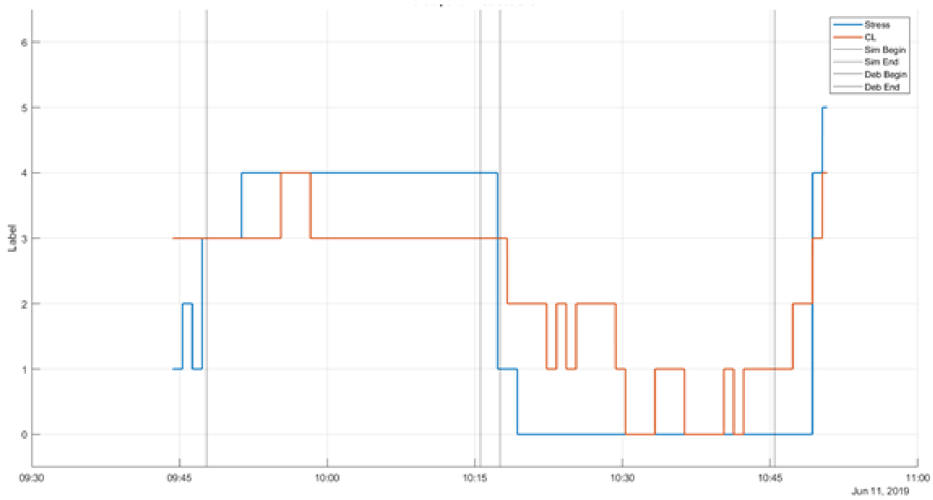


*Figure 45: Comparison of Stress and Cognitive Load Levels during simulation and debriefing phase for a randomly selected subject*

The same trend of stress and CL levels (between simulation phase and debriefing phase) can be found in most of the subjects (as later shown in *Figure 46* and *Table 11*). Indeed, this first inspective signal analysis reveals the expected stress decrement from the simulation to the debriefing phase. Concerning the levels of cognitive load, *Figure 45* shows that while it is quite constant and high during the simulation, it varies along with the debriefing mostly according to the student's interaction with the teacher.

#### 4.1.6.1. Stress and Cognitive Load Analysis for Simulation Phases

First, through Phasya's algorithm, the stress and CL levels were defined, for all the 78 participants who wore the smart devices, for the entire simulation and debriefing. Then, stress and CL levels were analysed for the most stressful simulation phases. Indeed, according to the literature (Endedijka, et al., 2018), the most stressful phases are the initial one (arrival of the patient and definition of roles) and the final one (resolution of the case with critical patient conditions). Moreover, a preliminary statistical analysis confirmed that the most stress-related phases are the A ($p<.01$) and the D ($p<.05$), and that also the cognitive load is mostly linked to phase A ($p<.01$). Therefore, phases A and D were identified through video analysis and events recorded for each scenario in every simulation session. The preparatory phase (that is between the beginning of the simulation and the beginning of phase A) was also discerned to be used as a 'comparison'. The mean stress and CL values for each phase were calculated through Phasya's algorithm.

Results shown in *Figure 46* and *Table 11* refer to 68 subjects. Indeed, two signal acquisitions were discarded because they presented a lack of integrity. Other eight subjects were excluded from the analysis because they performed cardiac massage during the simulation, and the variations in physiological parameters were related to the physical activity rather than to stress and CL.

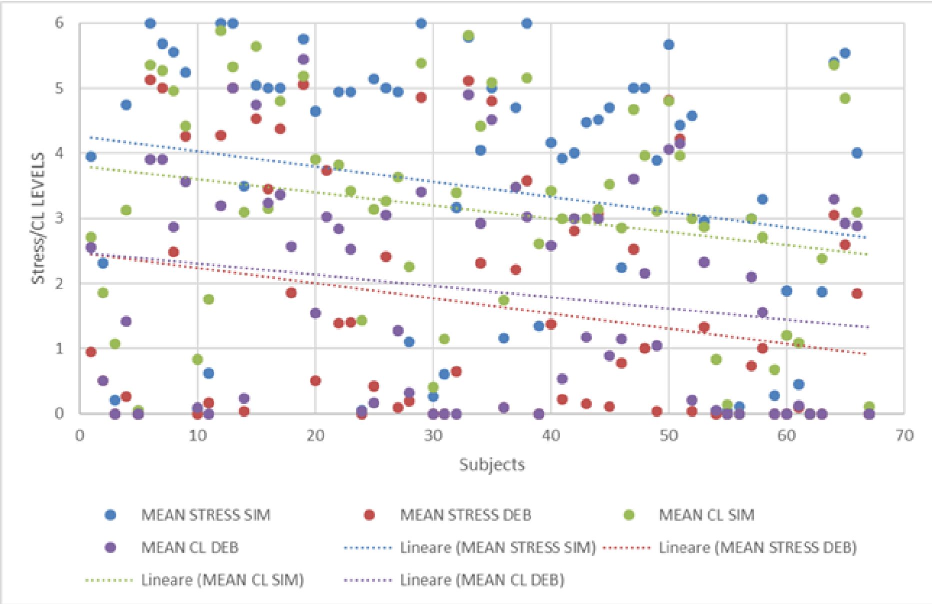


*Figure 46: Mean stress and CL levels during simulation and debriefing for each participant*

*Table 11: Mean stress and CL levels during the preparatory phase, phase A, phase D, and during the entire simulation and debriefing*

| | PREP PHASE | PHASE A | PHASE D | ENTIRE SIMUL | ENTIRE DEBRIEF |
|---|---|---|---|---|---|
| STRESS | 3.39 | 3.54 | 2.88 | 3.40 | 1.69 |
| CL | 2.98 | 3.19 | 2.44 | 3.05 | 1.89 |

Stress and CL levels range between 0 and 6. *Table 11* and *Figure 46* show that:

- On average, stress decreases from simulation to debriefing;
- On average, cognitive load decreases from simulation to debriefing;
- On average, from simulation to debriefing the decrease in stress is higher than the decrease in cognitive load;

- On average, during the simulation, stress is higher than cognitive load;
- On average, during the debriefing, cognitive load is higher than stress;

Moreover, from *Figure 46*, since subjects are sorted by simulation date, stress and CL seem to decrease over time.

From *Table 11*, it is confirmed that the average stress and CL in phase A are higher than the average stress and CL in the rest of the simulation.

In conclusion, if we define stress and CL values as very low/low for levels 1 and 2, medium for levels 3 and 4, and high/very high for levels 5 and 6, it is possible to assume that stress and CL levels are maintained on average on a medium level for the entire simulation. Moreover, the debriefing results effective on students' cognitive state, since during this practice stress and CL become both low. Post-simulation stress is well managed, and cognitive overload is avoided. Moreover, concerning the CL variations during the debriefing phase, a preliminary visual signal inspection was accomplished for each participant. The analysis revealed that CL changes during the debriefing according to teacher interaction in 64% of cases. The students' cognitive load increases while talking with the instructor about clinical decisions and actions undertaken during the simulation, thus confirming the correctness of the algorithm used to discriminate the CL.

#### 4.1.6.2. Stress and Cognitive Load Analysis for Simulation Scenarios

Another analysis has been done comparing mean stress and CL levels among the three different scenarios (S1, S2, and S3) of increasing difficulty. *Figure 47* shows that stress gradually decreases from scenario 1 to scenario 3 during both the simulation and the debriefing. Students performing the first scenario are more stressed than the others because they did not have the choice to see someone else doing the simulation and they were insecure and worried about how to approach the critical clinical case. On the other hand, students performing the third scenario were less stressed (even if it was the hardest and complicated scenario) because watching the previous two scenarios and debriefings helped them to understand how to organise themselves and manage the clinical case.

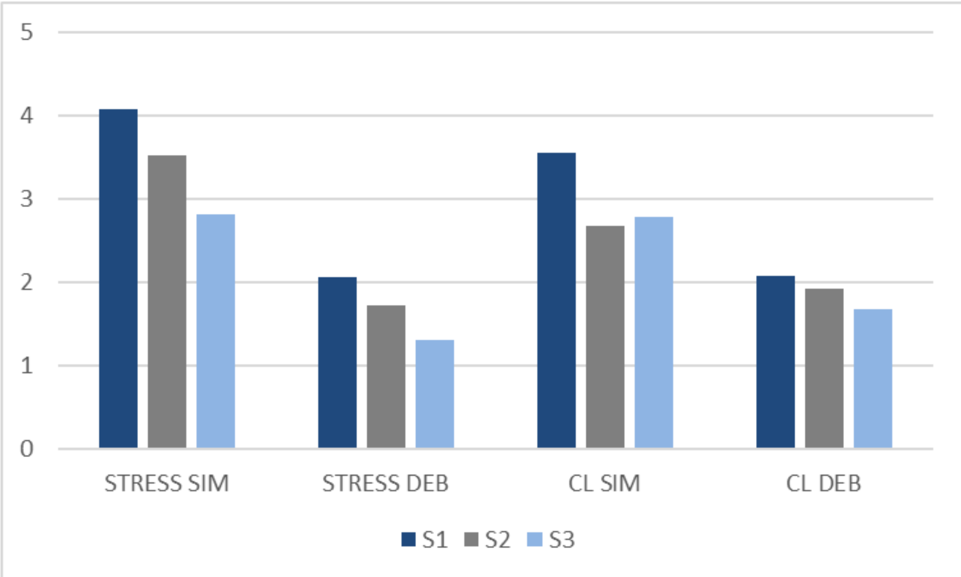


*Figure 47: Mean stress and CL levels during simulation and debriefing for scenarios S1, S2, and S3*

Concerning the cognitive load, the same trend as stress is visible for the debriefing, but not for the simulation. Indeed, during the simulation, the CL decreases from S1 to S2 and then increases from S2 to S3 (*Figure 47*). This may be due to the fact that, even if students performing S3 had already seen

S1 and S2 (and this allowed the stress reduction in subjects performing S3), the difficulty and complexity of the clinical case did not allow a further reduction of the mental demand.

*Figure 48* shows the mean stress levels in the preparation phase, phase A and phase D, in the three different scenarios.

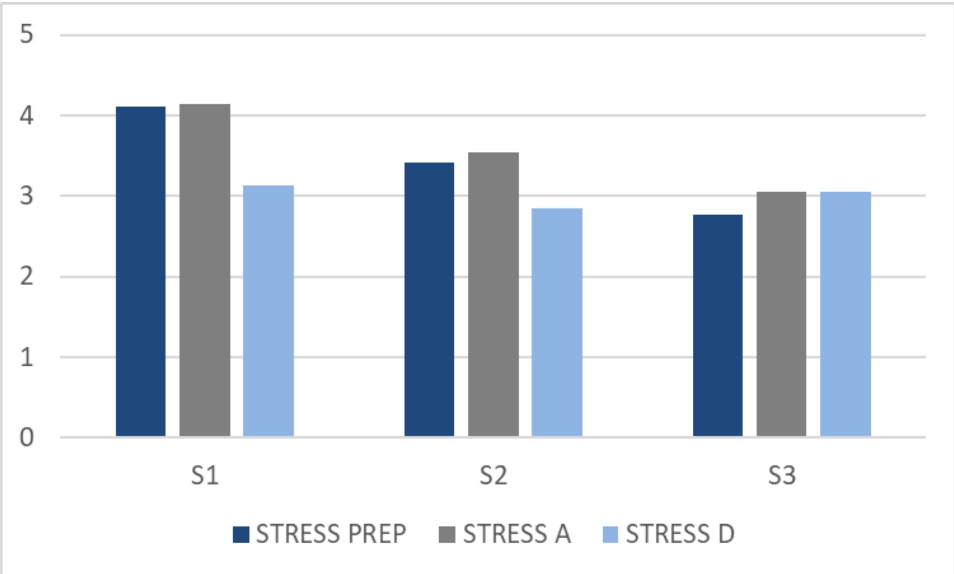


*Figure 48: Mean stress levels in scenarios S1, S2, and S3 divided by simulation phases (preparatory phase, phase A, and phase D)*

Stress in phase A is higher than in the other simulation phases in S1 and S2, while it is equal to stress in phase D in S3. This is explainable because of the greater complexity of the critical emergency case in S3, resulting in worse patient outcomes and, consequently, in higher stress in the re-assessment of the consciousness of the patient.

On the other hand, as it emerges in *Figure 49*, in all three scenarios, cognitive load is always higher in phase A.

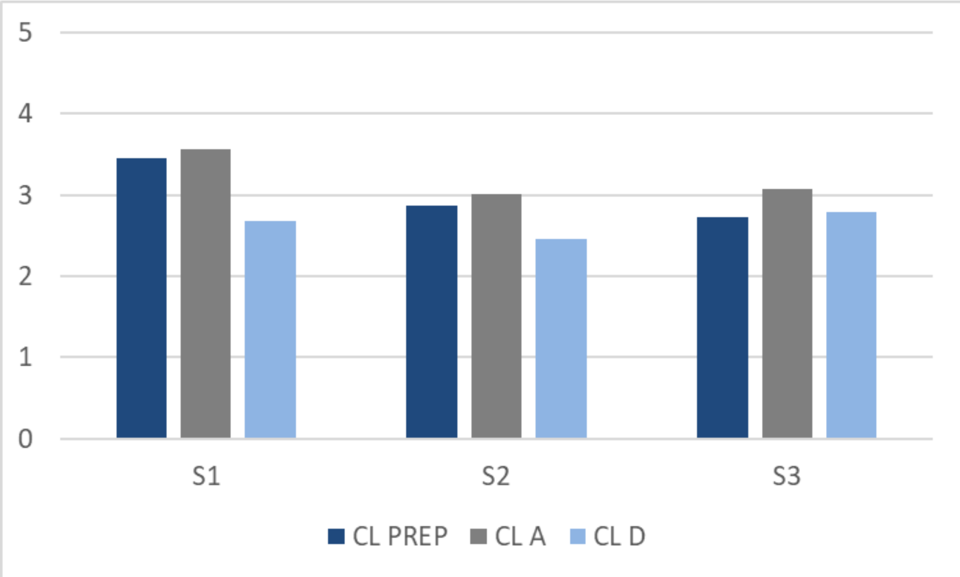


*Figure 49: Mean levels of cognitive load in scenarios S1, S2, and S3 divided by simulation phases (preparatory phase, phase A, and phase D)*

4.1.6.3. Intrinsic and Extraneous Cognitive Load

Signals analysis was integrated with video analysis to evaluate the events related to ICL and ECL. *Figure 50* shows the data analysis workflow for the discernment of the two different types of cognitive load.

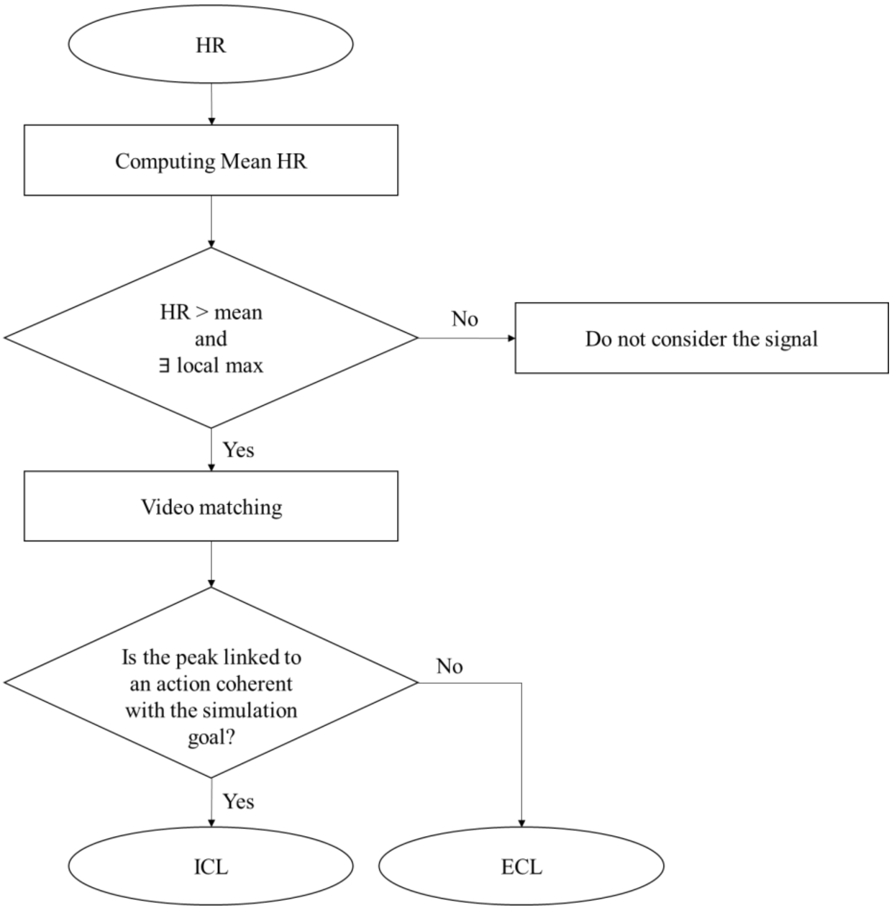


*Figure 50: ICL and ECL analysis workflow*

For each subject, the mean value of the HR signal was calculated, and HR peaks were counted every time the HR value was above the mean value. By matching signal analysis and simulation video recordings, only the peaks related to specific events were considered and classified as ICL or ECL.

*Table 12* shows the results of this analysis.

*Table 12: Events related to ICL and ECL*

| | |
|---|---|
| INTRINSIC COGNITIVE LOAD (ICL) | Emergency call reception |
| | Calls to/from other clinicians |
| | Ambulance arrival |
| | Patient arrival |
| | Oropharyngeal cavity inspection |
| | Cervical collar application |
| | Oxygen administration/Cannula application |
| | Send test tubes and request |
| | Pressure, ECG, pulse oximeter |
| | Breath sounds auscultation |
| | Fluids and/or blood administration |
| | Logroll for eco-fast, x-ray, vomit aspiration |
| | Cardiopulmonary resuscitation (CPR) |
| | Defibrillation |
| | Thermal blanket |
| | Decision making |
| EXTRANEOUS COGNITIVE LOAD (ECL) | Climb over a wire |
| | Problems with medical instrumentation |
| | No space for moving |
| | Exchange of roles |
| | Errors in calling or sending requests |
| | Calls/actions not performed by subject with sensors, but that he/she can listen/see |

The analysis of the events related to ECL is useful for the simulation optimisation. The room layout, the instrumentation distribution, the students' roles can be re-organised to minimise the ECL and increase the GCL.

Further analysis has been carried out to compare the percentage of ICL and ECL that occurred in groups that performed different scenarios (S1, S2, and S3). To complete an overall analysis about cognitive load, the percentage of CL (calculated through Phasya's algorithm), perceived mental demand (derived by NASA-TLX), and performance have been computed (*Figure 51*).

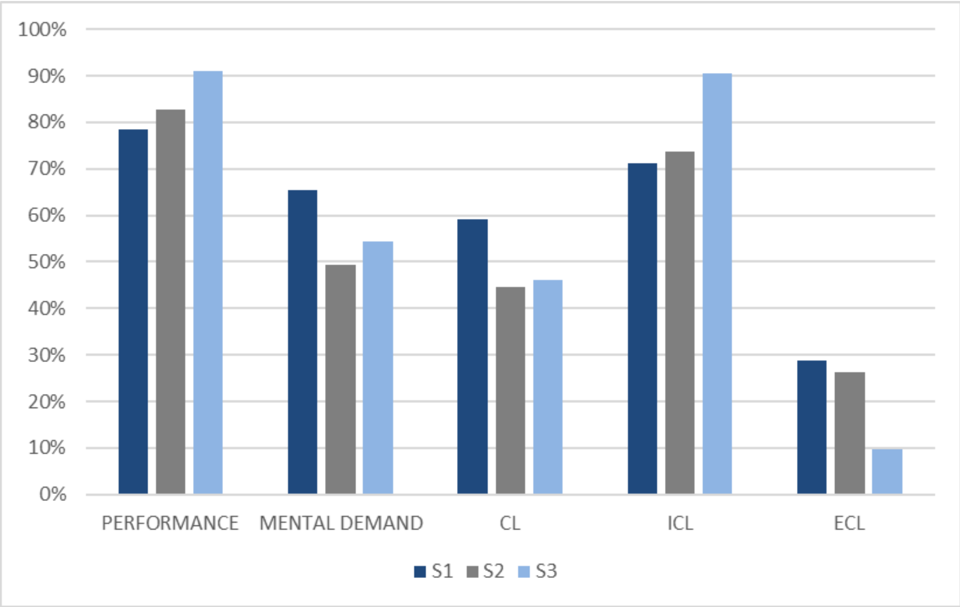


*Figure 51: Overall Cognitive Load analysis for scenarios S1, S2, and S3*

The results show good achievements in students' performances which are always over 75%. They increase from groups solving case S1 to groups assigned to case S3 with mean values of tasks correctly performed equal to 78.53% for S1, 82.83% for S2, and 91.02% for S3. This confirms again that attending the previous simulations and debriefings is very useful from a learning point of view.

As already seen for the physiological cognitive load, also the perceived mental demand considerably decreases from S1 (65.34%) to S2 (49.38%) and then increases for S3 (54.34%).

As expected, the number of ICL peaks is substantially higher than the ECL peaks. Moreover, ICL increases during simulation repetition (S1: 71.10%, S2: 73.57%, S3: 90.43%) while ECL decreases (S1: 28.90%, S2: 26.43%, S3:

9.57%). The increment of ICL and decrement of ECL can be explained by the sequential simulation scenarios: in fact, students can identify distracting elements from the previous simulations and then avoid them, focusing only on the important events to solve the clinical case (Brunzini, et al., 2020).

An outcome of a student-by-student analysis that deserves attention is related to the team working and leader/roles definition. Indeed, it has been demonstrated that the roles exchange during the simulation may cause an increment of CL and a decrement of CE (Kalakoski, et al., 2019).

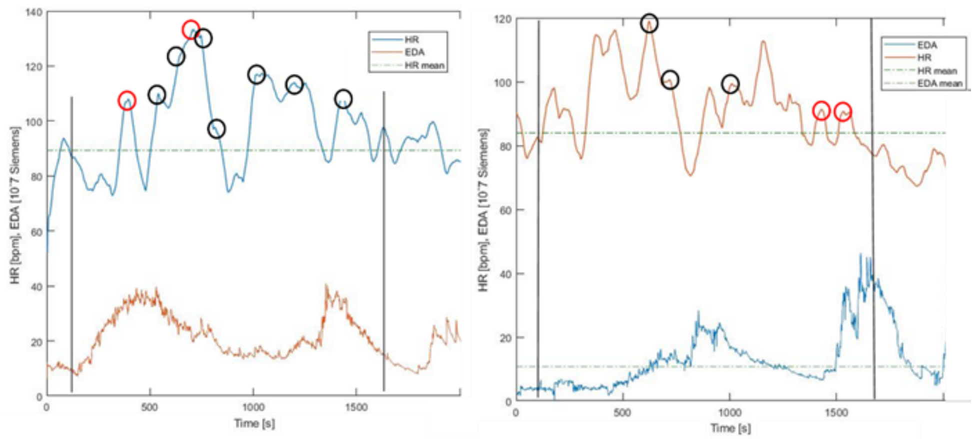

*Figure 52: ICL (black circles) and ECL (red circles) on HR and EDA trend: a subject with a team leader (left) and subject without a team leader (right)*

*Figure 52* shows a comparison between two students: the one on the left performed S3 with a leader, and the one on the right performed S1 without a leader. Although both subjects present 2 peaks related to ECL (red ones), the subject on the left, who performed a more difficult simulation and was more active than the other one, presents a greater number of HR peaks related to

ICL (black ones, 7 peaks vs 3 peaks ICL on the right). This is an indication of the fact that despite the difficulty of HFS, self-organization, team-working, and roles definition help the learners in performing a better simulation and acquiring technical and soft skills. Moreover, the leader can manage the tasks of other participants, resulting in a more effective resolution of the clinical case and better patient outcomes (Brunzini, et al., 2020).

#### 4.1.6.4. Additional considerations on physiological signals for stress detection

The values of salivary cortisol [ug/L] in *Figure 53* must be read with attention. Indeed, these are the mean values computed over the 144 subjects who correctly answered the NAS, without considering sex, menstrual cycle, the moment of the day (morning or afternoon).

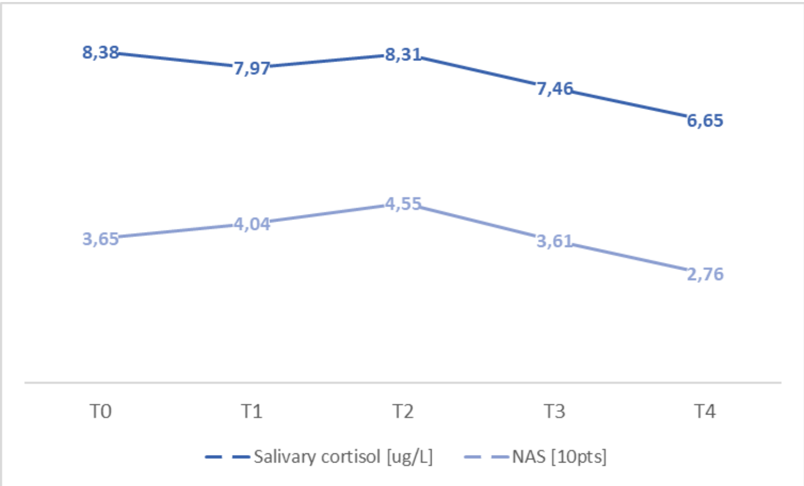


*Figure 53: Trend comparison between salivary cortisol and NAS, before and after the HFS*

The focus should be placed only on the trend along with the five acquisitions (at the arrival in the classroom (T0), ten minutes later (T1), and ten (T2), twenty (T3), and thirty (T4) minutes after the end of HFS).

Salivary cortisol was always collected together with NAS. From *Figure 53* it is possible to notice that cortisol and perceived stress have the same trend except for the first acquisition T0. It is interesting to see the stress peak after the end of the simulation (T2), the stress decrement during the debriefing (T3 and T4), and the effectiveness and benefit of debriefing that reduce the stress under the basal level (T4<T0 and T4<T1). It is worth to underline also the correspondence between physiological and perceived stress.

Finally, based on the literature review, the most important HRV parameters have been analysed, to verify their potential decrement in case of stress increment. The cardiac features considered in the time domain and non-linear domain were pNN50, pNN20, RR, SDRR, RMSSD, and D2.
Mean values of subjects who did not perform cardiac massage were analysed (68 subjects). Variations in these features (Δ), between debriefing and simulation, were computed. Results show a substantial increment of pNN20 and pNN50 for stress decrement from simulation to debriefing. While this increase is remarkable also for the RR signal, it is lower for RMSSD, SDRR, and D2.

*Table 13: Cardiac features variations related to stress*

| | Variation (Debriefing-Simulation) |
|---|---|
| Δ pNN50 | 6.94 |
| Δ pNN20 | 16.17 |
| Δ RMSSD | 0.01 |
| Δ SDRR | 0.01 |
| Δ RR | 0.13 |
| Δ D2 | 0.03 |

### *4.1.7. Statistical Analysis*

Statistical analysis was performed to understand the relationship among performance, stress, and cognitive load, and to overall evaluate cognitive ergonomics during high-fidelity simulations.

In particular, linear regression analysis was accomplished to comprehend which variables affect students' performance, perceived cognitive states, and physiological responses, during simulation training. Several models of single and multiple linear regression were computed to discover which factors are significantly related to each other. Indeed, linear regression analysis is used to estimate the relationships between the dependent variable (i.e. the main factor to understand and predict) and the independent variable(s) (i.e. the factor(s) which is(are) supposed to influence the dependent variable).

The linear regression analysis was performed employing the least-squares' method. A total of 53 variables related to 68 students was analysed (only participants with the complete performance, subjective, and physiological analysis were considered). Variables were related to demographic characteristics, performance, self-assessment questionnaires, physiological stress, and cognitive load detected through Phasya's algorithm (see Appendix C).

Concerning the surveys composed of multiple questions, a single total score was calculated, for both pre- and post-simulation responses. If some answers were missing, the total score was computed, for each questionnaire, as follow:

- Experience: the sum of the different experiences (i.e. group simulation, residency, professional training activity, training course, working experience, health volunteer activities, invasive procedure), for a total score ranging from 0 to 7;

- Aptitude to simulation and team-working: for each section of the survey, the sum of the scores was multiplied by the number of questions and divided by the number of answered questions. If for a participant, the answer rate was lower than 70%, that survey was discarded from the statistical analysis;
- STAI: for each section of the survey (State and Trait), the sum of the scores was multiplied by the number of questions and divided by the number of answered questions (Spielberger, et al., 1983). If for a participant, the number of answers was lower than 18, that survey was discarded from the statistical analysis;
- NASA-TLX: the total score was computed as expected by the questionnaire. The single items (mental, physical, temporal demands, performance, effort, and frustration) were weighted. All the students correctly completed the survey, thus there was no need to adjust the result.

*Table 14*, *Table 15*, and *Table 16* summarise the results of several single and multiple linear regression models, by showing the models and the variables which resulted significantly related to the dependent variable. P-values indicate statistical significance when they are lower than 0.05. All p-values were evaluated two-sided.

The following tables show the dependent variables, the p-value of the model (i.e. the significance F), and the significant independent variables. The p-value representation with asterisks indicates $^{*}p<.05$; $^{**}p<.01$; $^{***}p<.001$.

4.1.7.1. Performance

*Table 14* shows the results of the best linear regression models for the analysis of the variables that influence students' performance.

*Table 14: Linear regression analysis about performance (*p<.05; **p<.01; ***p<.001)*

| **Dependent Variable** | **Significance F** | **P-Value < 0,05** |
|---|---|---|
| TOTAL PERFORMANCE | 3,34E-11*** | DATE (-) ***<br>SESSION (+) *<br>GROUP (+) ***<br>SIM ATT PRE (+) *<br>NASA MENTAL (-) *<br>NASA PERF (+) ** |
| PERF PREP | 0,0004*** | DATE (-) ***<br>GROUP (+) **<br>SIM ATT PRE (+) *<br>NASA FRUSTR (-) * |
| PERF A | 0,033* | CL A (+) * |
| PERF B | 0,002** | DATE (-) *** |
| PERF C | 0,004** | DATE (-) *<br>GROUP (+) **<br>WEIGHT (-) * |
| PERF D | 0,009** | SESSION (+) *<br>GROUP (+) *** |
| PERF E | 7,42E-05*** | GROUP (+) ***<br>SIM ATT PRE (+) **<br>HEIGHT (+) * |

From *Table 14*, some interesting observations can be highlighted. First, performance is not influenced by stress. Indeed, neither perceived stress nor anxiety, neither physiological stress detected by Phasya's algorithm are statistically related to total or phase's performance. However, performance in Phase A seems to be statistically directly related to the physiological cognitive load (*p<.05*) measured in the same phase.

Total Performance, Performance of Preparatory Phase, Phase C, Phase D, and Phase E improve from scenario 1 to scenario 3. Indeed, in these cases, the "group" variable has a *p-value<.01*. Moreover, performance is higher in the morning (session) and the first days of simulation training (date).
Performance in the preparation phase directly depends on the attitude that students have toward group simulations.
It is also worth to see that the total performance is better for a lower perceived cognitive load.

4.1.7.2. Stress and Emotional State

*Table* ***15*** shows the results of the best linear regression models for the analysis of the variables that influence students' stress and emotional state.

*Table 15: Linear regression analysis about stress and emotional state (*p<.05; **p<.01; ***p<.001)*

| Dependent Variable | Significance F | P-Value < 0,05 |
|---|---|---|
| STAI TRAIT | 5,32E-18*** | STAI PRE (+) *** |
| Δ STAI | 0,0007*** | PERF A (+) **<br>PERF D (-) *<br>NASA MENTAL (-) *<br>NASA FRUSTR (+) *** |
| Δ NAS (2-1) | 2,72E-06*** | Δ STAI (+) ***<br>NASA-TLX (+) * |
| Δ NAS (4-1) | 6,39E-06*** | Δ STAI (+) *** |
| NASA PERFORMANCE | 0,001** | STAI TRAIT (-) **<br>PERF A (+) *<br>PERF B (-) *<br>PERF C (-) **<br>PERF D (+) * |

| | | |
|---|---|---|
| | | PERF E (+) * |
| NASA EFFORT | 0,048* | BIRTH (+) ** |
| NASA FRUSTRATION | 0,0002** | STAI TRAIT (+) **<br>Δ STAI (+) *<br>PERF PREP (-) *<br>PERF A (-) * |
| STRESS SIM | 0,009 ** | WEIGHT (-) **<br>HEIGHT (+) *<br>EXPERIENCE (-) * |
| STRESS D | 0,007 ** | WEIGHT (-) **<br>HEIGHT (+) **<br>EXPERIENCE (-) **<br>PERF D (-) ** |

The first line in *Table 15* shows that STAI Trait significantly depends on STAI Pre. This means that the students' answers about their anxiety in daily life, are strongly influenced by the anxiety perceived at that moment. For the same reason, variations in Δ STAI, Δ NAS (2-1), and Δ NAS (4-1) are mutually related to one another. The perception and discernment among similar emotional states such as anxiety or stress are not easy to subjectively evaluate. However, it can be observed that higher anxiety is associated with better performance in Phase A ($p<.01$) and worse performance in Phase D ($p<.05$). Moreover, the anxiety variation before and after the simulation depends inversely on the mental demand ($p<.05$) and directly on the sense of frustration ($p<.001$).

Increment of stress from Δ NAS (2-1) is not related to performance but is related to higher perceived workload ($p<.05$). However, it is not related to the mental demand itself.

Frustration is statistically directly related to performance in the preparation phase and Phase A ($p<.05$). Frustration is also higher in anxious subjects.

Physiological stress measured during the simulation is statistically inversely related to the previous experience of the student. Thus, subjects without previous experiences felt more stressed.

Concerning the stress in different simulation phases, it results that only in Phase D the mean stress is inversely related to performance (in Preparation Phase and Phase A, stress is not related to performance). Thus, worse performance in Phase D induces greater stress in students. This could be explained by the fact that the re-assessment of the consciousness of the patient and the resolution of the clinical case are strong stressful events.

#### 4.1.7.3. Cognitive Load

*Table 16* shows the results of the best linear regression models for the analysis of the variables that influence students' cognitive load.

*Table 16: Linear regression analysis about cognitive load (*p<.05; **p<.01; ***p<.001)*

| Dependent Variable | Significance F | P-Value < 0,05 |
|---|---|---|
| NASA MENTAL | 9,05E-05*** | STAI TRAIT (-) *<br>Δ STAI (-) **<br>Δ NAS (2-1) (+) **<br>PERF A (+) ** |
| CL SIM | 0,028 * | DATE (-) *<br>PERF A (+) * |
| CL A | 0,028 * | HEIGHT (+) *<br>PERF A (+) ** |
| CL D | 0,011 * | WEIGHT (-) **<br>HEIGHT (+) **<br>SIM ATT PRE (+) *<br>STAI PRE (-) **<br>NAS 1 (+) **<br>PERF D (-) * |

Concerning the perceived cognitive load, results in *Table 16* show that the mental demand is statistically directly related to performance in Phase A (*p<.01*). Moreover, higher perceived mental demand is also directly influenced by stress and inversely related to anxiety. Thus, subjects who feel not anxious but very stressed, have the perception of higher mental effort.

Increment of physiological cognitive load is related to better performance in Phase A (*p<.05*). Physiological cognitive load was higher in the first simulation sessions.

Concerning the CL measured in different simulation phases, it is directly related to performance in Phase A, and inversely related to performance in Phase D (as for the stress). Mean CL in Phase D is also directly related to the student's attitude to simulation and team-working. This means that the cognitive load is higher when subjects have a greater attitude toward group simulation.

#### 4.1.7.4. Additional analysis for cortisol and cardiac features

Linear regression analysis was done also to understand which variables may affect the salivary cortisol. Several multiple linear regression models were computed but only the ones with physiological stress and CL as independent variables were statistically significant. Indeed, cortisol 2, fold increase, and AUCi result directly related only to physiological stress and CL (*p<.01*) measured during the simulation. AUCi also depends on the physiological stress and CL collected during the debriefing (*p<.05*). All the other variables (such as performance, self-assessment measures) do not influence salivary cortisol.

Concerning the statistical analysis about the cardiac features useful for the detection of physiological stress and cognitive load, linear regression models revealed that the HRV pNN20 feature is the only one time-domain feature significant for the recognition of stress level ($p<.05$). pNN20 is highly significant also for the recognition of cognitive load ($p<.001$). Other significant time-domain cardiac features strictly related to cognitive load are pNN50, SDRR, and RMSSD ($p<.01$).

It must be noted that stress and cognitive load are always inversely related to HRV features. This is in line with the fact that in stress conditions and for high cognitive load, the HRV features are expected to decrease.

### *4.1.8. Main Findings HFS: Summary*

In this paragraph, a summary of the main findings is reported.

#### 4.1.8.1. Performance

Some highlights related to performance are hereunder listed:

- Performace decreases toward the end of the simulation, during the patient's consciousness assessment and exposure. More theory and practice should be provided about these aspects. However, goods results are achieved, with a mean total score around 85/100 points.
- Performance generally improves from scenario 1 to scenario 3, confirming that observing previous scenarios is extremely useful.
- Performance is higher in the morning and the first days of simulation training.

- Performance is not influenced by stress, nor subjective, neither physiological.
- Performance in Phase A (airway assessment) is higher with higher physiological cognitive load measured in the same phase. Nevertheless, total performance is higher for a lower perceived cognitive load. This data highlights the difference between perceived and measured cognitive load, between what the learners feels/believes and what he/she experiences. Therefore, the importance of measuring physiological parameters to assess stress and CL, that can be considered objective (against the subjective self-assessment), becomes evident.

#### 4.1.8.2. Stress and Emotional State

Some highlights related to Stress and Emotional State are hereunder listed:

- Students are generally more anxious in their life (STAI Trait) than during the lesson (STAI State). As expected, their feeling of anxiety decreases after the simulation.
- Perceived stress (from NAS) follows a bell-shaped trend, with a peak just at the end of the simulation. The effectiveness of the debriefing is confirmed by the fact that, at the end of it, participants feel less stressed than when they arrived in the classroom.
- Variations in Δ STAI, Δ NAS (2-1), and Δ NAS (4-1) are mutually related to one another. The perception and discernment among similar emotional states such as anxiety or stress are not easy to subjectively evaluate. For

this reason, it is important to include physiological, objective measurements in the analysis of cognitive and emotional conditions.

- Physiological signal analysis shows that, on average, from simulation to debriefing the stress decreases and the decrement in stress is higher than the decrement in CL, thus confirming the effectiveness of the debriefing.
- On average, during the simulation, stress is higher than cognitive load, and during the debriefing, cognitive load is higher than stress.
- The average stress in phase A is higher than the average stress in the rest of the simulation.
- Students feel greater stress when their performance is worse in Phase D. This could be explained by the fact that the re-assessment of the consciousness of the patient and the resolution of the clinical case are strong stressful events.
- Stress is maintained on average on a medium level for the entire simulation.
- Stress gradually decreases from scenario 1 to scenario 3 during both the simulation and the debriefing, confirming the usefulness of watching the previous scenarios and debriefings.
- Subjects without previous experiences felt more stressed (from physiological measurements).

It is worth to note that, contrary to the self-assessment measures, the analysis of stress from the physiological parameters allows distinguishing stress between simulation and debriefing, and among simulation phases.

4.1.8.3. Cognitive Load

Some highlights related to Cognitive Load are hereunder listed:

- NASA-TLX after high-fidelity simulations is on average high (equal to 65.77). Students perceive the simulation as high demanding activity with the greatest weight given to the mental demand.
- Physiologically measured cognitive load is quite constant and high during the simulation and varies along with the debriefing, mostly according to the student's interaction with the teacher, thus confirming the correctness of the algorithm used to discriminate the cognitive load by the physiological signals.
- On average, cognitive load decreases from simulation to debriefing (but less than the stress).
- On average, cognitive load is lower than stress during the simulation, and higher than stress during the debriefing.
- Cognitive load is higher when subjects have a greater attitude toward group simulation.
- Physiologically measured cognitive load in phase A is higher than in the rest of the simulation.
- Mental demand and physiological cognitive load are statistically directly related to performance in Phase A.
- CL level is maintained on average on a medium level for the entire simulation.
- Physiological CL and perceived mental demand decrease from scenario 1 to scenario 2 and then increases from scenario 2 to scenario 3. This may

be because, even if students performing scenario 3 had already seen scenario 1 and scenario 2, the difficulty and complexity of the clinical case did not allow a further reduction of the mental demand.

- The simulation repetition in three different scenarios allows the increment of ICL and decrement of ECL, from scenario 1 to scenario 3. The analysis of ICL and ECL is useful for the simulation optimisation. The room layout, the instrumentation distribution, the students' roles should be re-organised to minimise the ECL and increase the GCL.
- Despite the difficulty of HFS, self-organization, team-working, and roles definition help the learners in performing a better simulation and acquiring technical and soft skills. Moreover, the leader can manage the tasks of other participants, resulting in a more effective resolution of the clinical case and better patient outcomes (Brunzini, et al., 2020).

Even in this case, it is worth to note that the analysis of cognitive load from the physiological parameters allows distinguishing CL between simulation and debriefing, and among simulation phases. It allows also discerning between Intrinsic and Extraneous Cognitive Load. It would not be possible using only the self-assessment measures.

#### 4.1.8.4. Additional Considerations

Some additional considerations are hereunder listed:

- The perception of the simulation's usefulness in dealing with situations that can cause anxiety and fear increases after the simulation. Therefore, during the HFS, students felt realistic emotional involvement, related both to the patient and to the emergency and critical situation.

- Conversely, after the HFS, students' opinion about simulation benefit toward the development of decision-making skills gets worse. This issue should be solved during the debriefing.
- After the HFS, students strengthened their belief in the relevance of asking for help, summarising patient test results, establishing "who-does-what", and, above all, paraphrasing and repeating the instruction received.
- Conversely, after the HFS, a slight worsening concerns the students' points of view related to the relevance of inter-professional education. Also this issue should be solved during the debriefing.
- Concerning the salivary cortisol analysis, it is possible to notice the correspondence in trend between cortisol, physiological, and perceived stress, confirming the correctness of the algorithm used for the analysis of physiological parameters for the stress assessment.
- Linear regression analysis underlined that only physiological measurements statistically affect the salivary cortisol. All the other variables (such as performance, self-assessment measures) do not influence salivary cortisol.
- Concerning the most important HRV parameters, the analysis shows a substantial increment of pNN20 and pNN50 for stress decrement from simulation to debriefing. While this increase is remarkable also for the RR signal, it is lower for RMSSD, SDRR, and D2.
- Linear regression models revealed that the HRV pNN20 feature is the only one time-domain feature significant for the recognition of stress level

($p<.05$). pNN20 is highly significant also for the recognition of cognitive load ($p<.001$). Other significant time-domain cardiac features strictly related to cognitive load are pNN50, SDRR, and RMSSD ($p<.01$). It must be noted that stress and cognitive load are always inversely related to HRV features.

These findings validate the algorithm used for stress and CL detection from physiological signals. For this reason, in order to simplify the procedure, the salivary cortisol analysis can be avoided in the next simulation training assessments.

After a deepened analysis of physiological features to be collected for the stress and CL detection, a unique wearable device could be chosen for the assessment of training in practice. Moreover, the surveys to be administered could be selected based on the kind of analysis to be accomplished (effectiveness assessment, cognitive load assessment, stress assessment, …).

## 4.2. Low-Fidelity Simulation

### *4.2.1. Rachicentesis Simulation in Low-Fidelity*

The low-fidelity simulation (LFS) is usually used for the training and enhancement of practical, technical skills. The one considered in this study concerns the training of the rachicentesis with the lumbar puncture trainer by Gaumard®. Rachicentesis is a lumbar puncture type whit the aim of taking a CFS sample to be analysed for diagnostic purposes. It may also have the therapeutic purpose of draining any liquor excess. Thus, rachicentesis is a surgical technique that involves inserting a thin needle (usually 22 gauge, 75mm long) into the space between the arachnoid meninx and the pia mater (i.e. in the subarachnoid space that contains the cerebrospinal fluid). This insertion occurs between the third and fourth lumbar vertebrae or more commonly between the fourth and fifth. To find this space more easily, it is recommended to draw an imaginary line between the 2 iliac crests.
For the execution of this procedure, the patient can be positioned in the lateral foetal position (with the trunk and the head flexed to allow the space between the vertebrae to be more opened) or sitting on the bed with the back arched forward. Rachicentesis requires maximum asepsis because, with this manoeuvre, the internal space of the central nervous system is put in communication with the environment. Thus, the operator can proceed with the needle insertion (into the subarachnoid space, crossing the skin, subcutis, muscles, flavum ligament, and dura mater), only after having identified the landmarks and ensured disinfection.

The simulated procedure can be divided into three consecutive phases:

- The preparatory phase: involves the reception and positioning of the patient, the palpation of anatomical landmarks, the disinfection, and the correct grabbing of the needle;
- The puncture phase: involves threading the needle in the correct point, extracting the needle stylet, and waiting for the liquor spilling out;
- The ‘end of procedure’ phase where the learner takes the test tube and collects the CFS, re-puts the stylet into the needle, removes the needle, and tampons the skin.

The goal of the simulation is to learn how to perform rachicentesis and become able to let the liquor spilling out. Indeed, if the needle is non inserted in the right place and with the right depth, the CFS would not come out.

Each student performed the simulation once. The simulation duration varied according to the student’s skill. Sometimes it lasted a few minutes, sometimes it took even around half an hour (mean duration of 6.3 ± 4.8 minutes). Rachicentesis simulation was performed singularly by each student, with the other participants observing it.

The room layout included a desk with the skill trainer and, on the right, all the instrumentation useful and needed for the practice.

The teacher and the other students were in the same room with the subject performing the simulated rachicentesis. Thus, even in this case, the room layout was not optimal, since students performing the simulation, students watching the simulation, and the teacher should not be in the same room. Indeed, their presence may influence the performance of the student who is practicing the rachicentesis, and also his/her feeling of stress and pressure.

In this case study, the standard simulation flowchart was not followed. Indeed, LFS was preceded by the teacher briefing, but it was not followed by the debriefing.

Low-fidelity simulations were recorded through a video camera. A video frame is shown in *Figure 54*.

*Figure 54: A student, equipped with wearable sensors, performing the low-fidelity simulation of rachicentesis*

### *4.2.2. Data Acquisition Workflow*

The same 182 students of the 6th year of Medicine and Surgery Course were asked to participate in this other case study. As for the HFS, the study was carefully explained, including implications and eventual contraindications in the use of smart devices. Then, the forms for the informed consent and the consent for the processing of personal data were administered to the students. The same 148 students accepted to be involved in the study. Also, in this case, students who did not sign the informed consent were excluded from the study but not from the simulation.

Students were divided into 26 groups of 7 subjects each. For each simulation session, two groups were consecutively convened, one group at a time. Simulation sessions lasted about four hours and a half, from 9:00 a.m. to 1:30 p.m. in the morning and from 2:30 p.m. to 7:00 p.m. in the afternoon (about two hours per group).

At the arrival of each group in the simulation room, the workflow in *Figure 55* was followed.

As expected by the proposed procedure for the assessment of stress and cognitive load, at the arrival in the classroom, students answered the NAS scale to record the basal level of perceived stress. Then, the survey session about aptitude to technology and its relevance in the simulation context was administered. This time, the STAI questionnaire was given to the students to assess only their perceived level of anxiety in that precise moment (STAI State). For the STAI Trait, the answer previously collected were considered, as they should not change in a few days. Ten minutes after the first sample, another NAS score was collected (NAS 1). It is worth to notice that, in this case, the salivary cortisol was not sampled. Indeed, it was necessary as a gold standard in the first case (HFS) to validate the stress assessment. Indeed, the HFS was more appropriate for the stress assessment validation. The low-fidelity simulation is less stressful, and the debriefing is not applied. For these reasons, salivary cortisol was not collected here.

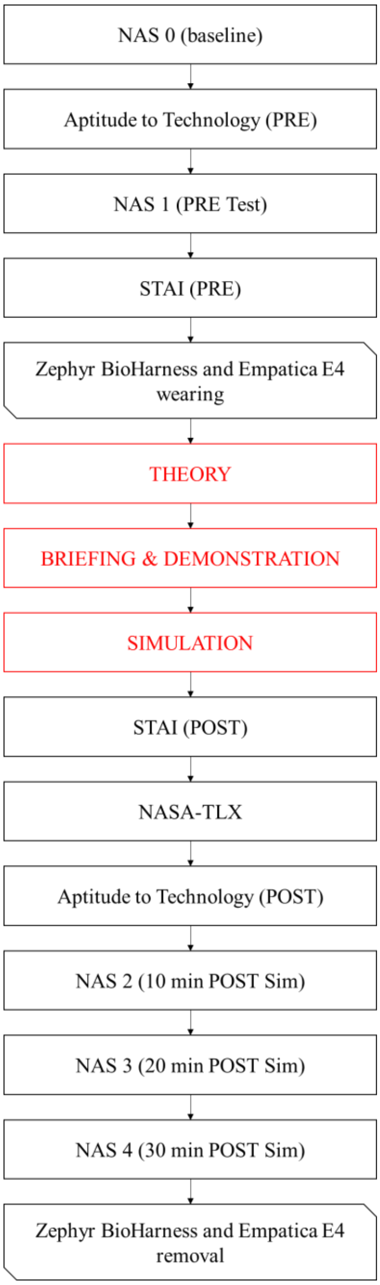


*Figure 55: Workflow for the analysis of stress and cognitive load during LFS*

After the self-assessment pre-simulation phase, participants were equipped with wearable devices for the collection of physiological parameters. Their biometric parameters were recorded during the theory explanation (rest), during the briefing and teacher demonstration, and obviously during their simulation activity. At the end of the demonstration, a random sequence to perform the simulation was established. To avoid eventual influences on students' anxiety, stress, and cognitive states, the sequence was not revealed. Students were called one by one to practice, while the other ones were observing.

After each rachicentesis simulation, the STAI State, the NASA-TLX, and the survey about the relevance and aptitude to technology were administered. Perceived stress was assessed through NAS, by administering it to each participant 10, 20, and 30 minutes after the end of the simulation.

In this case, the performance was assessed for each student using a checklist to evaluate each task as 'correctly performed', 'non correctly performed', or 'not performed'. Times (i.e. patient preparation time, time to succeed, total time), number of errors (i.e. touching the needle in the wrong place, do not re-insert the stylet, coming in-and-out with the needle in the skill trainer), number of attempts to succeed, and teacher interference were also recorded and assessed.

A skill questionnaire was administered before and after the simulation.

### *4.2.3. Participants*

The 148 students involved in this case study were the same who performed the high-fidelity simulation. Thus, their characteristics have been previously reported in *Table 7*, with the only difference that, in this case, the percentage of students developing a thesis pertinent with the LFS was equal to 54.05% (while 21.62% did not answer). In this case, the courses considered pertinent to LFS were: anaesthesia and intensive care, cardiology, every surgery, haematology, endocrinology, emergency, gastroenterology, geriatrics, paediatrics, obstetrics and gynaecology, infectious diseases, internal medicine, nephrology, orthopaedics, pneumology, rheumatology, radiology, and urology. It is also worth knowing that 27.70% of them have already had experience in practicing invasive procedures.

*Table 17* shows the mean subjective students' perception of stress and anxiety in their life (STAI Trait) and before executing the simulation (STAI State PRE, NAS 0, and NAS 1). In comparison with the subjective psychological state before HFS, in this case, students felt more relaxed (STAI State of 41.07 against 46.14, NAS 0 and NAS 1 of 2.7 respectively against 3.7 and 4.0). This was quite foreseeable since the high-fidelity simulation is more stressful and induces greater performance anxiety, respect to the low-fidelity simulation.

*Table 17: Subjective psychological state of participants before LFS*

| | MEAN | STANDARD DEV. |
|---|---|---|
| STAI TRAIT | 48.06 | 9.95 |
| STAI STATE PRE | 41.07 | 10.92 |
| NAS 0 | 2.73 | 2.14 |
| NAS 1 | 2.71 | 1.92 |

*Figure 56* and *Figure 57* show the results of the survey about aptitude to the use of technology, described in *Table 3*, and administered before the

LFS. As already explained, it is divided into two sections regarding the personal use of technological devices and the relevance of technology in the simulation context.

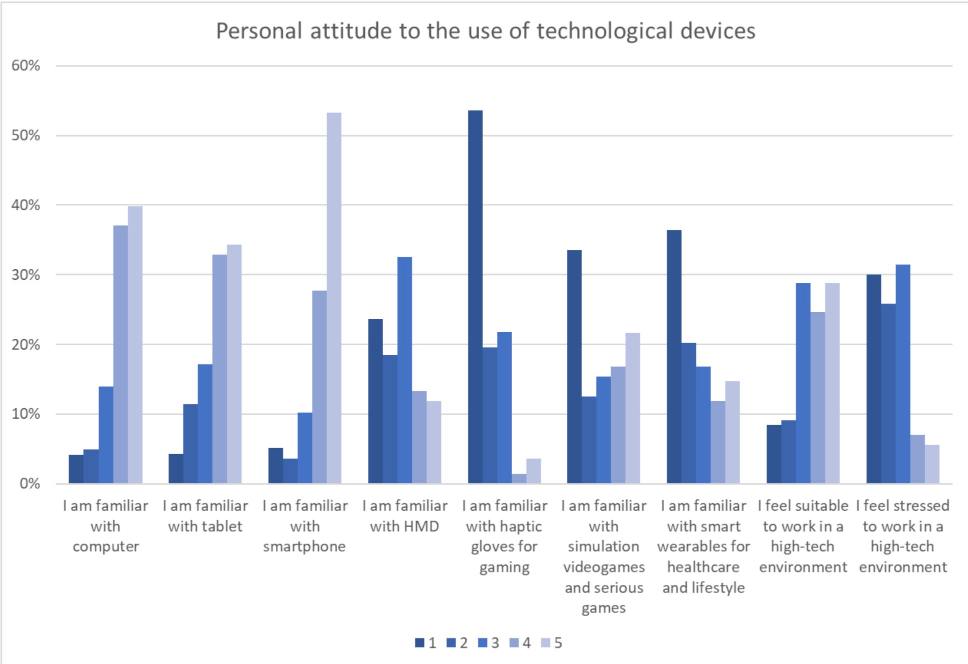

*Figure 56: Results of the survey about aptitude to technology usage (Section: Personal attitude to the use of technological devices)*

From *Figure 56*, it emerges that most students are very familiar with the use of computers, tablets, and smartphones. Conversely, only 25.19% are familiar with the use of head-mounted displays, and the percentage falls to 5.07% for students who use haptic gloves for gaming. However, 38.46% are used to play with simulation videogames and/or serious games.

Concerning the smart wearables for monitoring healthcare and lifestyle, only 26.57% of students declare to use them.

In conclusion, even if 12.59% of students think that would feel stressed to work in a high-tech environment, 53.52% of them would appreciate working in it.

Indeed, most students (over 62%) think that technological devices (such as virtual reality glasses, gloves with haptic feedback, etc.) are a valuable tool for learning during training. Moreover, over 70% of them state that multisensory interaction (tactile, visual, and auditory), and feedback provided through technological devices promote learning in the simulation (*Figure 57*). Students assert also that a high degree of immersion during the simulation has a positive effect on learning (81.12%) and psychological component (76.76%).

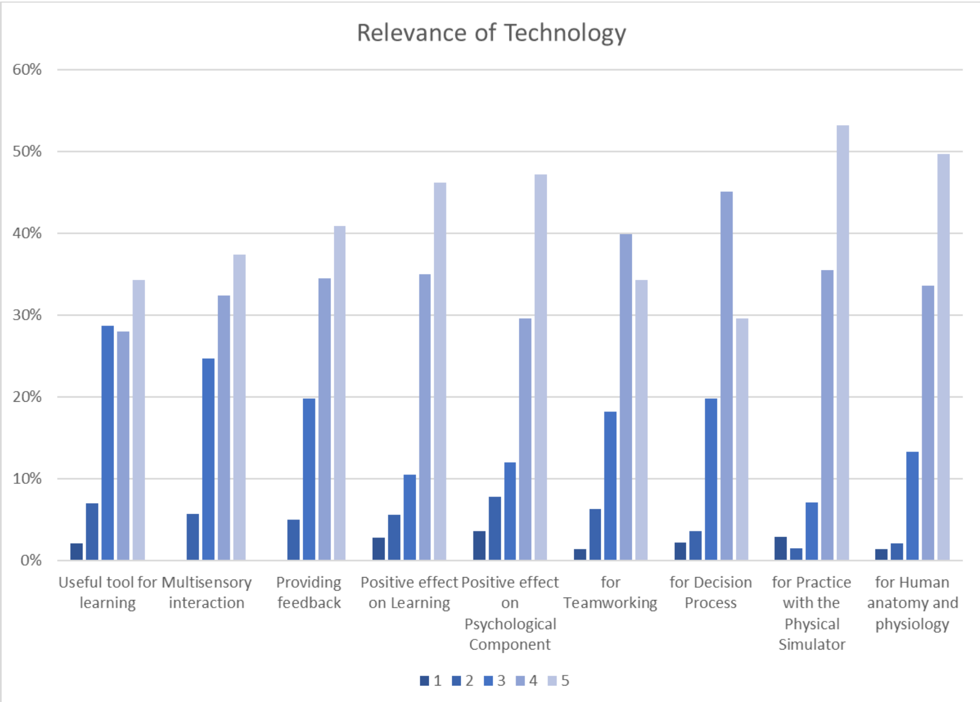


*Figure 57: Results of the survey about aptitude to technology usage (Section: Relevance of technology in the simulation context)*

In conclusion, most participants would like to be supported, during simulations, by high-tech devices for team working, decision process, practice with the physical simulator, and understanding of human anatomy, physiological and pathological processes. However, among these items, the support with high-tech devices is mainly desired (88.65%) for the practice with physical manikins, to improve the low-fidelity simulations.

### *4.2.4. Performance Analysis*

The performance was assessed for all the 148 students enrolled in the study. 90.85% of them reached success, i.e. they were able to correctly perform the puncture and make the CFS pouring out.

The mean number of attempts to collect the liquor was equal to 3.61 (±3.98), while the mean number of errors committed by each student, during the rachicentesis procedure, was equal to 0.54 (±0.65). Indeed, as shown in *Figure 58*, the percentage of errors for the total sample of students is quite low. The 21.13% touched the needle where it was not allowed (for hygienic reasons), and none of them extracted the needle without re-inserting the stylet. However, 32.39% of students completely came in-and-out with the needle during the procedure (always for hygienic reasons, the needle should be extracted only once at the end of the procedure; even if it has been inserted in the wrong position, the learner should try to improve the needle incline without extracting it).

*Figure 59* shows the percentage of correctly/incorrectly/not performed tasks, calculated over the 148 participants, and divided between the preparation phase, and tasks after the liquor pouring out (end).

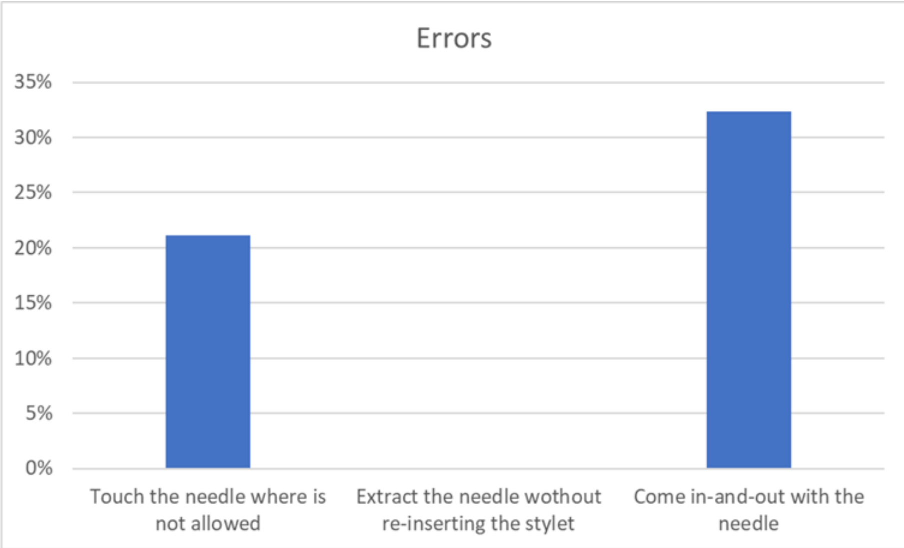


*Figure 58: Percentage of committed errors during LFS*

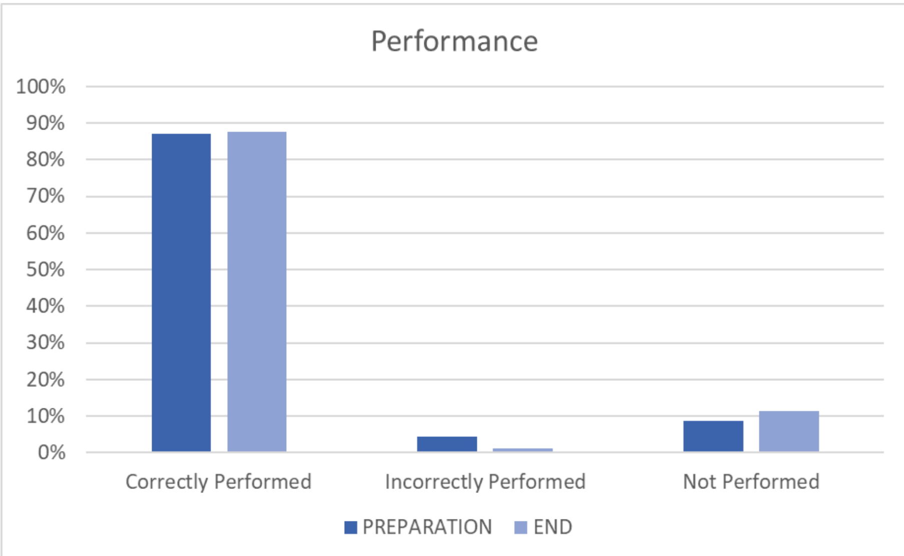


*Figure 59: LFS students' performance per phases*

In both phases, most students (more than 87%) correctly performed the tasks. The percentage of students who incorrectly executed the tasks is very low (under 4%), while the percentage of participants who did not perform some tasks is equal to 8.69% for the preparation phase and 11.27% for the end of the simulation. Thus, students understand how to execute the tasks but sometimes they forget some steps.

For the 90.85% of students who reached success, mean execution times were computed:

- Preparation Time: 1,74 ± 0,81 minutes
- Time to success: 3,30 ± 4,50 minutes
- Time to finish: 1,54 ±3,04 minutes
- Simulation Total Time: 6,26 ±4,81 minutes

Moreover, in this case study, students' skills were assessed also using a questionnaire composed of five queries. It was administered before and after the simulation. Only 58.11% of students entirely answered the questionnaire before and after the LFS. Considering only these students' percentage, while the mean number of correct answers before the simulation was equal to 3.73 ± 0.95, after the simulation it increased to 4.71 ± 0,51.

#### *4.2.5. Psychometric Analysis*

In this paragraph, results related to subjective questionnaires for the assessment of stress, cognitive load, and, more in general, perceived involvement in the simulation, are illustrated.

4.2.5.1. The relevance of technology in the simulation context

*Figure 60* shows the comparison of the relevance of technology in the simulation context, between students' opinion before (PRE) and after (POST) having taken part in the LFS. 132 participants answered both the pre- and post-simulation survey; thus, results in *Figure 60* represent the mean values of each item (on the 5-points Likert scale), computed over the answers of 89.19% of participants.

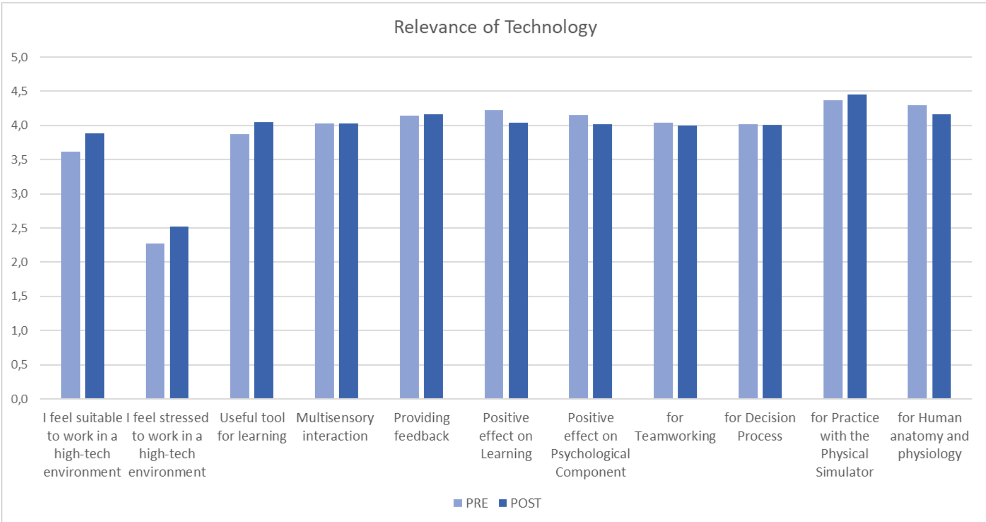


*Figure 60: Relevance of technology in the simulation context: comparison between opinions collected before and after the LFS*

Students' point of view remains approximately the same before and after the simulation. This may be attributed to the fact that, during the LFS, participants did not use high-tech devices such as head-mounted displays or haptic gloves, and, consequently, their multisensorial interaction is not improved.

The main difference between pre- and post- simulation regards the way the students feel in relation to a high-tech environment. Indeed, after having executed the low-fidelity simulation (without the use of AR but wearing the smart devices) they feel more suitable to work in a high-tech environment, but, at the same time, they also perceived greater stress.

However, regarding the kind of training they would like to be supported by high-tech devices, they confirm the practice with physical manikin as the simulation that would benefit the most by the use of haptic gloves, head-mounted displays, and applications in extended reality.

#### 4.2.5.2. State Trait Anxiety Inventory (STAI)

Excluding participants who did not answer the STAI questionnaire before or after the simulation, results in *Table 18* show the mean values (20 questions on a 4-points Likert scale for each session, for a maximum possible total score of 80) calculated on 137 responses (92.57% of participants).

*Table 18: STAI results before and after LFS*

| | MEAN | STANDARD DEV. |
|---|---|---|
| STAI TRAIT | 47.72 | 9.88 |
| STAI STATE PRE | 41.36 | 10.98 |
| STAI STATE POST | 38.95 | 10.63 |

Even in this case, as forecast, perceived anxiety decreases after the simulation (mean STAI State Post is lower than mean STAI State Pre). The mean STAI State Post agrees with the value expected for college students (Spielberger, et al., 1983), while the mean STAI State Pre is few points higher. Thus, the STAI State Pre confirms the actual increment of anxiety perceived by the students before performing the simulation.

4.2.5.3. Numeric Analog Scale (NAS)

Results in *Figure 61* are based on 121 students' (81.76%) answers. Students who omitted one or more responses were excluded from the analysis.

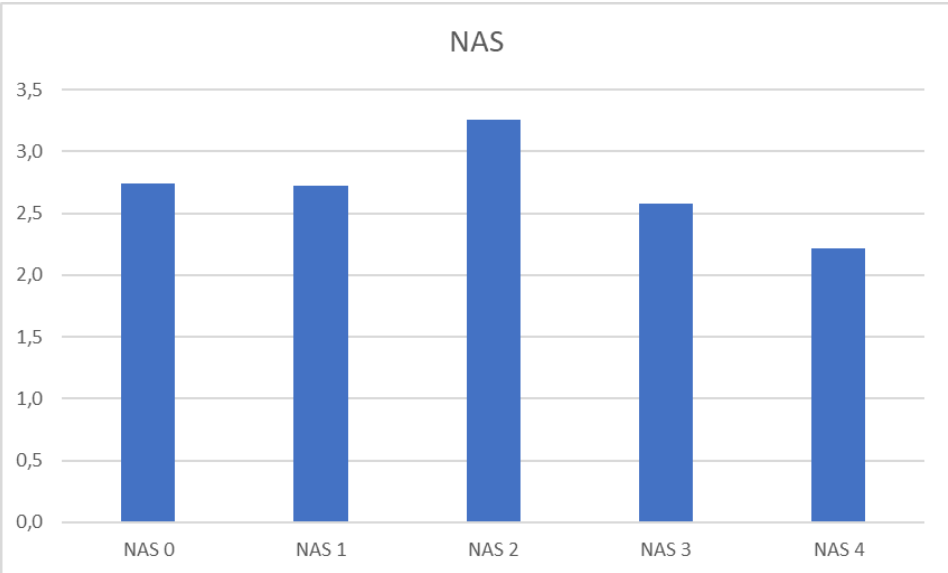

*Figure 61: Mean NAS results (10-points Likert scale) for LFS*

Even in this case, mean NAS results confirm the expectations. The peak in NAS 2 (10 minutes after the end of the simulation) refers to the stress due to the execution of the simulated rachicentesis. Even if, in this case study, the debriefing is not provided, the perceived level of stress after the simulation decreases below the one perceived at the arrival in the simulation room (NAS 3 and NAS 4 are lower than NAS 0 and NAS 1). Indeed, the low fidelity simulation is not as emotionally engaging as the high-fidelity one. For this reason, the maximum perceived stress is low (mean of 3.3 on a 10-points scale, against the mean NAS 2 equal to 4.5 in HFS), and the debriefing is not always necessary (mean NAS 4 in LFS without the debriefing is lower than mean NAS 4 in HFS with debriefing).

#### 4.2.5.4. NASA Task Load Index (NASA-TLX)

*Table 19* shows the mean values of the total score for the perceived workload related to LFS, and the weight assigned to the six indexes. The response rate of the NASA-TLX survey is equal to 95.95%.

*Table 19: NASA-TLX results for LFS*

| | MEAN | STD. DEV. |
|---|---|---|
| NASA-TLX | 52.51 | 18.75 |
| Mental Demand | 175.14 | 122.74 |
| Physical Demand | 65.65 | 91.21 |
| Temporal Demand | 73.37 | 80.72 |
| Performance | 237.79 | 125.86 |
| Effort | 135.67 | 115.03 |
| Frustration | 99.99 | 133.18 |

According to the NASA-TLX score interpretation of (Sugarindra, et al., 2017), the perceived workload after the low fidelity simulation is on average high (high for 50-79 total score). In contrast to the results related to high-fidelity simulation, in this case the greatest weight is assumed by the performance (in HFS it was assumed by the mental demand). Thus, the importance of resulting in a correct and good performance has a great impact on the perceived workload. However, even the mental demand and the effort gain high values. The perceived physical demand is higher with respect to the one perceived in the HFS. Conversely, the frustration is lower than in the HFS.

#### *4.2.6. Biometric Analysis*

In this paragraph, levels of stress and cognitive load, detected through Phasya's algorithm, are discussed.

Physiological signals were collected for the entire sample of 148 students. However, due to missing and erroneous data, results refer to 107 students (72,3%).

Stress and CL levels were computed for simulation phases. Levels range between 0 and 6.

The mean stress level calculated over the entire simulation is equal to 2.94 (±1.91), while the mean CL level is 2.82 (±1.34), resulting both lower than the mean levels in high-fidelity simulation.

In this case, the stress is on average higher than the cognitive load involved in the execution of the simulated rachicentesis. However, stress becomes lower than CL toward the end of the simulation.

From *Figure 62*, it is possible to note that the peak levels of stress and CL are during the preparation phase. Then, while stress constantly decreases from phase to phase, CL remains approximately constant (with a light decrement from the preparation phase to the end of the simulation). The stress and CL levels during the rest period after the simulation, are lower than the ones calculated in the rest period before the simulation. This confirms the lack of debriefing necessity. Moreover, stress and CL during the execution of the procedure can be considered medium, not high.

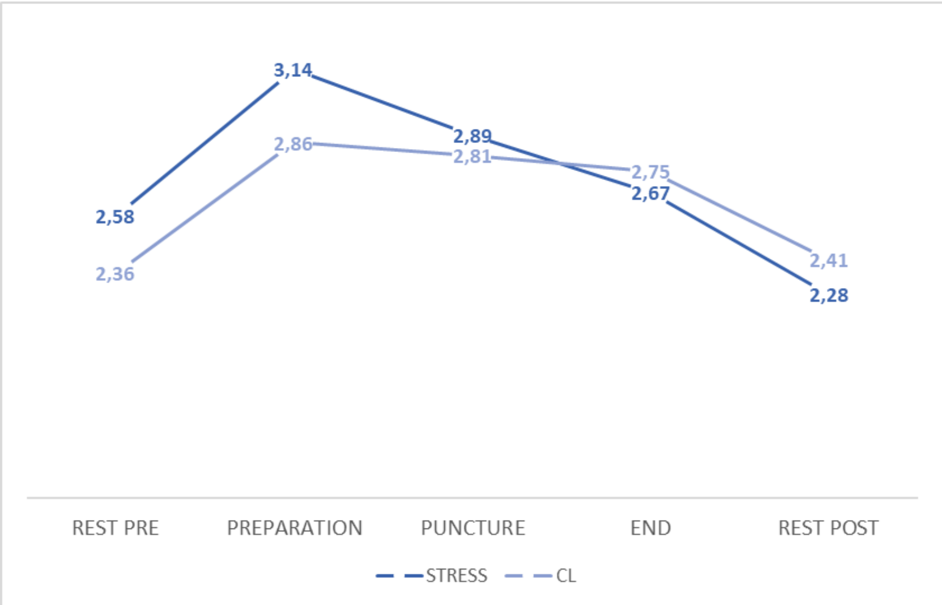


*Figure 62: Stress and CL during LFS phases*

### 4.2.7. Statistical Analysis

Linear regression analysis was performed to understand the relationships among different variables. The main aim was to analyse the influence that some independent variables may have on performance, stress, and cognitive load. This analysis is particularly useful to understand on which variables it's useful to work to improve the simulation outcome, to balance the stress felt by the students, and to manage the mental effort avoiding cognitive overload.

The 55 considered variables are described in Appendix C. They relate to demographic characteristics, performance, subjective questionnaires, stress, and cognitive load measured through Phasya's algorithm. To cope with missing data concerning different variables, several single and multiple linear regression models were computed with a different number of observations, depending on the considered variables. The number of observations refers to the number of students considered for that specific model.

As for the HFS, concerning the surveys composed of multiple questions, a single total score was calculated, for both pre- and post-simulation responses. In the case of missing answers, the total score was computed as explained in paragraph 4.1.7.

*Table 20*, *Table 21*, and *Table 22* show the most interesting and significant models, gathered in the three items: performance, stress, and cognitive load. Statistical significance is present when p-values are lower than 0.05. The following tables show the dependent variables, the number of observations (obs) per model, the p-value of the model (i.e. the significance F), and the significant independent variables. The p-value representation with asterisks indicates **p<.05; **p<.01; ***p<.001*.

4.2.7.1. Performance

*Table 20* shows the results of the best linear regression models for the analysis of the variables that influence students' performance in LFS.

*Table 20: Linear regression analysis about performance in LFS (*p<.05; **p<.01; ***p<.001)*

| Dependent Variable | Obs | Significance F | P-Value < 0,05 |
|---|---|---|---|
| Δ SKILLS | 79 | 0,026* | GENDER (+) *<br>SESSION (+) *<br>STAI TRAIT (-) * |
| TOTAL PERFORMANCE | 128 | 0,00045 *** | DATE (+) *<br>SESSION (+) *<br>STAI PRE (-) * |
| SUCCESS | 123 | 8,1E-16 *** | STAI TRAIT (-) **<br>GROUP (-) **<br>PREP PERF (+) *<br>PREP TIME (+) *<br>TEACHER (-) *** |
| ATTEMPTS | 110 | 0,009 ** | HEIGHT (+) *<br>DATE (-) **<br>ERRORS (+) * |
| ERRORS | 80 | 0,023 * | STAI TRAIT (-) **<br>NASA MENT (+) * |
| | 105 | 0,003 ** | STAI TRAIT (-) *<br>THESIS (-) *<br>SESSION (-) *<br>ATTEMPTS (-) *<br>TEACHER (-) ** |
| PREP TIME | 61 | 0,002 ** | SEQUENCE (-) ***<br>EXPERIENCE (+) *<br>SIM ATT PRE (-) * |
| SUCCESS TIME | 122 | 5,4E-32 *** | DATE (-) ***<br>GROUP (+) *<br>ATTEMPTS (+) ***<br>CL LP (+) * |
| TOT TIME | 80 | 1,31E-21 *** | GROUP (+) **<br>TECH ATT (-) *<br>ATTEMPTS (+) *** |

The first highlight is that students' performance is not influenced by perceived and physiological stress. Indeed, none of the dependent variables is statistically related to NAS and stress detected through Phasya's algorithm. Instead, some performance items are statistically related to cognitive load. The number of committed errors is directly related to the mental demand of NASA-TLX. In other words, an excessive perception of mental effort, which may result in cognitive overload, provokes an increment of errors. Students may feel to have too many new notions to remember during the simulation and this may cause confusion and loss of concentration, resulting in an increment of wrong actions. In the same way, cognitive load influences also the time to success. Indeed, students take more time to make the liquor pouring out (i.e. to make a correct puncture) when the level of physiological cognitive load during the puncture phase is higher.

Total performance is influenced by the date and moment of the day. Indeed, students perform better in the morning and the last days of simulation training. Moreover, Total Performance depends on the anxious state of the students before beginning the simulation session: greater anxiety causes worse performance. The students' anxious trait (STAI Trait) also influences performance. Indeed, success in collecting the CFS depends on STAI Trait (no success for subjects very anxious). However, very anxious students commit fewer errors. Errors depend also on the thesis developed by the students (fewer errors for subjects with a thesis relevant for the rachicentesis simulation), the number of attempts, and the teacher's help. Indeed, students helped by the teacher committed fewer errors. However, success is not reached if the teacher practically assists the learner ($p<.001$). Success also depends on preparation time and performance. Finally, the number of attempts is directly related to errors.

Concerning performance time, the preparation time is longer for subjects without previous experience and for participants performing the LFS first (sequence has *p*< *.001*). Indeed, successive students have the opportunity to watch and learn from previous ones. Time to success obviously depends on the number of attempts, and Total time also depends on the students' attitude to technology (subjects more familiar with technological devices take less time).

Even the improvement in acquired skills (Δ SKILLS) does not depend on perceived and physiological stress and cognitive load. However, it is inversely related to STAI Trait: less anxious subjects learn better in LFS.

4.2.7.2. Stress and Emotional State

*Table 21* shows the results of the best linear regression models for the analysis of the variables that influence students' stress and emotional state.

*Table 21: Linear regression analysis about stress and emotional state in LFS (*p<.05; **p<.01; ***p<.001)*

| Dependent Variable | Obs | Significance F | P-Value < 0,05 |
|---|---|---|---|
| STAI TRAIT | 140 | 2,44E-14 *** | STAI PRE (+) *** |
| STAI POST | 81 | 4,88E-10 *** | GROUP (-) *<br>STAI TRAIT (+) ***<br>Total Performance (-) **<br>NAS 2 (+) ***<br>STRESS SIM (+) *** |
| NAS 2 | 82 | 0,00019 *** | STAI POST (+) ***<br>PERF TOT (+) *<br>STRESS (+) * |

| | | | |
|---|---|---|---|
| Δ STAI | 117 | 3,63E-12 *** | Δ NAS (2-1) (+) ***<br>GENDER (-) *<br>TECH ATT (-) *<br>NASA-TLX Frustration (+) * |
| Δ NAS (2-1) | 116 | 0,002 ** | TOT TIME (+) **<br>NASA-TLX Frustration (+) ** |
| STRESS SIM | 80 | 0,001** | STAI PRE (+) ***<br>ERRORS (+) * |
| NASA TEMP | 114 | 0,025 * | SMOKER (-) *<br>SESSION (-) * |
| NASA PERF | 132 | 6,3E-05 *** | SUCCESS (+) ** |
| NASA EFF | 126 | 0,004 ** | HEIGHT (+) **<br>TECH ATT (-) *<br>ATTEMPTS (-) *<br>TOT TIME (+) * |
| NASA FRUSTR | 104 | 0,0007 *** | TOT TIME (+) *<br>Δ NAS (2-1) (+) * |

As for the HFS statistical analysis, even in this case, STAI Trait is strongly directly related to STAI State PRE ($p<.001$). This means that the perception that subjects have about their anxious state in life is highly influenced by the anxiety they feel in that moment.

After the LFS, the sensation of anxiety highly depends on the general anxiety trait ($p<.001$), and on the stress measured ($p<.001$) and perceived ($p<.001$) during and ten minutes after the simulation. Moreover, even the Total Performance influences the anxiety felt after the simulation ($p<.01$): subjects who performed better are less anxious.

The anxiety variation between pre- and post- simulation directly depends on the variation in subjective stress ($p<.001$) and on the frustration felt during the simulated rachicentesis ($p<0.5$). The anxiety change is also influenced by gender (women are more anxious than men) and attitude to the use of technology (subjects who are not familiar with technological devices feel more anxious).

The perceived stress ten minutes after the LFS directly depends on how much anxiety subjects feel at that moment (*p<.001*) and on Total Performance (*p<0.5*). It is also influenced by the physiological stress during the simulation: for greater stress levels during the simulation, higher perceived stress after the LFS. Moreover, the variation in self-assessed stress before and after the simulation is directly related to the Total Time of the performance (*p<0.1*) and the felt frustration (*p<0.1*).

Lastly, the level of stress during the simulation, computed through Phasya's algorithm, directly depends on the subjects' anxious state before the LFS (*p<.001*) and on the errors committed during the simulated rachicentesis procedure (*p<0.5*).

Concerning the indexes related to perceived workload, self-assessed through NASA-TLX, it results that the weight attributed to temporal demand depends on the period of the day in which students performed LFS (temporal demand is perceived higher in the morning).

While Performance has a greater weight when students succeed in collecting the liquor, the effort is higher for tall subjects (*p<.01*) and for long LFS execution times (*p<0.5*). Perceived effort is also higher for a low number of attempts and low attitude toward technology.

The sense of frustration directly depends on the total duration of the simulation and the variation of perceived stress (*p<0.5*).

4.2.7.3. Cognitive Load

*Table 22* shows the results of the best linear regression models for the analysis of the variables that influence students' cognitive load.

*Table 22: Linear regression analysis about cognitive load in LFS (*p<.05; **p<.01; ***p<.001)*

| Dependent Variable | Obs | Significance F | P-Value < 0,05 |
|---|---|---|---|
| CL SIM | 102 | 0,0005 *** | GENDER (+) **<br>BIRTH (-) *<br>GROUP (-) *<br>TOT TIME (+) *<br>Total Performance (+) *<br>ERRORS (+) **<br>STAI PRE (+) ** |

Perceived cognitive load (through NASA-TLX, Mental Demand) resulted in no significative linear regression models. This means that there are no variables that statistically influence the subjects' perception of mental effort. However, the physiological cognitive load, measured through Phasya's algorithm, is statistically related to some independent variables.

An interesting linear regression model suggests that CL directly depends on gender (women have higher CL during the LFS) ($p<.01$), sense of anxiety before the simulation ($p<.01$), and on parameters related to performance such as the total time of the simulation ($p<.05$), the total performances ($p<.05$), and the committed errors ($p<.01$). Thus, the more the subject commits errors, the more try to concentrate and focus on the execution of the LFS. Moreover, cognitive load depends also on the simulation group: groups practicing later in the morning or in the afternoon have a lower CL because maybe they feel fatigued by what they have done earlier.

### *4.2.8. Main Findings LFS: Summary*

#### 4.2.8.1. Performance

Some highlights related to performance are hereunder listed:

- More than 90% of students were able to correctly perform the puncture and make the CFS pouring out.
- The percentage of committed errors is quite low (21.13% of students touched the needle where it was not allowed, and 32.39% completely came in-and-out with the needle during the procedure).
- The percentage of students who incorrectly executed the tasks is under 4%, while the percentage of participants who did not perform some tasks is equal to 8.69% for the preparation phase and 11.27% for the end of the simulation. Thus, students understand how to execute the tasks but sometimes they forget some steps.
- Performance is not statistically influenced by perceived and physiological stress.
- Some performance items are statistically related to cognitive load.
- An excessive perception of mental effort, which may result in cognitive overload, provokes an increment of errors.
- Greater anxiety causes worse performance, less anxious subjects learn better. However, very anxious students commit fewer errors.
- Students helped by the teacher committed fewer errors. However, success is not reached if the teacher practically assists the learner.

- Preparation time is longer for subjects without previous experience and for participants performing the LFS first.

#### 4.2.8.2. Stress and Emotional State

Some highlights related to Stress and Emotional State are hereunder listed:

- Perceived anxiety decreases after the simulation.
- Perceived level of stress (NAS) after the simulation decreases below the one perceived at the arrival in the simulation room even if, in this case study, the debriefing is not provided. Indeed, the LFS is not as emotionally engaging as the HFS.
- The mean stress level physiologically calculated over the entire simulation is lower than the mean level in high-fidelity simulation.
- The physiological stress is on average higher than the cognitive load.
- The peak level of stress is during the preparation phase. The mean stress level during the rest period after the simulation, is lower than the one calculated in the rest period before the simulation. This confirms the lack of debriefing necessity.
- The perceived stress ten minutes after the LFS directly depends on Total Performance. It is also influenced by the physiological stress during the simulation: for greater stress levels during the simulation, higher perceived stress after the LFS. The variation in self-assessed stress before and after the simulation is also directly related to the Total Time of the procedure and the felt frustration.

- The level of physiological stress during the simulation directly depends on the subjects' anxious state before the LFS and on the errors committed during the procedure.
- The effort is higher for tall subjects (physical ergonomics must be considered), long LFS execution times, low number of attempts, and low attitude toward technology.
- The sense of frustration directly depends on the total duration of the simulation and the variation of perceived stress.

#### 4.2.8.3. Cognitive Load

Some highlights related to Cognitive Load are hereunder listed:

- The perceived workload after the low fidelity simulation is on average high, with the greatest weight is assumed by the performance.
- The mean physiological CL level results lower than the mean level in HFS.
- Physiological CL is on average lower than physiological stress during the LFS.
- The peak level of CL is during the preparation phase even if it remains approximately constant during the simulation. The CL level during the rest period after the simulation is lower than the one calculated in the rest period before the simulation. This confirms the lack of debriefing necessity.
- While there are not variables that statistically influence the subjects' perception of mental effort, the physiological cognitive load results

statistically related to gender, sense of anxiety before the simulation, total time of the simulation, total performances, and committed errors. This result confirms the relevance of using physiological (objective) measures together with self-assessment measures.

## 4.3. Mixed Reality Simulation

### *4.3.1. Rachicentesis Simulation in Mixed-Reality*

The pilot study for the effectiveness analysis of mixed reality simulations has been conducted with the developed MR prototype for rachicentesis. Therefore, the simulation procedure, phases, tasks, objectives were the same described in paragraph 4.2.1. As for the low fidelity simulation, while one subject performed the simulation, the other participants were in the same room, observing the performance. Even the room layout and equipment were the same as the LFS, with the only addition, in this case, of the two HMD Vox Gear Plus and Hololens. Therefore, the main difference with the LFS consists of a greater immersion in the simulation, a higher realism, and a better interaction with the simulated patient. Indeed, the AR patient gives auditory and visual stimuli to the learner, in response to his/her action, as described in Chapter 3.2.

*Figure 63: A student, equipped with wearable sensors, performing MR rachicentesis simulation with Vox Gear Plus*

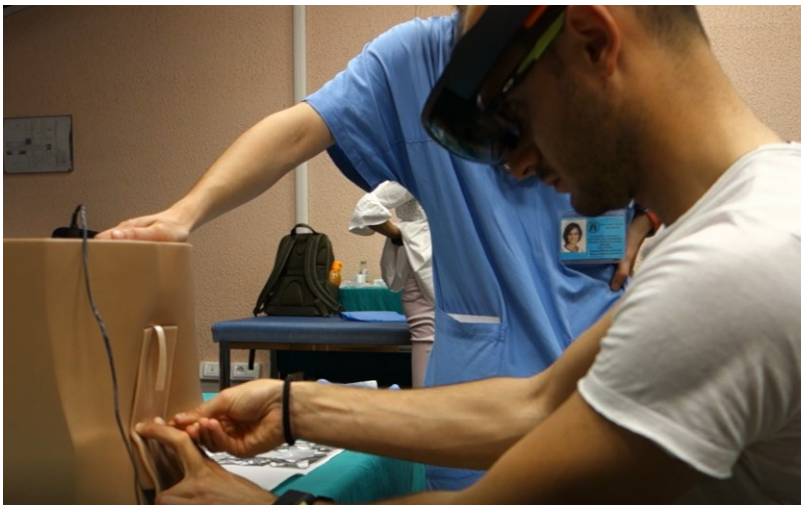

*Figure 64: A student, equipped with wearable sensors, performing MR rachicentesis simulation with Hololens*

The MR simulation was video-recorded, and *Figure 63* and *Figure 64* show two students performing the rachicentesis in mixed reality using the different HMD (Vox Gear Plus and Hololens).

### *4.3.2. Data Acquisition Workflow*

Forty students of the 6th year of Medicine and Surgery Course (of the next year respect to those who participated in the HFS and LFS) were randomly selected for the pilot study. After the explanation of the research study and the mixed reality simulation with the eventual contraindications, thirty-six students signed the informed consent and the form for the processing of personal data. The thirty-six students were randomly divided into two groups and eighteen of them were assigned to the MR simulation with the Vox Gear Plus, and the other eighteen participants to the MR simulation with the Hololens. Each group was consecutively divided into 3 subgroups of six students each. Thus, for each simulation, in turns, one student performed the rachicentesis while the other five participants looked at him/her.

For the assessment of stress and cognitive load, the workflow in *Figure 55* was followed. However, due to the technical limitations of the device, it was impossible to complete the rachicentesis procedure with the Vox Gear Plus, since it was not safe for the participants. Conversely, with Hololens, it was possible to accomplish the entire simulation procedure. Therefore, data and results in the following paragraphs refer only to the self-assessment questionnaires, physiological parameters, and performance of the eighteen students who practiced the MR rachicentesis with Hololens (data collected before the simulation with the Vox Gear Plus was discarded). Performance and skills were acquired as explained in paragraph 4.2.2.

### *4.3.3. Participants*

The characteristics of the eighteen participants enrolled in the pilot study with Hololens are described in *Table 23*.

In this case, men (55.55%) were more than women (44.45%), and the mean age was of 25.6 (± 0.6) years. 33.33% of them was a smoker and 83.33% of them was working on a degree thesis pertinent to the rachicentesis simulation (as for the LFS we considered pertinent: anaesthesia and intensive care, cardiology, every surgery, emergency, gastroenterology, obstetrics and gynaecology, orthopaedics, pneumology, radiology, and urology). Moreover, 38.89% have already had previous experience with invasive procedures.

*Table 23: MR simulation - Participants characteristics*

| | % OF PARTICIPANTS |
|---|---|
| GENDER | |
| • Male<br>• Female | 55.55 %<br>44.45 % |
| SMOKER | |
| • Yes<br>• No | 33.33 %<br>66.66 % |
| PERTINENT THESIS | |
| • Yes<br>• No | 83.33 %<br>16.66 % |
| PREVIOUS EXPERIENCE IN | |
| • Group simulation<br>• Residency<br>• Professional training activity<br>• Training course (e.g. BLS, ATLS)<br>• Working experience<br>• Health volunteer activities<br>• Invasive procedure | 83.33 %<br>50.00 %<br>94.44 %<br>55.55 %<br>11.11 %<br>11.11 %<br>38.89 % |

Concerning the self-assessed psychological state of the eighteen participants before the MR simulation, mean values of STAI (Trait and State PRE) and NAS (at the arrival in the simulation room, and ten minutes after) are shown in *Table 24*.

*Table 24: Subjective psychological state of participants before MR training*

| | MEAN | STANDARD DEV. |
|---|---|---|
| STAI TRAIT | 45.17 | 8.92 |
| STAI STATE PRE | 39.23 | 10.27 |
| NAS 0 | 2.17 | 1.29 |
| NAS 1 | 2.33 | 1.24 |

Even in this case, on average, participants are more anxious in their life than in the classroom before the simulation (STAI Trait > STAI State PRE). However, even if attention must be paid in comparing the data about MR and LF simulations due to the different sample dimension, it is possible to note that, in this case, participants are less anxious with respect to the ones enrolled in the LFS (both STAI Trait and State are lower). Moreover, as in the previous case studies, before the simulation, students do not feel stressed.

Concerning the aptitude of the subjects to technology, the percentages of the answers to the survey, described in *Table 3* and administered before the MR simulation, are visible in *Figure 65* (section regarding the personal use of technological devices) and *Figure 66* (section regarding the relevance of technology in the simulation context).

Most students (more than 72%) are familiar with the use of everyday life technological devices such as computers, tablets, and smartphones. Conversely, only 16.67% have used haptic gloves for gaming. However, 55.56% of participants are used to play with simulation videogames and/or serious games and are familiar with the use of head-mounted displays.

Regarding the use of smart wearables for healthcare and lifestyle, only 16.67% are familiar with them.

However, 61.11% of students feel suitable to work in a high-tech environment, and only 22.22% of them would feel stressed to work in it.

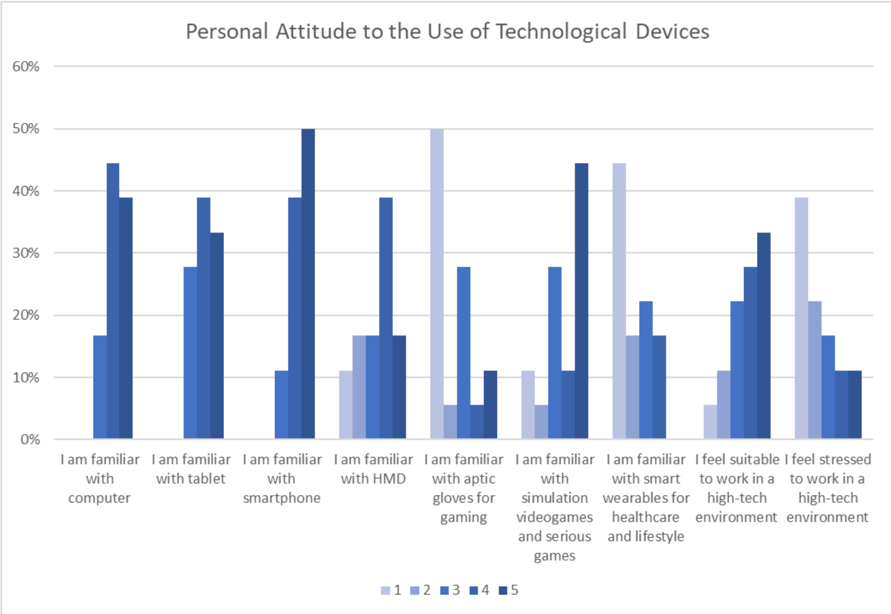


*Figure 65: Results of the survey about aptitude to technology (Section: Personal attitude to the use of technological devices) before MR simulation*

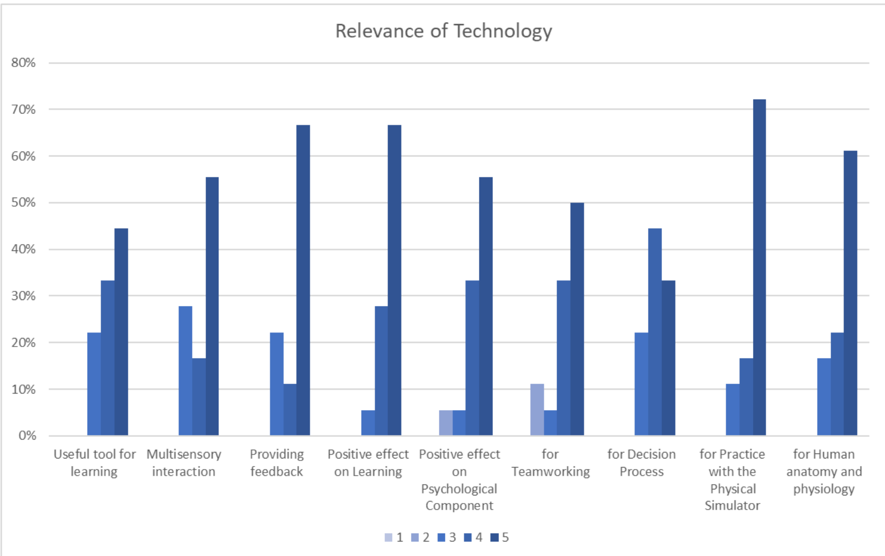


*Figure 66: Results of the survey about aptitude to technology (Section: Relevance of technology in simulation context) before MR simulation*

Technological devices (such as virtual reality glasses, gloves with haptic feedback, etc.), multisensory interaction (tactile, visual, and auditory), and feedback provided through technological devices are considered useful tools for learning during training by the majority of participants (respectively 77.78%, 72.22%, and 77.78%). Moreover, a high degree of immersion during the simulation is thought to have a positive effect on learning (94.44%) and psychological component (88.89%).

Having the opportunity to choose on which fields being supported, during simulations, by high-tech devices, 83.33% of participants would prefer team-working, 77.78% decision process activities, 83.33% understanding of human anatomy, physiological and pathological processes, but, most of all, 88.89% desire to be supported for the practice with physical manikins, to improve the low-fidelity simulations.

#### *4.3.4. Performance Analysis*

Performance and skills were assessed for all the eighteen participants enrolled in the pilot study with Hololens.

Students' skills were assessed employing the same 5-items questionnaire used for the low-fidelity simulation. It was answered before and after the training, and, while the mean number of correct responses before the simulation was equal to 3.61 (± 1.04), after the simulation is increased to (4.72 ± 0.57).

Regarding the performance, success was reached in 94.45% of cases. Thus, most participants were able to correctly perform the puncture and make the liquor pouring out. The mean number of attempts to reach the success was 2.21 (± 2.03), and the mean number of errors per subject was 0.33 (± 0.49).

*Figure 67* shows the percentage of each error for the total sample of students. 27.78% came in-and-out with the needle during the procedure, and only 5.56% touched the needle where it is not allowed (both actions should be avoided for hygienic reasons). No one extracted the needle without re-inserting the stylet.

*Figure 68* shows the percentage of correctly/incorrectly/not performed tasks, divided per phases (preparation and tasks after the CFS pouring out). In both phases, most tasks (more than 94%) were correctly performed. The percentage of incorrectly performed or not executed tasks is under 6%.

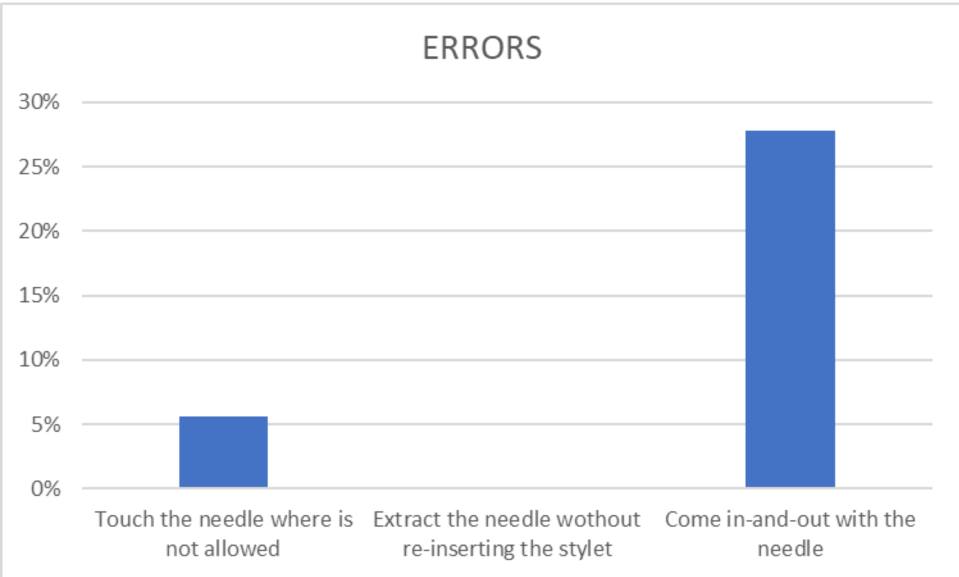


*Figure 67: Percentage of committed errors during MR simulation*

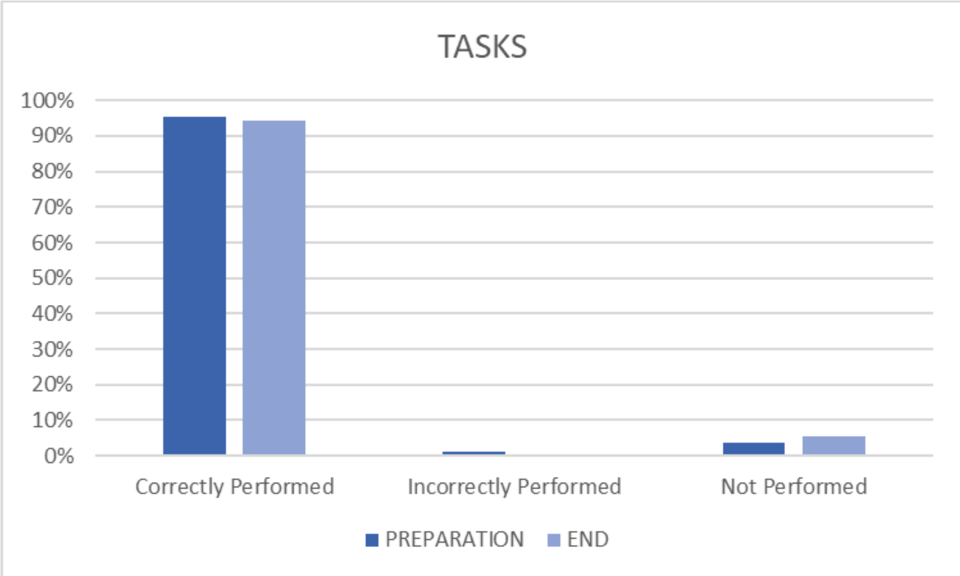


*Figure 68: Students' performance per phases during MR simulation*

Finally, mean execution times were calculated for students who reached success:

- Preparation Time: 1.16 ± 0.50 minutes
- Time to success: 3.15 ± 2.87 minutes
- Time to finish: 0.92 ± 1.79 minutes
- Simulation Total Time: 5.92 ± 3.09 minutes

On average, students' performance in MR rachicentesis simulation is slightly better than the one achieved by participants in traditional low fidelity simulation. However, since the great difference in sample dimension, this difference in performance cannot be attributed to the use of mixed reality but only to reasons intrinsic to the participants involved in the studies. However, the trends in tasks correctly performed, committed errors, and simulation times are comparable between LFS and MRS.

#### *4.3.5. Psychometric Analysis*

This paragraph describes the results of self-assessment questionnaires for the analysis of perceived anxiety, stress, and cognitive load, and differences in students' opinion about the relevance of technology, after the mixed reality simulation.

##### 4.3.5.1. The relevance of technology in the simulation context

*Figure 69* shows the comparison between students' point of view, before and after the MR training with Hololens. All the eighteen participants correctly answered the survey before (PRE) and after (POST) the simulation. Results in *Figure 69* represent the mean values of each item (on the 5-points Likert scale).

First, it should be kept in mind that in this case the AR application was thought to improve the sense of realism and immersion, giving visual and auditive realistic feedbacks at the interaction between learner and patient. Thus, it was not designed to give informative feedbacks or to make the simulation easier.

After having experienced the rachicentesis simulation with the AR responsive patient, students continue to think that technological devices and multisensory interaction are valuable tools for learning. However, their opinion about the utility of providing feedback through technological devices decreases a bit.

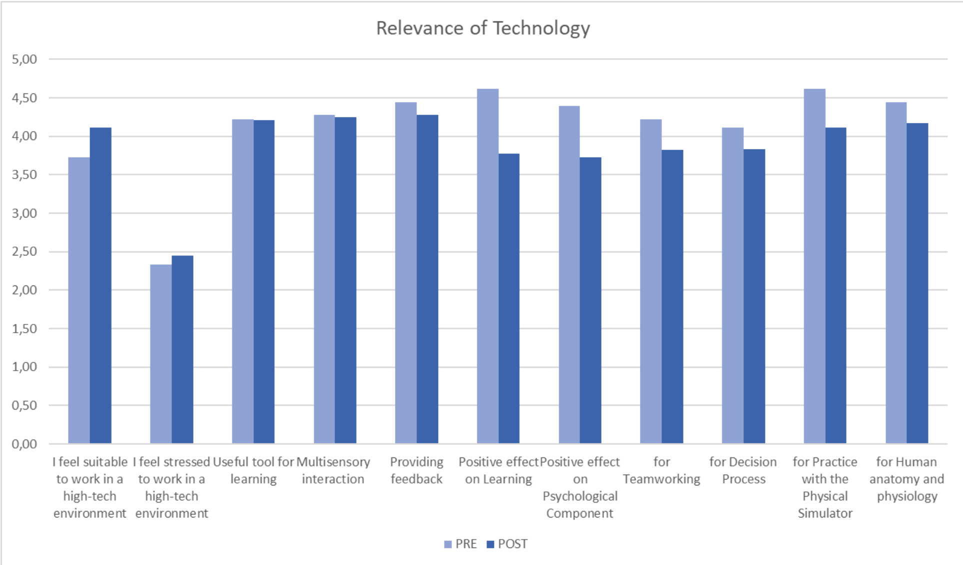


*Figure 69: Relevance of technology in the simulation context: comparison between opinions collected before and after the MR simulation*

Overall, after the MR simulation, even if participants feel a little bit more stressed to work in a high-tech environment, they also feel quite a lot more suitable to work in it.

Unfortunately, their opinion about the positive effect of the high degree of immersion on learning and psychological component decreases. Even their desire of being supported by high-tech devices, during the simulation, has an overall decrement. However, the practice with physical manikin is confirmed as the simulation that would benefit the most by the use of high-tech devices, together with the understanding of human anatomy, physiology, and pathology.

4.3.5.2. State Trait Anxiety Inventory (STAI)

*Table 25* shows the mean values of STAI, before and after the simulation in mixed reality, computed over fifteen participants (83.33%). Indeed, three subjects missed some responses after the training.

*Table 25: STAI results before and after MR simulation*

| | MEAN | STANDARD DEV. |
|---|---|---|
| STAI TRAIT | 45.17 | 8.92 |
| STAI STATE PRE | 40.81 | 10.14 |
| STAI STATE POST | 37.53 | 9.69 |

As in the previous cases, perceived anxiety decreases after the simulation (mean STAI State Post < mean STAI State Pre). The mean STAI State Post agrees with the value expected for college students (Spielberger, et al., 1983), while the mean STAI State Pre is few points higher, confirming the increment of anxiety felt before the simulation.

4.3.5.3. Numeric Analog Scale (NAS)

Results in *Figure 70* refer to the mean values of perceived stress, on a 10-points Likert scale, calculated on the eighteen subjects, at the arrival in the simulation room (NAS 0), ten minutes later, and ten, twenty, and thirty minutes after the end of the MR simulation (respectively NAS 2, NAS 3, and NAS 4).

Even in this case study, the peak in NAS 2 corresponds to the perceived stress during the rachicentesis training. However, here, even if the feeling of stress constantly decreases from NAS 2 to NAS 4, thirty minutes after the end of the simulation the perception of stress is still high. Indeed NAS 4 is higher than NAS 0. Maybe, a debriefing should be suggested, to keep the perceived stress under the basal level.

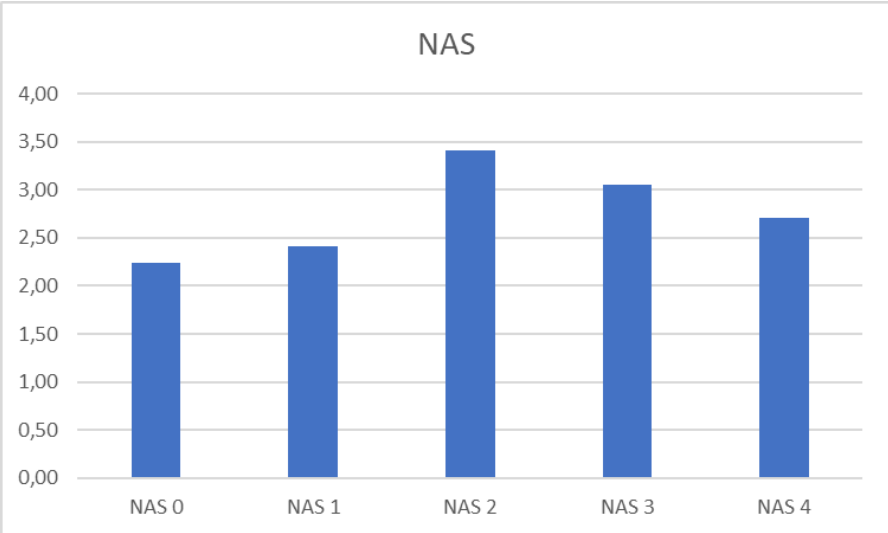

*Figure 70: Mean NAS results (10-points Likert scale) for MR simulation*

Always considering the difference in the number of participants, and the impossibility of making correct comparisons, it can be observed that mean NAS 2, NAS 3, and NAS 4 in MRS are higher than in the LFS. This could be an index of the enhanced realism and immersion in the simulation that causes an increment of perceived stress through the interaction with the AR patient, who responds to the learner's actions.

4.3.5.4. NASA Task Load Index (NASA-TLX)

The response rate of NASA-TLX was equal to 100%. *Table 26* shows the mean values of the total score for the perceived workload related to MR rachicentesis, and the weight assigned to the six indexes (mental, physical, temporal demands, performance, effort, and frustration).

*Table 26: NASA-TLX results for MR simulation*

| | MEAN | STD. DEV. |
|---|---|---|
| NASA-TLX | 50.41 | 19.52 |
| Mental Demand | 188.50 | 126.82 |
| Physical Demand | 59.61 | 68.51 |
| Temporal Demand | 108.28 | 116.26 |
| Performance | 203.50 | 120.56 |
| Effort | 136.67 | 88.30 |
| Frustration | 59.67 | 103.81 |

The mean perceived workload for the simulation in mixed reality is on the inferior boundary of high-range (high for 50-79 total score), according to literature's score interpretation (Sugarindra, et al., 2017).

Even if the small sample of participants does not allow for comparisons with LFS, some preliminary considerations can be outlined. First, the total score of NASA-TLX is two points lower than in the LFS: the use of high-tech devices seems to not cause increments of perceived workload. However, the weight attributed to the mental demand is a few points higher than in the low-fidelity simulation. Nevertheless, as for the LFS, the performance received the greatest weight. Even the effort obtained high values, in agreement with low-fidelity simulation, but the use of Hololens and AR does not increase the effort to accomplish the tasks. Moreover, the perceived physical demand and the frustration are on average lower than in the LFS.

#### *4.3.6. Biometric Analysis*

In this paragraph, levels of stress and cognitive load, measured through Phasya's algorithm, are presented.

Stress and CL levels, that range between 0 and 6, were computed for simulation phases.

Even if physiological signals were collected for the entire sample of participants, results refer to fifteen subjects (83.33%), because some data were affected by errors.

The mean stress and CL levels, computed over the entire MR rachicentesis, are respectively equal to 2.24 (± 2.06) and 2.20 (± 1.28). Thus, the stress is on average a bit higher than the cognitive load.

However, as visible from *Figure 71*, the cognitive load is higher than the stress during the puncture execution. It is worth to note that the cognitive load increases from preparation to puncture phase, and then decreases toward the end of the simulation until the rest after the training. Conversely, the stress increases from the basal level before the simulation to the preparation phase, and then constantly decreases during the following phases. The stress and CL levels after the training on average turn back to the basal levels (even if debriefing is not performed). On average, stress and CL during the execution of the procedure in mixed reality can be considered medium, not high (on the 6 levels scale).

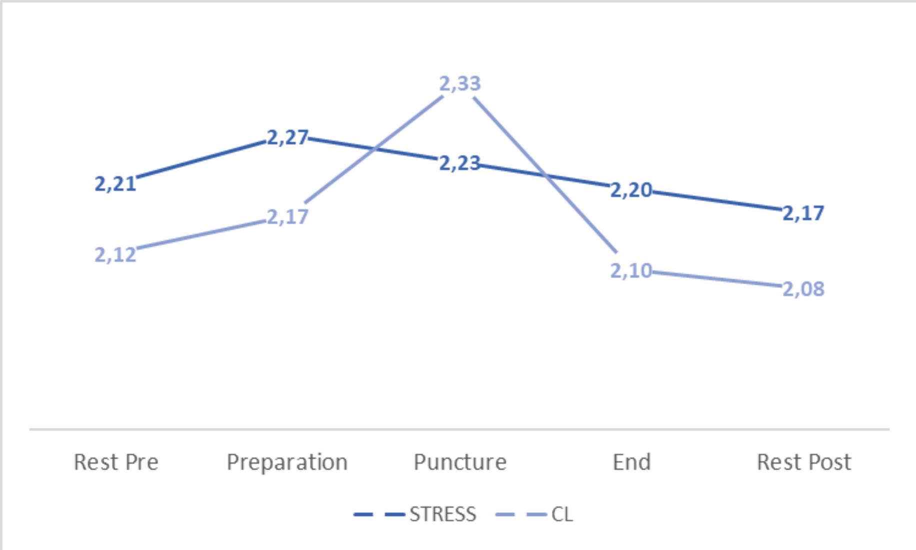


*Figure 71: Stress and CL during MR simulation phases*

Preliminary observations can be done on the difference in stress and CL relationship between LFS and MRS. While for the rest period before the simulation and for the preparation phase stress is higher than CL in both simulations, from the lumbar puncture phase to the rest after the simulation, their relationship is inverted. Indeed, in MRS, conversely to the LFS, during the last phase of the simulation and the rest period after the simulation, stress is higher than CL. This could be an index of the fact that the use of MR, and the enhanced realism, causes an increment of stress (also confirmed by the NAS). However, even with the use of AR, it remains stable on a medium level.

### *4.3.7. User Experience Evaluation*

In this paragraph, the main advantages and drawbacks of both devices (Vox Gear Plus and Hololens), tested in the mixed reality rachicentesis simulation, are outlined. The results of a semi-structured interview, for user experience evaluation, are hereafter reported.

#### 4.3.7.1. Vox Gear Plus

Vox Gear Plus is a low-cost solution, consisting in a headset containing a smartphone. This technology gives the user the impression of being immersed in a virtual world, leaving the opportunity to actively interact with the surrounding real environment. However, it presents several limitations:

- Patient Tracking: despite the "extended tracking" option activated, when the patient's target gets out of the user's field of view for long periods, the 3D AR model tends to misalign or disappear. Indeed, the smartphone uses the accelerometers to perform extended tracking, and thus to estimate the AR patient's position based on the movements of the user's head, once the target has been lost. Unfortunately, this technique has limited accuracy. Rachicentesis is a complex operation and requires the student to use both hands and focus on the area of lumbar puncture. Thus, it is difficult to find an area, always free from obstacles, on which positioning the target.
- Loss of Depth: the smartphone is equipped with a single camera and thus is unable to record the real scene in a stereoscopic manner. Therefore, the user sees the same duplicated images in front of the left and right eye. For

this reason, the user can see a unique image, but the sense of depth is lost (the sensation is to do the operation with only one open eye). This makes impossible the execution of an invasive procedure such as rachicentesis.

- Lag and nausea: these are typical problems of all the extended reality systems. Using the device for a few minutes, it is possible to feel a sense of nausea due to the delay (LAtency Gap) between what our eyes see and our movements.
- Obstruction: the headset device, even if light, ergonomic and comfortable, constitutes an obstacle.

#### 4.3.7.2. Hololens

The use of Hololens allows to solve some of the issues highlighted above:

- Patient tracking: Hololens are equipped with depth sensors that allow, in a few moments, to reconstruct a relatively detailed mesh of the surrounding environment. Thanks to this and other features, the tracking of fixed elements in space is optimal, even after losing the image target from the user's field of view, for a long time.
- Depth Perception: Hololens allows seeing through the lenses of the device; the vision is stereoscopic, and the depth perception is optimal.

However, even Hololens presents various limitations:

- Narrow Field of View (FOV): the Hololens' FOV is only 30° x 17.5°; this means that only a small rectangle in front of the user can be actually populated by holograms. Consequently, during the rachicentesis simulations, the area available to view the 3D AR model of the patient is that of the real manikin, which, therefore, at a close distance, will never appear complete. The user is forced to move his/her head to see the rest of the AR patient's body, thus reducing the sense of realism and immersion.

- Slow Needle Tracking: although patient tracking is optimal, the tracking of the needle is much slower and more responsive than with Vox Gear Plus.
- Obstruction: even in this case, despite the lightness and ergonomics of the devices, it is however a possible obstacle.

4.3.7.3. Semi-structured survey

All the thirty-six students involved in the pilot study, after having performed the rachicentesis simulation in mixed reality, repeated it in the low-fidelity manner. At the end of both the executions, a semi-structured survey for the assessment of user experience was administered. Results in *Table 27* and *Table 28* show the mean values calculated over eighteen subjects who used Hololens and eighteen subjects who used Vox Gear Plus.

*Table 27: User Experience about the use of AR in simulation training [5-points Likert scale]*

| | **Hololens** | **Vox Gear Plus** |
|---|---|---|
| AR use has increased the feeling of immersion and involvement compared to the use of the manikin alone | 3,00 | 3,82 |
| AR field of view is satisfying for the simulation execution | 2,64 | 3,53 |

*Table 28: User Experience about the use of AR in simulation training [Agree/Disagree]*

| | **Hololens [%Agree]** | **Vox Gear Plus [%Agree]** |
|---|---|---|
| AR allowed me to carry out checks and actions that I could not do with the only physical manikin | 24% | 29% |
| AR prevented me from carrying out checks and actions that I could carry out with the only physical manikin | 18% | 35% |
| AR made the operation more difficult | 41% | 76% |
| The visive feedback is effective | 94% | 96% |
| The auditory feedback is effective | 24% | 20% |

Therefore, although the use of both devices in the mixed reality simulation increased the feeling of immersion and involvement respect to the low-fidelity simulation, the use of Vox Gear Plus was more appreciated from this point of view. Vox Gear Plus also has a more satisfying field of view to maintain the realism of the clinical case, respect to Hololens. The percentage of students who thought that AR allowed to carry out checks and actions that could not be done with the only physical manikin is under 30% for both devices. However, while 18% of students who used Hololens stated that AR prevented them from carrying out checks and actions that could be done with the only physical manikin, the percentage rises to 35% for students who used Vox Gear Plus. In the same way, while 41% of students who used Hololens declared that AR made the procedure more difficult, for the same item the percentage increases to 76%.

Finally, no great difference emerges between the devices regarding the most effective and useful feedback: the visive feedback is preferred against the auditory feedback with a percentage of over 90% from both kinds of devices.

Other considerations about the use of Vox Gear Plus and Hololens are respectively reported in *Table 29* and *Table 30*.

*Table 29: Considerations about the use of Vox Gear Plus*

| **PRO** | **CONS** |
|---|---|
| The feeling of operating on a real patient | Very distorted depth perception |
| Very immersive and realistic | Difficult eye-hand coordination |
| Attention also shifts to the patient with whom we can interact | Nausea and dizziness |

*Table 30: Considerations about the use of Hololens*

| **PRO** | **CONS** |
|---|---|
| The vision of the whole body helps to orientate yourself | Attention was sometimes caught more by AR itself than by the procedure |
| Feeling of three-dimensionality | It is initially difficult to get used to |
| I needed to calm the patient down | The physical bulk of the device |
| Sometimes I looked at the patient's head to see if she was calm or agitated | Since the AR field of view is narrow and the procedure is focused on a specific area, it is like simulating without the AR device |

### *4.3.8. Main Findings MRS: Summary*

The main issues observed in the pilot MR study are related to current technology limitations (e.g. narrow field of view, unrealistic depth perception, slow tracking of moving object, etc).

These technology limitations, as well as training environment, may alter the emotional state and cognitive load of students compared to real clinical situation. They may cause frustration or stress and bias the students' emotional state and cognitive load.

Indeed, learners' opinion about the positive effect of the high degree of immersion on learning and psychological component decreases after having experienced the MRS. Even their desire of being supported by high-tech devices during the simulation has an overall decrement, since they declared to feel stressed working in a high-tech environment.

Whatever how advanced a simulation technique is, simulated training is always different from reality. It is therefore important to identify these differences and to understand how these differences affect the training. For this reason, a comparative study between traditional and mixed reality simulations need to be accomplished. However, the results of this pilot study give some valuable hints, hereunder summarised.

#### 4.3.8.1. Performance

Some highlights related to performance (using Hololens) are hereunder listed:

- Students' skills improved after the MRS.

- 94.45% of participants were able to correctly perform the puncture and make the liquor pouring out.
- 27.78% of participants came in-and-out with the needle during the procedure, and only 5.56% touched the needle where it is not allowed.
- The percentage of incorrectly performed or not executed tasks was under 6%.
- Although the great difference in sample dimension, the trends in tasks correctly performed, committed errors, and simulation times are comparable between LFS and MRS.

#### 4.3.8.2. Stress and Emotional State

Some highlights related to Stress and Emotional State, detected using Hololens, are hereunder listed:

- Perceived anxiety decreases after the simulation and agrees with the value expected for college students.
- Even if the perceived stress constantly decreases from NAS 2 to NAS 4, thirty minutes after the end of the simulation the perception of stress is still high (NAS 4 is higher than NAS 0). This may be due to the use of HMD. However, a debriefing should be suggested to keep the perceived stress under the basal level.
- Always considering the different sample dimension, it could be noted that the perceived stress in MRS is higher than in the LFS. This could be due to the enhanced realism and immersion in the simulation, and to the interaction with the AR patient, who responds to the learner's actions.

- The mean physiological stress is on average a bit higher than the cognitive load.
- The mean physiological stress increases from the basal level before the simulation to the preparation phase, and then constantly decreases during the following phases. After the training it turns back to the basal levels (even if debriefing is not performed).
- During the last phase of the simulation and the rest period after the simulation, stress is higher than CL. Maybe the use of HMD, MR application, and the enhanced realism, causes an increment of stress (also confirmed by the NAS).

#### 4.3.8.3. Cognitive Load

Some highlights related to Cognitive Load are hereunder listed:

- The mean perceived workload is on the inferior boundary of high range.
- The use of high-tech devices seems to not cause increments of perceived workload (the total score of NASA-TLX is two points lower than in the LFS). However, the weight attributed to the mental demand is a few points higher than in the low-fidelity simulation.
- The use of Hololens and AR does not increase the effort to accomplish the tasks.
- The physiological cognitive load is higher than the stress during the puncture execution.
- The CL level after the training on average turn back to the basal levels (even if debriefing is not performed).

# 5. Concluding Remarks

Simulation training is an instructional method always more used in the medical field. However, the measurement of its effectiveness in terms of performance, stress, emotional, and cognitive states still lacks a standard and consistent method.

This work presented a new, well-structured methodology suitable for the assessment of every kind of medical simulation. It allows fulfilling an overall and precise analysis of simulation effectiveness considering learners' performance, subjective feelings (especially anxiety, stress, and mental demand, but also frustration, effort, perceived workload), and physiological parameters related to stress and cognitive load (from an objective point of view). In the simulation context, not only performance but also stress and CL dimensions assume a great weight. Indeed, as it is important to achieve high performance, it is also fundamental to assure a good stress balance: simulation should bring learners to a stress level similar to that one in the real practice, and, at the same time, it should not provoke excessive stress which may damages performance, consequently resulting in increased risk for the patient. The same reasoning should be done for the cognitive load. Indeed, to guarantee better performance and the best skills memorisation, the mental effort should not be too low, and, at the same time, cognitive overload must be avoided.

Moreover, in the last years, the use of extended reality systems in medical simulation is becoming to take hold. Even in this case, the effect of these technologies on performance, stress, and cognitive load are still not clear. For this reason, a mixed reality prototype (AR + skill trainer) for rachicentesis simulation was implemented. The main aim of the AR application was to improve the realism and immersion of the simulation, by reproducing the rest of the patient body and by furnishing visive and auditory feedbacks related to patient movements and complaints. Informative feedbacks with additional indications, on how to perform the simulation, were not provided, to avoid helping students with additional information that they would not have in the real case.

The methodology for the assessment of simulation effectiveness was applied to three different case studies: the first one was the high-fidelity simulation for emergency management, the second one was the low-fidelity simulation for the practice of rachicentesis, and the third one was the pilot study about rachicentesis simulation in mixed reality. In addition to the descriptive analysis of performance, emotional, and cognitive conditions, while in the first two cases a statistical analysis was done to comprehend the relationships among different variables, in the third case (pilot study) a user experience analysis was accomplished.

This effectiveness analysis is useful to understand the weaknesses - and their causes - of the various simulations, and to draft some guidelines to re-design and optimise them. In *Table 31* and *Table 32* a synthesis of HFS and LFS issues are respectively reported and possible solutions are drafted.

Concerning the high-fidelity simulation, the analysis of performance, psychometric, and biometric measures revealed that the HFS was overall effective, and perceived and physiological cognitive states were appropriate both during the simulation and after the debriefing (also thanks to the debriefing itself). However, some issues related to high-fidelity simulation can be pointed out. *Table 31* drafts possible solutions to solve the problems related to HFS internal factors and improve the simulation outcome.

*Table 31: Possible solutions to HFS issues*

| | **Solutions to issues** |
|---|---|
| **Simulation** | |
| Simulation is perceived as not useful to decision-making | Decision-making skills could be more discussed during the debriefing |
| Simulation is not sufficiently perceived relevant for interprofessional education | Students of different degree course (nursing, lab technicians, …) could be included in the simulation training |
| **Performance** | |
| Performance is worse if aptitude to simulation is low | Simulation training could be introduced earlier in the academic path |
| Performance is worse if perceived CL is higher | Elements and notion in the simulation should be balanced to avoid overload of perceived mental effort and/or cognitive underload |
| Performance is worse if physiological CL is low | |
| Performance is worse in the last phases of the scenario | Do not introduce too many difficulties that may cause performance worsening |
| Performance is worse in the afternoon | If possible, try to organise the HFS in the morning |
| **Cognitive and Emotional States** | |
| Perceived stress is higher if the perceived workload is higher | Do not introduce too many elements that may cause cognitive overload and, consequently excessive stress |
| Physiological stress is higher if previous experiences are few | More practical experiences could be introduced in the academic path |
| Physiological stress is higher if performance is lower | The teacher could lead the students to positively solve the clinical case |
| Frustration is higher if performance is lower | |

Moreover, from the ECL analysis, it is clear that, in this specific case, a simulation room re-design is mandatory. It includes the creation of a control room for the teacher, separated from the simulation room. In this way, communication is forced to occur between the teacher and one student, and among students of the same group. The other students, who observe the simulation, should wait in another room, and not have the possibility to interact with the participants simulating the scenario. The disposition of the medical instrumentation should be also improved. All these elements would enhance the development of ICL against ECL.

The simulation room should also have a dedicated audio and video recording system. This would help the discussion in the debriefing phase.

Regarding the low-fidelity simulation, from psychological and physiological analysis, particular critical issues do not emerge. Stress and CL levels are well balanced during the simulation and they turn back to basal levels after the end of the LFS. Moreover, from statistical analysis, it is highlighted that students' performance is not influenced by perceived and physiological stress. These results confirm the non-necessity of debriefing in this kind of low-fidelity simulation.

However, statistical analysis drew attention to some issues that can be solved and optimised through a simulation re-design. In *Table 32* some basic guidelines are proposed.

First, an additional evaluation of physical ergonomics, concerning the impact of the workplace layout should be accomplished. Indeed, from this analysis, the height of the desk respects to the height of the learner seems to have an influence on the perceived effort. It could be interesting even to analyse the gestures kinematics during the simulation (by using motion capture) and study the comfort.

*Table 32: Possible solutions to LFS issues*

| | **Solutions to issues** |
|---|---|
| **Performance** | |
| Increment of errors for (too) high cognitive load | Feedback about the tasks' sequence and execution modality could be provided during the simulation |
| Increment of simulation time for (too) high cognitive load | |
| Success is not reached if the teacher practically assists the learner | The teacher should only assist the learner with advice, without interrupting his/her simulation execution |
| Performance is worse in the afternoon | Lesson contents could be divided into two sessions: theory in the morning and simulation in the afternoon (real cases could happen every time of the day) |
| **Cognitive and Emotional States** | |
| Anxiety before the simulation compromises performance and learning | The instructor should calm down the students during the briefing |
| Stress is higher when more errors are committed | If too many errors are committed, the instructor should give theoretical help |
| Stress increases for a long simulation duration | The simulation duration could be reduced (and consequently even stress, effort, and frustration) by providing feedback that could help the learners in the correct execution of the tasks |
| Frustration increases for long simulation duration | |
| Effort increases for long simulation duration | |
| The effort is higher for tall subjects | Physical ergonomics should be studied and adjusted |

This analysis reveals that rachicentesis simulation comprehends too many tasks. Indeed, students have to remember the task sequence and how to correctly execute them. This may cause a chain of negative events: an excessive number of tasks to be remembered may lead to cognitive overload, working memory overstepping, decrease in learning and performance, increase in committed errors, increased simulation time, an increment of stress, effort, and frustration. A good solution to avoid cognitive overload (and related consequences) could be providing informative feedback during the simulation. Feedbacks could be supplied by the instructor or through augmented reality applications developed ad hoc for that kind of simulation.

Moreover, since the time needed to succeed obviously depends on the number of attempts, other kinds of feedback could be provided to foster and speed up the lumbar puncture execution. For example, various haptic feedbacks could be added by integrating the skill trainer with force or pressure sensors. Learner's performance could be captured to provide him/her with positive or negative real-time feedback using immersive MR applications (Linde, et al., 2019). The combination of haptic and visual feedbacks through AR could guide and assist the learner in the procedure. This would reduce the number of attempts in stinging the lumbar spinal canal, and consequently also the time to success, the stress, the frustration, and the effort.

AR applications designed to provide feedback already exist in literature and on the market. Thus, in this work, the mixed reality application was designed and developed with another goal: to improve the sense of realism and immersion in the simulation, excluding elements that learners would not have or see in the real clinical case.

Using the developed mixed reality simulator, the performance was good, stress was not high, and cognitive load was adequately low. From this MR pilot study, it emerges that the main issues are related to technology limitations. Future work will include the validation process of the MR application, considering all the five steps (face validity, content validity, construct validity, concurrent validity, and predictive validity), and involving a statistically significant number of participants (the sample will be similar to the one enrolled in the assessed HFS and LFS).

The experimental study of the effectiveness of MR rachicentesis simulation will consist of a randomized controlled trial with two arms: some students will carry out the low-fidelity simulation with the traditional skill trainer, while other randomly extracted students will perform the simulation with the mixed reality prototype. In this way, it will be possible to analyse the effectiveness of MR simulations, and its impact on stress and CL, by comparing it with controls (LFS). Then the cross-over of the two arms will be carried out. Who tried the LFS will then try the MR prototype and vice-versa. In this way, it will be possible to see how stress, CL and acquired skills vary when repeating the procedure with one or the other kind of simulation.

The proposed standard and structured procedure for the assessment of the overall simulation effectiveness could also be applied to analyse the impact of every kind of extended reality applications, implementable both in high-fidelity and low-fidelity simulations. As explained in paragraph 4.1.8.4, the proposed procedure could also be easily simplified for its use in practice (i.e. for the assessment of every kind of training), by reducing the number of administered surveys and collected physiological parameters.

# References

**Akçayır M and Akçayır G** Advantages and challenges associated with augmented reality for education: A systematic review of the literature [Journal] // Educational Research Review. - 2017. - Vol. 20. - pp. 1-11.
**AL Sabei S. D. and Lasater K** Simulation debriefing for clinical judgment development: A concept analysis [Journal] // Nurse Education Today. - 2016. - Vol. 45. - pp. 42-47.
**AlRomi N** Human factors in the design of medical simulation tools [Conference] // 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, Procedia Manufacturing. - 2015. - Vol. 3. - pp. 288 – 292.
**Arora S [et al.]** The impact of stress on surgical performance: a systematic review of the literature [Journal] // Surgery. - 2009. - Vol. 147. - pp. 318–330.
**Atalay KD [et al.]** Assessment of mental workload and academic motivation in medical students [Journal] // Journal Of Pakistan Medical Association. - 2016. - Vol. 66.
**Barsom E.Z., Graafland M and Schijven M.P.** Systematic review on the effectiveness of augmented reality applications in medical training [Journal] // Surgical Endoscopy. - 2016. - Vol. 30. - pp. 4174-4183.
**Bhandary SP [et al.]** NASA Task Load Index Scale to Evaluate the Cognitive Workload during Cardiac Anesthesia Based Simulation Scenarios [Journal] // International Journal of Anesthesiology & Research (IJAR). - 2016. - Vol. 4. - pp. 300-304.
**Bjørshol CA [et al.]** Effect of socioemotional stress on the quality of cardiopulmonary resuscitation during advanced life support in a randomized manikin study [Journal] // Critical care medicine. - 2011. - Vol. 39. - pp. 300-304.
**Blanco M [et al.]** The impact of secondary task cognitive processing demand on driving performance [Journal] // Accident Analysis & Prevention. - 2006. - Vol. 38. - pp. 895-906.
**Boles DB [et al.]** Predicting dual-task performance with the multiple resources questionnaire (MRQ) [Journal] // Human Factors: The Journal of the Human Factors and Ergonomics Society. - 2007.

**Bond W. F. [et al.]** The Use of Simulation in Emergency Medicine: A Research Agenda [Journal] // Academic Emergency Medicine. - 2007. - Vol. 14. - pp. 353-364.
**Bosse HM [et al.]** The benefit of repetitive skills training and frequency of expert feedback in the early acquisition of procedural skills [Journal] // BMC Medical Education. - 2015.
**Bottino A [et al.]** Holo-BLSD: an Augmented Reality Self-directed Learning and Evaluation System for Effective Basic Life Support Defibrillation Training [Conference] // IMSH 2018 Proceedings. - 2018.
**Boucsein W** Electrodermal activity [Book]. - [s.l.] : Springer, 2012.
**Brunzini A [et al.]** Cognitive Load and Stress Assessment of Medical High-Fidelity Simulations for Emergency Management [Conference] // AHFE 2020. - 2020. - pp. 343–350.
**Brunzini A [et al.]** How to Improve Medical Simulation Training: A New Methodology Based on Ergonomic Evaluation [Conference] // AHFE 2019, AISC 963. - 2019. - pp. 145–155.
**CAEHealthcare** Patient simulation [Online]. - https://caehealthcare.com/patient-simulation/juno/.
**Cain B** A Review of the Mental Workload Literature [Journal]. - 2007.
**Carswell CM, Clarke D and Seales WB** Assessing mental workload during laparoscopic surgery [Journal] // Surgical Innovation. - 2005. - Vol. 12. - pp. 80-90.
**Castaldo P [et al.]** Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis [Journal] // Biomedical Signal Processing and Control. - 2015. - Vol. 18. - pp. 370–377.
**Chaballout B. [et al.]** Feasibility of Augmented Reality in Clinical Simulations: Using Google Glass With Manikins [Journal] // JMIR Medical Education. - 2016. - Vol. 2.
**Charles RL and Nixon J** Measuring mental workload using physiological measures: a systematic review [Journal] // Applied Ergonomics. - 2019. - Vol. 74. - pp. 221-232.
**Chi DL, Pickrell JE and Riedy CA** Student learning outcomes associated with video vs. paper cases in a public health dentistry course [Journal] // Journal of Dental Education. - 2014. - Vol. 78. - pp. 24-30.

**Clements AD** Salivary cortisol measurement in developmental research: where do we go from here? [Journal] // Developmental Psychobiology. - 2013. - Vol. 55. - pp. 205-220.
**Coles TR, John NW and Gould DA** Integrating Haptics with Augmented Reality in a Femoral Palpation and Needle Insertion Training Simulation [Conference] // IEEE TRANSACTIONS ON HAPTICS. - 2011. - Vol. 4. - pp. 199-209.
**Cook DA [et al.]** Measuring achievement goal motivation, mindsets and cognitive load: validation of three instruments’ scores [Journal] // Medical Education . - 2017. - Vol. 51. - pp. 1061–1074.
**Curtis M. T., DiazGranados D and Feldman M** Judicious Use of Simulation Technology in Continuing Medical Education [Journal] // Journal of Continuing Education in the Health Professions. - 2012. - Vol. 32. - pp. 255-260.
**Dawson S. L. and Kaufman J. A.** The imperative for medical simulation [Conference] // Proceedings of the IEEE. - 1998. - Vol. 86. - pp. 479-483.
**De Waard D** The Measurement of Drivers’ Mental Workload, PhD Thesis [Report]. - [s.l.] : The Traffic Research Centre VSC, University of Groningen, The Netherlands, 1996.
**DeMaria S [et al.]** Adding emotional stressors to training in simulated cardiopulmonary arrest enhances participant performance [Journal] // Medical education. - 2010. - Vol. 44. - pp. 1006-1015.
**Di Stasi LL [et al.]** Quantifying the cognitive cost of laparo-endoscopic single-site surgeries: gaze-based indices [Journal] // Applied Ergonomics. - 2017. - Vol. 65. - pp. 168–174.
**Dias RD [et al.]** Physiological synchronization and entropy as measures of team cognitive load [Journal] // Journal of Biomedical Informatics. - 2019. - Vol. 96.
**Dias RD [et al.]** Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload [Journal] // British Journal of Surgery. - 2018. - Vol. 105. - pp. 491-501.
**Elshama S. S.** How to apply Simulation-Based Learning in Medical Education? [Journal] // IberoAmerican Journal of Medicine. - 2020. - Vol. 02. - pp. 79-86.
**Endedijka M [et al.]** Using sensor technology to capture the structure and content of team interactions in medical emergency teams during stressful

moments [Journal] // Frontline Learning Research. - 2018. - Vol. 6. - pp. 123-147.
**Evain JN [et al.]** Residual anxiety after high fidelity simulation in anaesthesiology: An observational, prospective, pilot study [Journal] // Anaesthesia Critical Care & Pain Medicine. - 2017. - Vol. 36. - pp. 205-212.
**Feldman Barrett L and Russell JA** Independence and bipolarity in the structure of current affect [Journal] // Journal of Personality and Social Psychology. - 1998. - Vol. 74. - pp. 967–984.
**François C [et al.]** Tests of a New Drowsiness Characterization and Monitoring System Based on Ocular Parameters [Journal] // International Journal of Environmental Research and Public Health. - 2016. - 174 : Vol. 13.
**Fraser K [et al.]** Emotion, cognitive load and learning outcomes during simulation training [Journal] // Medical Education. - 2012. - Vol. 46. - pp. 1055–1062.
**Fraser K, Huffman J and Ma I** The Emotional and Cognitive Impact of Unexpected Simulated Patient Death [Journal] // Chest Journal. - 2014. - Vol. 145. - pp. 958-963.
**Garzón J., Pavón J. and Baldiris S.** Augmented Reality Applications for Education: Five Directions for Future Research [Conference] // Augmented Reality, Virtual Reality, and Computer Graphics. AVR 2017. Lecture Notes in Computer Science. - 2017. - Vol. 10324. - pp. 402-414.
**Gaumard** Advanced patient [Online]. - https://www.gaumard.com/s402-100-smash-advanced-patient-iv-training-arm.
**Geeraerts T [et al.]** Physiological and self-assessed psychological stress induced by a high fidelity simulation course among third year anesthesia and critical care residents: An observational study [Journal] // Anaesthesia Critical Care & Pain Medicine. - 2017. - Vol. 36. - pp. 403-406.
**Gerup J, Soerensen C. B. and Dieckmann P** Augmented reality and mixed reality for healthcare education beyond surgery: an integrative review [Journal] // International Journal of Medical Education. - 2020. - Vol. 11. - pp. 1-18.
**Goldberg MB [et al.]** Optimizing performance through stress training - An educational strategy for surgical residents [Journal] // American Journal of Surgery. - 2018. - Vol. 216. - pp. 618-623.

**Gouin A [et al.]** Evolution of stress in anaesthesia registrars with repeated simulated courses: An observational study [Journal] // Anaesthesia Critical Care & Pain Medicine. - 2017. - Vol. 36. - pp. 21-26.
**Green TRG and Hoc JM** WHAT IS COGNITIVE ERGONOMICS? [Journal] // Le Travail Humain/COGNITIVE ERGONOMICS: multidisciplinary stakes. - 1991. - Vol. 54. - pp. 291-304.
**Gutierrez-Puerto E [et al.]** Augmented Reality Central Venous Access Training Simulator [Conference] // HCI International 2015 - Posters' Extended Abstracts - Communications in Computer and Information Science. - 2015. - Vol. 528.
**Haji FA [et al.]** Measuring cognitive load during simulation-based psychomotor skills training: sensitivity of secondary task [Journal] // Advances in Health Science Education: Theory and Practice. - 2015. - Vol. 20. - pp. 1237-1253.
**Haji FA [et al.]** Measuring cognitive load: performance, mental effort and simulation task complexity [Journal] // Medical Education. - 2015. - Vol. 49. - pp. 815–827.
**Haji FA [et al.]** Thrive or overload? The effect of task complexity on novices' simulation-based learning [Journal] // Medical Education. - 2016. - Vol. 50. - pp. 955-968.
**Hanshaw S. L. and Dic S. S.** High fidelity simulation evaluation studies in nursing education: A review of the literature [Journal] // Nurse Education in Practice. - 2020. - Vol. 46.
**Hart SG and Staveland LE** Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research [Journal] // Advances in Psychology. - 1988. - Vol. 52. - pp. 139-183.
**Hart SG** Nasa-task load index (NASA-TLX); 20 years later [Conference] // In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. - 2006. - Vol. 50. - pp. 904–908.
**Helton WS** Validation of a short stress state questionnaire [Conference] // Proceedings of the Human Factors and Ergonomics Society Annual Meeting. - 2004. - Vol. 48. - pp. 1238–1242.
**Herron J.** Augmented Reality in Medical Education and Training [Journal] // Journal of Electronic Resources in Medical Libraries. - 2016. - Vol. 13. - pp. 51-55.
**Holzinger A [et al.]** Learning performance with interactive simulations in medical education: Lessons learned from results of learning complex

physiological models with the HAEMOdynamics SIMulator [Journal] // Computers & Education. - 2009. - Vol. 52. - pp. 292–301.
**Hunt E. A., Nelson K. L. and Shilkofski N. A.** Simulation in Medicine: Addressing Patient Safety and Improving the Interface Between Healthcare Providers and Medical Technology [Journal] // Biomedical Instrumentation &Technology. - 2006. - Vol. 40. - pp. 399–404.
**Hunziker S [et al.]** Dynamics and association of different acute stress markers with performance during a simulated resuscitation [Journal] // Resuscitation. - 2012. - Vol. 83. - pp. 572-578.
**ISO 10075-1** Ergonomic principles related to mental workload - Part 1: General issues and concepts, terms and definitions [Journal]. - 2017.
**ISO 10075-3** Ergonomic principles related to mental workload // Part 3: Principles and requirements concerning methods for measuring and assessing mental workload. - 2004.
**ISO 9241-210** Ergonomics of human system interaction - Part 210: Human-centered design for interactive systems. - 2010.
**Issenberg S. B. [et al.]** Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review [Journal] // Medical Teacher. - 2005. - Vol. 27. - pp. 10-28.
**Julian LJ** Measures of anxiety: State-Trait Anxiety Inventory (STAI), Beck Anxiety Inventory (BAI), and Hospital Anxiety and Depression Scale-Anxiety (HADS-A) [Journal] // Arthritis Care & Research. - 2011. - Vol. 63. - pp. S467-472.
**Kalakoski V [et al.]** Cognitive ergonomics for data analysis. Experimental study of cognitive limitations in a data-based judgement task [Journal] // Behaviour & Information Technology . - 2019. - Vol. 38. - pp. 1038-1047.
**Karwowski W** International Encyclopedia of Ergonomics and Human Factors [Book]. - 2006.
**Kobayashi L [et al.]** Exploratory Application of Augmented Reality/Mixed Reality Devices for Acute Care Procedure Training [Journal] // Educational Advances. - 2017. - pp. 158-164.
**Kotranza A [et al.]** Mixed Reality Humans: Evaluating Behavior, Usability, and Acceptability [Journal] // IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS. - 2009. - Vol. 15. - pp. 369-382.
**Kotranza A [et al.]** Real-Time In-Situ Visual Feedback of Task Performance in Mixed Environments for Learning Joint Psychomotor-

Cognitive Tasks [Conference] // IEEE International Symposium on Mixed and Augmented Reality 2009 - Science and Technology Proceedings. - 2009. - pp. 125-134.
**Kotranza A, Scott Lind D and Lok B** Real-Time Evaluation and Visualization of Learner Performance in a Mixed-Reality Environment for Clinical Breast Examination [Journal] // IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS. - 2012. - Vol. 18. - pp. 1101-1114.
**Koziol LF and Budding DE** Procedural Learning // Encyclopedia of the Sciences of Learning. - 2012.
**Lazarus RS and Folkman S** Stress, appraisal, and coping [Book]. - [s.l.] : Springer Pub. Co., 1984.
**LeBlanc VR [et al.]** Examination stress leads to improvements on fundamental technical skills for surgery [Journal] // The American Journal of Surgery. - 2008. - Vol. 196. - pp. 114-119.
**Lee D [et al.]** Augmented Reality to Localize Individual Organ in Surgical Procedure [Journal] // Healthcare Informatics Research. - 2018. - Vol. 24. - pp. 394-401.
**Lee JY [et al.]** How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking [Journal] // Computers in Human Behavior . - 2019. - Vol. 99. - pp. 268–277.
**Lesage FX, Berjot S and Deschamps F** Clinical stress assessment using a visual analogue scale [Journal] // Occupational Medicine. - 2012. - Vol. 62. - pp. 600-605.
**Linde A.S. and Geoffrey T.M.** Applications of Future Technologies to Detect Skill Decay and Improve Procedural Performance [Journal] // Military Medicine. - 2019. - Vol. 184. - pp. 72-77.
**Luque-Casado A [et al.]** Heart rate variability and cognitive processing: The autonomic response to task demands [Journal] // Biological Psychology. - 2016. - Vol. 113. - pp. 83–90.
**Magee D, Zhu Y and Ratnalingam R** An augmented reality simulator for ultrasound guided needle placement training [Journal] // Medical & Biological Engineering & Computing. - 2007. - Vol. 45. - pp. 957–967.
**Margarido Mendes HC [et al.]** PI˜NATA: Pinpoint insertion of intravenous needles via augmented reality training assistance [Journal] // Computerized Medical Imaging and Graphics. - 2020. - Vol. 82.

**Marjanovic NS [et al.]** Evaluation of emotional excitation during standardized endotracheal intubation in simulated conditions [Journal] // Annals of Intensive Care. - 2018.
**Mehta V [et al.]** Applications of Augmented Reality in Emerging Health Diagnostics: A Survey [Conference] // International Conference on Automation and Computational Engineering (ICACE - 2018). - 2018. - pp. 45-51.
**Meola A. [et al.]** Augmented reality in neurosurgery: a systematic review [Journal] // Neurosurgical Review. - 2017. - Vol. 40. - pp. 537–548.
**Mercer S. J.** Using full immersive simulation to prepare trauma teams to work in a major trauma centre [Journal] // Trauma. - 2017. - Vol. 19. - pp. 277–285.
**Mikhail M, Mithani K and Ibrahim G. M.** Presurgical and Intraoperative Augmented Reality in Neuro-Oncologic Surgery: Clinical Experiences and Limitations [Journal] // World Neurosurgery. - 2019. - Vol. 128. - pp. 268-276.
**Milgram P [et al.]** Augmented Reality: A class of displays on the reality-virtuality continuum [Conference] // Proceedings Telemanipulator and Telepresence Technologies. - 1994. - Vol. 2351. - pp. 282-292.
**Miller EK and Cohen JD** An integrative theory of prefrontal cortex function [Journal] // Annula Review of Neuroscience. - 2001. - Vol. 24. - pp. 167-202.
**Mitchell AM, Crane PA and Kim Y** Perceived stress in survivors of suicide: psychometric properties of the Perceived Stress Scale [Journal] // Research in Nursing & Health. - 2008. - Vol. 31. - pp. 576-585.
**Müller M [et al.]** Excellence in performance and stress reduction during two different full scale simulator training courses: a pilot study [Journal] // Resuscitation. - 2009. - Vol. 80. - pp. 919-924.
**Munzer B.W. [et al.]** Augmented Reality in Emergency Medicine: A Scoping Review [Journal] // Journal of Medical Internet Research. - 2019. - Vol. 21.
**Muresan C [et al.]** Transfer of training in the development of intracorporeal suturing skill in medical student novices: a prospective randomized trial [Conference] // In: 11th World Congress of Endoscopic Surgery, Japan. - 2008. - pp. 537–541.

**Naismith LM [et al.]** Limitations of subjective cognitive load measures in simulation-based procedural training [Journal] // Medical Education. - 2015. - Vol. 49. - pp. 805–814.
**Naismith LM and Cavalcanti RB** Validity of Cognitive Load Measures in Simulation-Based Training: A Systematic Review [Journal] // Academic Medicine. - 2015. - Vol. 90. - pp. S24–S35.
**Nasco** Products [Online]. - https://www.enasco.com/p/TruMan-Trauma-System%2BSB51022?searchText=thoracentesis.
**Nausheen F and Bhupathy R** Disrupting clinical education: point-of-care ultrasound merged with HoloLens augmented reality during early medical training [Journal] // The Clinical Teacher. - 2020. - Vol. 17. - pp. 146-147.
**Norman D** Cognitive engineering [Book]. - 1987.
**Norman G, Dore K and Grierson L** The minimal relationship between simulation fidelity and transfer of learning [Journal] // Medical Education. - 2012. - Vol. 46. - pp. 636–647.
**Paul G and Scataglini S** Open-source software to create a kinematic model in digital human modeling [Journal] // DHM and Posturography. - 2019. - pp. 201-213.
**Pheasant S** Body Space: Anthropometry, Ergonomics and the Design of Work [Book]. - [s.l.] : Taylor & Francis, 1999. - pp. 121-123.
**Pottier P [et al.]** Effect of extrinsic and intrinsic stressors on clinical skills performance in third-year medical students [Journal] // Journal of General Internal Medicine. - 2015. - Vol. 30. - pp. 1259-1269.
**Pruessner JC [et al.]** Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change [Journal] // Psychoneuroendocrinology. - 2003. - Vol. 28. - pp. 916–931.
**Rochlen LR, Levine R and Tait AR** First-Person Point-of-View-Augmented Reality for Central Line Insertion Training: A Usability and Feasibility Study [Journal] // Simulation in Healthcare. - 2017. - Vol. 12. - pp. 57-62.
**Roscoe AH** Assessing pilot workload. Why measure heart rate, HRV and respiration? [Journal] // Biological Psychology. - 1992. - Vol. 34. - pp. 259-287.
**Ruddy R. M. and Patterson M. D.** Medical simulation: a tool for recognition of and response to risk [Journal] // Pediatric Radiology. - 2008. - Vol. 38. - pp. S700–S706.

**Sahu S and Lata I** Simulation in resuscitation teaching and training: an evidence-based practice review [Journal] // Journal of Emergencies, Trauma and Shock. - Oct-Dec 2010. - 4 : Vol. 3. - pp. 378–384.
**Sandroni C [et al.]** Haemodynamic effects of mental stress during cardiac arrest simulation testing on advanced life support courses [Journal] // Resuscitation. - 2005. - Vol. 66. - pp. 39-44.
**Sawyer T and Gray M. M.** Procedural training and assessment of competency utilizing simulation [Journal] // Seminars inPerinatology. - 2016. - Vol. 40. - pp. 438-446.
**Scafà M [et al.]** Assessment of Students' Cognitive Conditions in Medical Simulation Training: A Review Study [Conference] // In: Cassenti D. (eds) Advances in Human Factors and Simulation. AHFE 2019. Advances in Intelligent Systems and Computing. - 2019. - Vol. 958. - pp. 224–233.
**Scataglini S [et al.]** Moving Statistical Body Shape Models Using Blender [Conference] // IEA 2018, Ad-vances in Intelligent Systems and Computing. - 2019. - Vol. 822. - pp. 28-38.
**Schmalstieg D and Höllerer T** Augmented reality: Principles and practice [Book]. - 2016.
**Seagull FJ** Human Factors Tools For Improving Simulation Activities In Continuing Medical Education [Journal] // Journal Of Continuing Education In The Health Professions. - 2012. - Vol. 32. - pp. 261–268.
**Sewell JL [et al.]** Trainee perception of cognitive load during observed faculty staff teaching of procedural skills [Journal] // Medical Education. - 2019. - Vol. 53. - pp. 925-940.
**Sherstyuk A [et al.]** Mixed Reality Manikins for Medical Education [Book Section] // Handbook of Augmented Reality. - [s.l.] : DOI:10.1007/978-1-4614-0064-6_23, 2011.
**Sigalet E, Donnon T and Grant V** Undergraduate students' perceptions of and attitudes toward a simulation-based interprofessional curriculum: the KidSIM ATTITUDES questionnaire [Journal] // Simulation in Healthcare. - 2012. - Vol. 7. - pp. 353-358.
**Singh S [et al.]** A cross-sectional assessment of stress, coping, and burnout in the final year medical undergraduate students [Journal] // Indian Psychiatry Journal. - 2016. - Vol. 25. - pp. 179-183.
**Spielberger CD and Gorsuch RL** State-trait anxiety inventory for adults : Sampler set: manual, test, scoring key [Book]. - [s.l.] : Mind Garden, Redwood City, Calif, 1983.

**Stawarczy D [et al.]** Drowsiness or mind-wandering? Fluctuations in ocular parameters during attentional lapses [Journal] // Biological Psychology. - 2020. - 107950 : Vol. 156.
**Stramler J.H.** The Dictionary for Human Factors/Ergonomics [Book]. - [s.l.] : CRC Press, Boca Raton, 1993.
**Stretton T, Cochrane T and Narayanb V** Exploring mobile mixed reality in healthcare higher education: A systematic review [Journal] // Research in Learning Technology. - 2018. - Vol. 26.
**Sugarindra M, Suryoputro MR and Permana AI** Mental workload measurement in operator control room using NASA-TLX [Conference] // IOP Conf. Series: Materials Science and Engineering. - 2017. - Vol. 277.
**Sweller J, van Merrienboer JJG and Paas FGWC** Cognitive Architecture and Instructional Design [Journal] // Educational Psychology Review. - 1998. - Vol. 10. - pp. 251-296.
**Szu-Chi Chen S and Duh HBL** Mixed Reality in Education: Recent Developments and Future Trends [Conference] // IEEE 18th International Conference on Advanced Learning Technologies. - 2018. - pp. 367-371.
**Szulewski A and Howes D** Combining First-Person Video and Gaze-Tracking in Medical Simulation: A Technical Feasibility Study [Journal] // The Scientific World Journal. - 2014.
**Tai Y, Wei L and Zhou H** Augmented-reality-driven medical simulation platform for percutaneous nephrolithotomy with cybersecurity awareness [Journal] // International Journal of Distributed Sensor Networks. - 2019. - Vol. 15.
**Takaia N [et al.]** Effect of psychological stress on the salivary cortisol and amylase levels in healthy young adults [Journal] // Archives of Oral Biology. - 2004. - Vol. 49. - pp. 963-968.
**Theodoraki MN [et al.]** Mental distress and effort to engage an image-guided navigation system in the surgical training of endoscopic sinus surgery: a prospective, randomised clinical trial [Journal] // European Archives of Oto-Rhino-Laryngology. - 2015. - Vol. 272. - pp. 905-913.
**Tomasko JM [et al.]** Sleep deprivation increases cognitive workload during simulated surgical tasks [Journal] // The American Journal of Surgery. - 2012. - Vol. 203. - pp. 37-43.
**Tremblay M. L. [et al.]** Simulation-based education for novices: complex learning tasks promote reflective practice [Journal] // Medical Education. - 2019. - Vol. 53. - pp. 380-389.

**Van Merrienboer Jeroen J G and Sweller J** Cognitive load theory in health professional education: design principles and strategies [Journal] // Medical Education. - 2010. - Vol. 44. - pp. 85-93.
**Vincent-Lambert C and Bogossian F** A guide for the assessment of clinical competence using simulation [Book]. - [s.l.] : U21 Health Sciences Group, 2017.
**Waller L [et al.]** Experiential learning as preparation for leadership: an exploration of the cognitive and physiological processes [Journal] // Leadership & Organization Development Journal. - 2017. - Vol. 38. - pp. 513-529.
**Wang S [et al.]** Augmented Reality as a Telemedicine Platform for Remote Procedural Training [Journal] // Sensors. - 2017. - Vol. 10.
**Webb A. K. [et al.]** A Systems Approach for Augmented Reality Design [Conference] // Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience. AC 2016. Lecture Notes in Computer Science. - 2016. - Vol. 9744. - pp. 382-389.
**WHO World Health Organization** Data and Statistics [Online] // Regional Office for Europe. - 2017. - http://www.euro.who.int/en/health-topics/Health-systems/patient-safety/data-and-statistics.
**Wickens CD, Gordon SE and Liu Y** An introduction to human factors engineering [Journal] // Upper Saddle River. - 2004. - pp. 64–90.
**Williams N** The borg rating of perceived exertion (RPE) scale [Journal] // Occupational Medicine. - 2017. - Vol. 67. - pp. 404-405.
**Wilson JR** Fundamentals of ergonomics in theory and practice [Journal] // Applied Ergonomics. - 2000. - Vol. 31. - pp. 557-567.
**Wilson MR [et al.]** Development and validation of a surgical workload measure: the surgery task load index (SURG-TLX) [Journal] // World Journal of Surgery. - 2011. - Vol. 35. - pp. 1961–1969.
**Woods B, Byrne A and Bodger O** The effect of multitasking on the communication skill and clinical skills of medical students [Journal] // BMC Medical Education. - 2018. - Vol. 18.
**Wucherer P [et al.]** Vertebroplasty Performance on Simulator for 19 Surgeons Using Hierarchical Task Analysis [Journal] // IEEE TRANSACTIONS ON MEDICAL IMAGING. - 2015. - Vol. 34. - pp. 1730-1737.

**Yerkes RM and Dodson JD** The relation of strength of stimulus to rapidity of habit formation [Journal] // Journal of Comparative Neurology. - 1908. - Vol. 18. - pp. 459–482.
**Zhu E. [et al.]** Augmented reality in healthcare education: an integrative review [Journal] // PeerJ. - 2014.

# Appendix A. State-Trait Anxiety Inventory

## A.1. STAI Form Y-1 (State)

**SELF-EVALUATION QUESTIONNAIRE** **STAI Form Y-1**

**Please provide the following information:**

Name______________________________ Date______________ S_____

Age______________ Gender (*Circle*) **M F** T_____

**DIRECTIONS:**

A number of statements which people have used to describe themselves are given below. Read each statement and then blacken the appropriate circle to the right of the statement to indicate how you feel *right* now, that is, *at this moment*. There are no right or wrong answers. Do not spend too much time on any one statement but give the answer which seems to describe your present feelings best.

| | NOT AT ALL | SOMEWHAT | MODERATELY SO | VERY MUCH SO |
|---|---|---|---|---|
| 1. I feel calm | 1 | 2 | 3 | 4 |
| 2. I feel secure | 1 | 2 | 3 | 4 |
| 3. I am tense | 1 | 2 | 3 | 4 |
| 4. I feel strained | 1 | 2 | 3 | 4 |
| 5. I feel at ease | 1 | 2 | 3 | 4 |
| 6. I feel upset | 1 | 2 | 3 | 4 |
| 7. I am presently worrying over possible misfortunes | 1 | 2 | 3 | 4 |
| 8. I feel satisfied | 1 | 2 | 3 | 4 |
| 9. I feel frightened | 1 | 2 | 3 | 4 |
| 10. I feel comfortable | 1 | 2 | 3 | 4 |
| 11. I feel self-confident | 1 | 2 | 3 | 4 |
| 12. I feel nervous | 1 | 2 | 3 | 4 |
| 13. I am jittery | 1 | 2 | 3 | 4 |
| 14. I feel indecisive | 1 | 2 | 3 | 4 |
| 15. I am relaxed | 1 | 2 | 3 | 4 |
| 16. I feel content | 1 | 2 | 3 | 4 |
| 17. I am worried | 1 | 2 | 3 | 4 |
| 18. I feel confused | 1 | 2 | 3 | 4 |
| 19. I feel steady | 1 | 2 | 3 | 4 |
| 20. I feel pleasant | 1 | 2 | 3 | 4 |

# A.2. STAI Form Y-2 (Trait)

**SELF-EVALUATION QUESTIONNAIRE**

**STAI Form Y-2**

Name________________________________________________Date__________

**DIRECTIONS**

A number of statements which people have used to describe themselves are given below. Read each statement and then blacken in the appropriate circle to the right of the statement to indicate you *generally* feel.

| | ALMOST NEVER | SOMETIMES | OFTEN | ALMOST ALWAYS |
|---|---|---|---|---|
| 21. I feel pleasant | 1 | 2 | 3 | 4 |
| 22. I feel nervous and restless | 1 | 2 | 3 | 4 |
| 23. I feel satisfied with myself | 1 | 2 | 3 | 4 |
| 24. I wish I could be as happy as others seem to be | 1 | 2 | 3 | 4 |
| 25. I feel like a failure | 1 | 2 | 3 | 4 |
| 26. I feel rested | 1 | 2 | 3 | 4 |
| 27. I am "calm, cool, and collected" | 1 | 2 | 3 | 4 |
| 28. I feel that difficulties are piling up so that I cannot overcome them | 1 | 2 | 3 | 4 |
| 29. I worry too much over something that really doesn't matter | 1 | 2 | 3 | 4 |
| 30. I am happy | 1 | 2 | 3 | 4 |
| 31. I have disturbing thoughts | 1 | 2 | 3 | 4 |
| 32. I lack self-confidence | 1 | 2 | 3 | 4 |
| 33. I feel secure | 1 | 2 | 3 | 4 |
| 34. I make decisions easily | 1 | 2 | 3 | 4 |
| 35. I feel inadequate | 1 | 2 | 3 | 4 |
| 36. I am content | 1 | 2 | 3 | 4 |
| 37. Some unimportant thought runs through my mind and bothers me | 1 | 2 | 3 | 4 |
| 38. I take disappointments so keenly that I can't put them out of my mind | 1 | 2 | 3 | 4 |
| 39. I am a steady person | 1 | 2 | 3 | 4 |
| 40. I get in a state of tension or turmoil as I think over my recent concerns and interests | 1 | 2 | 3 | 4 |

# Appendix B. NASA Task Load Index

## B.1. NASA-TLX Sources of Workload Comparison Card

| | | |
|---|---|---|
| EFFORT or PERFORMANCE | TEMPORAL DEMAND or FRUSTRATION | FRUSTRATION or EFFORT |
| PERFORMANCE or MENTAL DEMAND | TEMPORAL DEMAND or EFFORT | PHYSICAL DEMAND or FRUSTRATION |
| PERFORMANCE or TEMPORAL DEMAND | MENTAL DEMAND or EFFORT | PERFORMANCE or FRUSTRATION |
| PHYSICAL DEMAND or TEMPORAL DEMAND | MENTAL DEMAND or PHYSICAL DEMAND | EFFORT or PHYSICAL DEMAND |
| PHYSICAL DEMAND or PERFORMANCE | TEMPORAL DEMAND or MENTAL DEMAND | FRUSTRATION or MENTAL DEMAND |

# B.2. NASA-TLX Rating Sheet

*Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.*

| Name | Task | Date |
|---|---|---|
| | | |

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

# Appendix C. Statistical Variables Legend

| | VARIABLES | DESCRIPTION |
|---|---|---|
| Variables collected once | GENDER | Male or female |
| | BIRTH | Year of birth of the participant |
| | WEIGHT | Weight of the participant |
| | HEIGHT | Height of the participant |
| | SMOKER | If the participant smokes or not |
| | QUALIFICATION | Degree of education |
| | EXPERIENCE | If the participant has already performed group simulations, residency, professional training activity, training courses (BLS, ATLS, …), working experiences, health volunteer activities, invasive procedures, and how many of them |
| | STAI TRAIT | Total score per participant to the STAI TRAIT questionnaire |
| Variables collected twice (for HF and LF simulations) | MENSTR | Last menstruation date |
| | THESIS | If the Department where the participant does the thesis is relevant for the simulation |
| | DATE | Date of simulation |
| | SESSION | Simulation performed during morning or afternoon |
| | STAI PRE | Total score per participant to the STAI STATE questionnaire, answered before simulation |
| | STAI POST | Total score per participant to the STAI STATE questionnaire, answered after simulation |
| | Δ STAI | Difference between STAI POST and STAI PRE, per participant |
| | NAS 0 | Score per participant to NAS question answered at the arrival in the simulation room |
| | NAS 1 | Score per participant to NAS question answered 10 minutes after NAS 0 |

| | | |
|---|---|---|
| | NAS 2 | Score per participant to NAS question answered 10 minutes after the end of simulation |
| | NAS 3 | Score per participant to NAS question answered 20 minutes after the end of simulation |
| | NAS 4 | Score per participant to NAS question answered 30 minutes after the end of simulation |
| | Δ NAS (2-1) | Difference between NAS 2 and NAS 1 per participant |
| | Δ NAS (4-1) | Difference between NAS 4 and NAS 1 per participant |
| | NASA-TLX | Total score per participant of NASA-TLX questionnaire |
| | NASA MENTAL | Weight assigned from each participant to the item "mental demand" in NASA-TLX |
| | NASA PHYSICAL | Weight assigned from each participant to the item "physical demand" in NASA-TLX |
| | NASA TEMPORAL | Weight assigned from each participant to the item "temporal demand" in NASA-TLX |
| | NASA PERFOR | Weight assigned from each participant to the item "performance" in NASA-TLX |
| | NASA EFFORT | Weight assigned from each participant to the item "effort" in NASA-TLX |
| | NASA FRUSTR | Weight assigned from each participant to the item "frustration" in NASA-TLX |
| | STRESS SIM | Mean level of STRESS measured through physiological signals during simulation |
| | CL SIM | Mean level of Cognitive Load measured through physiological signals during simulation |
| | GROUP | To which one of the 3 different sessions of simulation of increasing |

| | | |
|---|---|---|
| Variables collected only for HFS | | difficulty, the student has participated to |
| | SIM ATT PRE | Total result per participant of the questionnaire about the attitude to simulation and team working, answered before the simulation |
| | SIM ATT POST | Total result per participant of the questionnaire about the attitude to simulation and team working, answered after the simulation |
| | Δ SIM ATT | Difference between SIM ATT POST and SIM ATT PRE, per participant |
| | TOTAL PERFORMANCE | Total performance score assigned to each HF scenario by the trainer |
| | PERF PREP | Performance score of preparation phase assigned to each HF scenario by the trainer |
| | PERF A | Performance score of phase A assigned to each HF scenario by the trainer |
| | PERF B | Performance score of phase B assigned to each HF scenario by the trainer |
| | PERF C | Performance score of phase C assigned to each HF scenario by the trainer |
| | PERF D | Performance score of phase D assigned to each HF scenario by the trainer |
| | PERF E | Performance score of phase E assigned to each HF scenario by the trainer |
| | STRESS PREP | Mean level of STRESS measured through physiological signals during preparation phase |
| | STRESS A | Mean level of STRESS measured through physiological signals during phase A |
| | STRESS D | Mean level of STRESS measured through physiological signals during phase D |

| | | |
|---|---|---|
| | STRESS DEB | Mean level of STRESS measured through physiological signals during debriefing |
| | CL PREP | Mean level of CL measured through physiological signals during preparation phase |
| | CL A | Mean level of CL measured through physiological signals during phase A |
| | CL D | Mean level of CL measured through physiological signals during phase D |
| | CL DEB | Mean level of Cognitive Load measured through physiological signals during debriefing |
| | Fold Increase T2/T1 | Cortisol variation between basal and peak measurements |
| | AUCi | Area under the curve with respect to increase (Cortisol measurement) |
| | Cortisol 2 | Salivary cortisol measured 10 minutes after the end of the simulation |
| Variables collected only for LFS | GROUP | To which one of the 2 different group of simulation the student has participated to |
| | SEQUENCE | In which order participants performed the simulation (first, second, third, fourth, ..) |
| | TECH ATT PRE | Total result per participant of the questionnaire about the attitude to the use of technologic devices, answered before the simulation |
| | TECH ATT POST | Total result per participant of the questionnaire about the attitude to the use of technologic devices, answered after the simulation |
| | Δ TECH ATT | Difference between TECH ATT POST and TECH ATT PRE, per participant |
| | SKILLS PRE | Competence of the participant about lumbar puncture before the simulation |
| | SKILLS POST | Competence of the participant about lumbar puncture after the simulation |

| | | |
|---|---|---|
| | Δ SKILLS | Difference between SKILLS POST and SKILLS PRE, per participant |
| | PREP TIME | Time the participant took to from the beginning of the simulation to the insertion of the needle |
| | SUCCESS TIME | Time the participant took from the insertion of the needle to the pouring out of the liquor |
| | TOT TIME | Total time of the simulation for each participant |
| | SUCCESS | If the participant succeeds in liquor collection or not |
| | ATTEMPTS | Number of attempts for each participant to collect the liquor |
| | PREP PERF | Performance in the preparatory phase |
| | POST PREP PERF | Performance in the lumbar puncture |
| | Total Performance | Total performance for each participant |
| | ERRORS | Number of errors for each participant |
| | TEACHER | If the teacher helped or not the participant during the simulation |
| | STRESS PREP | Mean level of STRESS measured through physiological signals during preparatory phase |
| | STRESS LP | Mean level of STRESS measured through physiological signals during lumbar puncture |
| | STRESS END | Mean level of STRESS measured through physiological signals after lumbar puncture |
| | CL PREP | Mean level of Cognitive Load measured through physiological signals during preparatory phase |
| | CL LP | Mean level of Cognitive Load measured through physiological signals during lumbar puncture |
| | CL END | Mean level of Cognitive Load measured through physiological signals after lumbar puncture |