Differences in visual behavior and cognitive load between experts and novices in medical emergency simulation

Master Thesis

For attainment of the academic degree of
Master of Science in Engineering (MSc)

in the Master Programme Digital Healthcare
at St. Pölten University of Applied Sciences

by
Dr. med. Univ. Christoph Franz Plank
1042290

First advisor: Dr. rer. nat. Vanessa Leung
Second advisor: Dr. habil. Roman Schmied

Gmunden, 11.09.2024

Declaration of Honour

I hereby declare that

- I have written the work at hand on my own without help from others and I have used no other resources and tools than the ones acknowledged.
- I have complied with the Standards of good scientific practice in accordance with the St. Pölten UAS' Guidelines for Scientific Work when writing this work.
- I have neither published nor submitted the work at hand to another higher education institution for assessment or in any other form as examination work.

Regarding the use of generative artificial intelligence tools such as chatbots, image generators, programming applications, paraphrasing and translation tools, I declare that

☐ no generative artificial intelligence tools were used in the course of this work.

☐ I have used generative artificial intelligence tools to proof-read this work.

☒ I have used generative artificial intelligence tools to create parts of the content of this work. I certify that I have cited the original source of any generated content. The generative artificial intelligence tools that I used are acknowledged at the respective positions in the text.

Having read and understood the St. Pölten UAS' Guidelines for Scientific Work, I am aware of the consequences of a dishonest declaration.

Preface

As a medical professional, lifelong learning and adapting to changing circumstances is essential. In today's rapidly evolving world, this is truer than ever. During my Digital Healthcare journey, I was fortunate to acquire new tools to meet this challenge. From the outset, I asked myself how I could use my knowledge to make a positive impact on others. With a deep passion for teaching, and after a semester under the guidance of my advisor Vanessa Leung, who introduced us to the potential of eye tracking and biosignal measurement, the idea for the "Simtrack" simulation study was born. Its ultimate aim: to enhance the learning of medical professionals and improve patient outcomes. While it's too early to tell if this goal will be fully realized, being part of this journey has been an incredibly rewarding experience.

Nonetheless, such a project is never a solo effort. I want to express my heartfelt gratitude to everyone involved. Nurses Daniela Beindling, Michael Höllwerth, Bernhard Ögger, and Monika Roither: thank you for your support and patience during two intense study days. Adrian Vulpe-Grigorasi: thank you for your dedication, especially for getting up at 4:30 a.m. and traveling to Upper Austria to assist with data collection. To all the participants: thank you for your enthusiasm and courage in having your performance recorded.

Special thanks to my thesis advisors, Vanessa Leung and Roman Schmied, for patiently answering my many questions, providing invaluable feedback, and offering constant guidance throughout the project.

Lastly, to my family and my girlfriend: thank you for enduring my moods, understanding my absence on numerous occasions, and offering your invaluable outside-the-box feedback. And my brother: For being the best sibling and study buddy one could hope for.

Abstract

Background: Medical simulation training has been shown to improve learning of medical professionals and outcome of patients. Research has thus long been trying to improve training even further to maximize this effect. With the advent of new technologies, such as eye tracking and biosignal measurement, this was taken to a more sophisticated level, providing insights into not only subjective experiences of trainees but also objectively measurable reactions of the body and mind.

Aim: This thesis aimed to evaluate visual behavior and cognitive load in a group of medical experts and novices during a simulated resuscitation scenario, and to provide a sample size estimate for future research and to assess feasibility of such an approach.

Methods: 7 novices and 5 experts – were included in this trial and underwent a standardized simulation scenario. Eye tracking glasses, a galvanic skin response (GSR) sensor and an electrocardiogram (ECG) chest strap were used to collect data. Additionally, NASA TLX and PAAS cognitive load questionnaires were completed by the participants after completion of simulation training. Data were then extracted and analyzed regarding visual behavior and cognitive load, comparing objective parameters such as velocity of eye movement and heart rate variability (HRV), with subjective experiences collected using the questionnaires mentioned above. Additionally, feasibility of this multi-sensor approach was also assessed, and sample size calculation was performed.

Outcome: Besides small differences, e.g. in duration of fixation during some parts of the scenario, visual behavior in general appeared mostly similar in both groups. HRV parameters and questionnaire results also did not differ greatly, suggesting a similar cognitive load. Differences were mostly found within groups during different parts of the simulation scenario, e.g. in HRV or saccadic velocity. Regarding study size, a size of N = 34 was estimated to provide sufficient power for statistical testing. The used multisensory approach was highly feasible and well accepted by probands, apart from GSR measurement due to strong motion artifacts.

Conclusion: A multisensor approach like in this work is highly feasible to collect data on cognitive load and visual behavior. Experts in a field do not necessarily seem to have a lower cognitive load. A study size of 34 participants was estimated to provide sufficient power for statistical testing.

Kurzfassung

Hintergrund: Medizinisches Simulationstraining verbessert nachweislich die Prognose von PatientInnen. Durch den Einsatz moderner Technologien wie Eye-Tracking und Biosignalmessungen können nun auch objektive physiologische Reaktionen erfasst werden.

Ziel: Ziel dieser Pilotstudie war es, das visuelle Verhalten und die kognitive Arbeitslast von medizinischen ExpertInnen und weniger erfahrenen MedizinerInnen während eines simulierten Notfallszenarios zu untersuchen. Zudem wurde eine Fallzahl für weiterführende Studien errechnet sowie die Machbarkeit dieser Methodik überprüft.

Methodik: Zwölf ProbandInnen – sieben ExpertInnen und fünf Neulinge – durchliefen ein Simulationsszenario, ausgestattet mit Eye-Tracking-Brillen, Sensoren zur Hautwiderstandsmessung und einem EKG-Brustgurt. Zusätzlich füllten sie Fragebögen zur subjektiven kognitiven Belastung aus. Die Daten des Blickverhaltens und der Herzfrequenz wurden zwischen den Gruppen und mit den subjektiven Angaben verglichen. Die Machbarkeit wurde ebenfalls mittels Fragebögen evaluiert.

Ergebnis: Es wurden kaum Unterschiede – mit Ausnahme zum Beispiel im Bereich der visuellen Fixationsdauer – im Blickverhalten zwischen ExpertInnen und Neulingen festgestellt. Diese traten eher innerhalb der Gruppen zu unterschiedlichen Zeitpunkten auf, beispielsweise in der Herzfrequenzvariabilität oder der Sakkadengeschwindigkeit. Auch die kognitive Belastung, gemessen durch Herzfrequenzvariabilität und Fragebögen, war ähnlich. Für weiterführende Studien wurde eine Stichprobengröße von 34 Teilnehmenden ermittelt. Der Multisensoransatz wurde durchweg positiv aufgenommen.

Conclusio: ExpertInnen und Neulinge unterscheiden sich nicht zwangsläufig in ihrem Blickverhalten und der kognitiven Last während eines Simulationstrainings. Zumindest 34 Personen sollten in eine weiterführende Studie eingeschlossen werden. Der hier verwendete Multisensoransatz wurde sehr gut angenommen.

Table of Content

1 Introduction

1.1 Problem

Becoming a medical professional comprises a (life)long way of learning, requiring the collection of experience. [1] Some of this experience is collected by treating actual patients, and some of it can and should be gained during simulated scenarios in a safe learning environment to avoid potentially life-threatening mistakes having an impact on an actual patient's life. [2]

For quite some time, research has tried to analyze how experts in a field make decisions and why and how the results of this study can be used to improve learning strategies for novices. [3].

Tackling this question, this thesis aimed to combine medical simulation and state-of-the-art technology, i.e. eye tracking as well as registration of galvanic skin response and heart rate, to analyze behavior during such a training scenario by assessment of gaze patterns and evaluation of cognitive load, that could be potentially used to improve novice learning.

1.2 Hypothesis and Method overview

Combining the aforementioned technologies, the working hypothesis that was to be evaluated was defined as follows:

Are there potential differences in visual behavior and cognitive load during medical simulation between novices and experts in the field?

To allow evaluation of this working statement, the following fundamental research questions were defined:

1) Do gaze patterns and areas of visual focus potentially differ between medical experts and medical novices during a simulated cardiac arrest training scenario?
2) Does cognitive load potentially differ between novices and experts during a simulation scenario?

3) What would be an appropriate sample size for further investigating differences in visual behavior and cognitive load?
4) Is the usage of eye tracking and cognitive load equipment feasible during such a scenario?

Regarding the method used to work on those questions, in brief, a prospective, observational pilot trial using two groups consisting of experts in the field of emergency medicine and emergency medicine novices, respectively, was designed. Both groups underwent a standardized pre-defined simulation scenario and eye tracking, GSR as well as heart rate data were collected. Subsequently, data were analyzed statistically. Also, feasibility of such a study setup was assessed using previously used questionnaires. [4], [5]

1.3 Motivation

The motivation to work on this topic derives from the personal learning and medical simulation experience of the author in combination with a growing interest in the medical teaching research community in the topic in general. [6] Summarized, by adding to the body of evidence in this field the ultimate goal is to provide a potential benefit to various groups:

1) The trainees themselves – novices as well as professionals by receiving a personalized first-person recapitulation of a simulation scenario.
2) Simulation / medical curriculum planners – by getting insights on what potential areas of education they might need to focus on in the future.
3) Patients – by potentially better chances of surviving medical emergencies due to better educated medical staff.
4) The healthcare system – by potentially profiting from more thoroughly trained personnel resulting in better decision-making processes and potential cost reduction by a decrease of unnecessary treatment/diagnostics/educational measures.
5) States/Society – by reduced economic loss due to individuals being more likely to survive medical emergencies and return to work life.

It shall be explicitly stated here that an attempt to translate the findings of the conducted study into a new learning strategy for young medical professionals was beyond the scope of this thesis, which merely, as already mentioned, aimed to add to the scientific foundation of future research in the field.

1.4 Frequently used terms and project outline

To allow a deeper understanding of the topic, the most relevant terms, which will appear repeatedly during this thesis, shall be defined here briefly:

- Cognitive load: Amount of workload that is posed on the human brain as an information processing system when dealing with information. [7]
- Cognitive load theory: Framework stating that, amongst others, human learning relies on so-called “working memory”, which processes incoming information. When the amount of cognitive workload exceeds the amount of working memory available, the learning effect diminishes. [8], [9]
- Galvanic skin response (GSR): “a change in the electrical properties (conductance or resistance) of the skin in reaction to stimuli, owing to the activity of sweat glands“ as defined by the American psychological association. [10]. Used as a surrogate parameter for cognitive load in research. [11], [12]
- Eye tracking: “the activity of studying the way that people's eyes move in order to discover what […] attracts their attention”. [13] This is performed using specialized devices, i.e. so-called “eye tracking glasses” to record human gaze and extract features such as gaze patterns, fixation frequencies, etc.

A more thorough explanation of the mentioned terms will be provided in the background section.

The remaining part of this thesis will be structured as follows:

- A background section providing in-detail information of the existing literature and work in the field.
- A section describing the requirements and methods necessary.
- An implementation section covering the actual translation of the designed study to a real-world simulation scenario and the achieved results.
- A thorough discussion of results in the context of the already existing research, weaknesses/potential bias as well as strengths of the used approach.

2 Background and Related Work

A broad spectrum of background information is necessary to spotlight the key aspects of this thesis, reaching from medical simulation, eye tracking technology, and human decision-making theories, as well as cognitive load, its quantification, and its influence on decisions and learning. Necessary facts will be laid out subsequently. As the respective subtopics are not necessarily obviously interwoven with each other, a summary integrating the fragments into a comprehensive concept will be provided at the end of this section.

2.1 Simulation in medical training

Over the last decades, medical simulation has become a fixed point in teaching medical skills, to certified medical professionals, such as doctors or nurses, as well as bystanders. [14], [15] What started with a low-fidelity model of a human torso in the 1960s – the publicly well-known “Resusci Anne” – has nowadays evolved into a highly complex setting involving simulation mannequins with an ever-growing set of features (Figure 1) and integration of e.g. augmented reality, machine learning and analysis of trainee’s vital parameters. [16], [17], [18] Simulation environments, where participants can learn crucial medical skills without direct consequences of potentially fatal errors for patients, now have become dedicated hospital units with state-of-the-art technology and a size of 1000m^2. [19]

With benefit on student learning and patient outcomes, as summarized in a review by Mundell et al. already in 2013, large medical societies, such as the European Council of Resuscitation (ERC), or the International Liaison Committee on Resuscitation (ILCOR), specifically recommend the use of simulation in teaching skills, in this case cardiopulmonary resuscitation (CPR). [20], [21] Furthermore, the regularly published ERC guidelines for resuscitation specifically dedicate a whole chapter to the usage of simulation, covering the recommended level of fidelity, the importance of “human factors” as well as the role of debriefing, which will be covered in more detail below. [20] In this context, it must be mentioned, that the evidence regarding the effect of high-fidelity techniques, e.g. high-end simulation units, in simulation is somewhat equivocal, with some work showing improved knowledge retention and skill performance in students, whereas other publications showed that skill performance at one year after a life support course was not superior in a high-fidelity setting in comparison to low-fidelity circumstances, or that high-fidelity could even lead to overestimation of knowledge gain while having no impact on life support skills [22], [23] Summarized, this means that more sophisticated technology does not automatically lead to more knowledge.

2.1.1 Debriefing in Simulation

Debriefing in the setting of simulation is defined as "[…] a discussion, reflection and analysis of a performance between individuals after resuscitation or training with the aim of improving future performance." by the ERC. [20] It nowadays plays an important role in the feedback process after simulation, but has also, with mixed outcomes, been evaluated for usage after real-world in-hospital cardiac arrest. [20], [24], [25] While the effect of debriefing itself on learning rates after simulation has been proven, the "optimal" type of debriefing is yet to be determined. [26] No debriefing strategy, e.g. oral-only, video-assisted feedback, or usage of special

Figure 1 - A high-fidelity simulation mannequin. The displayed model allows the introduction of intravenous lines, application of non-invasive ventilation as well as auscultation of lung sounds. U.S. Army photo by Patricia Deal, CRDAMC Public Affairs, approved for public use. The appearance of U.S. Department of Defense (DoD) visual information does not imply or constitute DoD endorsement.

techniques such as "debriefing for meaningful learning" (DML), which consists of a clearly defined process, spanning processing of emotions and critical review of the simulation scenario, has demonstrated a clear benefit over the other, as recently shown in a comprehensive review by Lee et al.[26], [27] Also, the timing and amount of debriefing have been under investigation, aiming to reduce the time consumed by debriefing, which is at the moment 20-30 minutes approximately.[28] More recently, with the advent of more powerful computational resources, new technologies, such as eye tracking, have found their way into simulation, debriefing, and medical education in general, a fact that will be discussed in more detail below. [6]

2.2 Eye tracking

The definition of eye tracking has been provided above. Some basic background, essential for further understanding of this master project, as well as eye tracking in a medical education environment, is now presented in this section.

2.2.1 Basic information

To enable understanding of eye tracking and how the gaze is connected to cerebral attention, a brief introduction into anatomy and how visual information is processed by humans shall be provided:

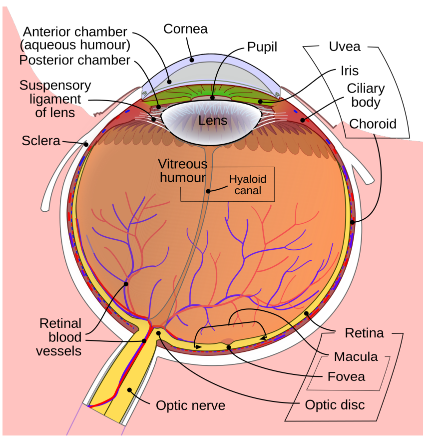

Figure 2 - Structures of the human eye. Image by Rhcastilhos, used under CC BY-SA 3.0

The human eye comprises essential components enabling vision: These include the *cornea*, acting as the transparent outer layer; the *iris*, regulating light entry via the pupil; the *lens*, focusing light onto the retina; and the *retina* itself, a light-sensitive layer housing photoreceptor cells known as rods and cones. [29] These cells convert light stimuli into electrical signals, which are then transmitted to the brain via specialized retinal cells and the optic nerve. [30], [31] Structures like the *vitreous humor* also contribute to maintaining the eye's shape. It's noteworthy that all structures preceding the retina are involved in optical refraction, ensuring proper light projection onto the retinal layer for clear vision. [32] Figure 2 provides an overview of the human eye, including other structures.

Rods and cones, distributed unevenly across the retina, are most densely packed within the *fovea*, a specialized region situated in the retinal *macula*, which represents the retinal center and the spot of most accurate vision. [32], [33] The fovea plays a pivotal role in achieving accurate vision and discerning fine details due to its exceptionally high density of cone cells. [34] This concentration enables precise visual perception necessary for tasks demanding detailed vision, such as reading. [35] Only if text is *foveated,* meaning, that vision is directed by the eye's muscles (Figure 3) in a way so that text hits the eye's fovea, it can be viewed sharply and thereby read and understood. [36] Upon transmission to the brain, visual information undergoes intricate processing. The optic nerve conveys electrical signals to the brain's visual cortex, particularly the occipital lobe, located at the very back of the brain, where primary visual processing occurs. [32] Information from the primary visual cortex (V1) is subsequently relayed to various brain regions for further processing. Current theories suggest that multiple brain regions must work in concert to facilitate the phenomenon of "visual consciousness". [37] More information on the connection between the eye and the mind is provided in a dedicated section below.

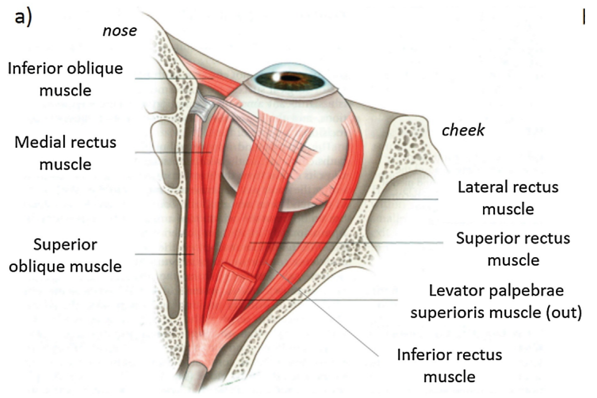

Figure 3 - Overview of eye the muscles. A human eye is displayed from above after removing parts of the skull. Used with permission from [47], licensed under CC BY-NC-ND 4.0

Human visual attention is complex and focusing on an object does not necessarily implicate that an individual is also always directing its complete conscious attention to it. [38] Nevertheless, it has been shown that eye movements and attention are connected and that visual attention even can precede and initiate eye movements. [39] Once the eye comes to a standstill, one can presume that an individual is directing visual attention to the current point of view, a process also known as *fixation* (which again leads to the foveation of a certain part of the visual field which seems of particular interest at that moment). [36] To move the eye(s) from one fixation point to another, fast movements, also called *saccades*, are performed by the eye muscles. [40] Frequently reported durations for fixations and saccades are 200-300 milliseconds (ms) for fixations and 30-80 ms for saccades. [36]

2.2.2 Eye tracking technique

Eye tracking is performed using special equipment following a participant's gaze. This equipment can be mounted e.g. on a computer screen or, as often the case, worn as a special pair of glasses connected to processing units. [41] Wearable eye trackers employ a multi-camera approach, with specific cameras directed toward the subject's eyes to conduct the tracking process. Meanwhile, other cameras operate concurrently, capturing the surrounding environment. This simultaneous recording enables the system to estimate not only the gaze point within the scene but also to detect if a participant is fixating on certain point and thereby foveating it, i.e. putting it in the center of visual attention. [42]

In contemporary gaze detection systems, the prevailing method is pupil illumination using infrared light. [36] This approach allows accurate tracking of ocular structures and, because infrared light is invisible to the human eye, does not disturb the subject during task performance, in contrast to visible light approaches. [43], [44] The appearance of the pupil in a captured image depends on the angle at which the light source intersects with the eye. If the light source aligns closely with the optical axis of the recording camera, the pupil appears bright, resembling the red-eye effect in photography. Conversely, if the light source is positioned more off-axis, the pupil appears dark. The used approach may vary between systems. [36] The eye tracking system can then detect the pupil, which is clearly distinguishable from its surroundings, and calculate an estimate of the direction of gaze by processing the acquired data with refined computational processes (Figures 4 and 5). Additionally, but not obligatory, some systems also use corneal reflections, also

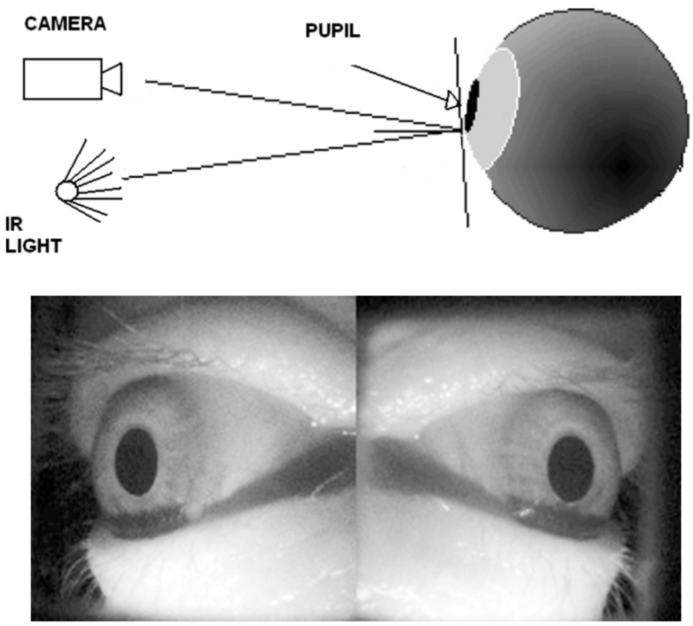

Figure 4 - Schematic representation of eye tracking using an infrared (IR) light source. The eyes are illuminated using an IR light source to avoid optical distractions for the participants. Illustration adapted with permission from [43] © 2011 IEEE Image of tracked pupil used with permission from Viewpointsystem GmbH, Vienna, Austria

called “Purkinje images”, in this case produced by the infrared light sources, as an additional reference point. [36], [45], [46] Using specialized software, a researcher can then proceed and analyze the recorded data in terms of various parameters. Those relevant to this work, including a definition, are provided in Table 1 in the methods section.

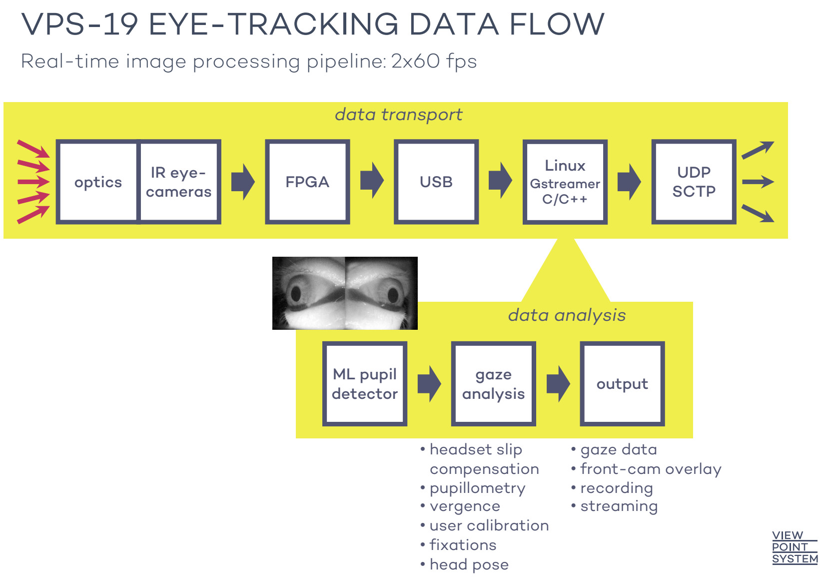

Figure 5 – Eye tracking data flow for the VPS 19 glasses used for this work. The incoming optical signal is registered by IR eye-cameras. After being processed by an integrated FPGA, information is then transmitted to the smart unit via USB, where further signal analysis, including ML-based pupil detection and gaze analysis, is performed. Afterwards, the output can be recorded, or transmitted for further usage via standard network protocols. FPGA: field programmable gate array, IR: infrared, ML: machine learning, SCTP: stream control transmission protocol, UDP: user datagram protocol, USB: universal serial bus, VPS: Viewpointsystem

2.2.3 Eye tracking and the mind

Various experiments have shown a connection between ocular activity and thought processes. [47], [48], [49] Already in 1967, Al'fred Yarbus showed that people attend to different kinds of sections of a picture when asked to answer questions related to it. The image used showed a scene of people in a house. When the subjects were asked to estimate the age of the displayed individuals, they focused more on their faces. However, when the task was to remember people's positions, they paid less attention to specific details and more to the general location of objects. [50] Fascinatingly, Yarbus used an apparatus consisting of suction caps that were attached to people's eyes to make eye tracking possible. [51] From the mentioned examples and many more in his book *Eye Movements and Vision*, Yarbus concludes that "Eye movements reflect the human thought processes". [50]

Since then, technology has become more sophisticated, and modern research dissected ocular movement into its finer subparts, individually analyzing these to gain insights into human cognition. [49] For example, Velichkovksy et al. have split up fixation durations in an experimental setting, aiming to distinguish between "preattentive scanning" and actual cognitive analysis of a (critical) optical stimulus. This work showed that once a critical – and thereby mentally challenging – event, occurs, fixation duration increases. [52]

In a comprehensive review, Eckstein et al. present evidence of how eye tracking metrics have been used in the past to draw conclusions about the human mind, ranging from cognitive approaches to problem-solving, reading or learning and, even connections between ocular behavior, such as frequency of blinks, and their connection to neurotransmitters, e.g. dopamine. [49]

Of particular interest from this summary to this thesis, and thus briefly presented here as examples, is the primary work of Grant et al. as well as Lai et al: Grant's workgroup showed, in an experiment with university students, that eye tracking can detect differences in fixation areas between students who successfully inferred a solution on a presented problem and those who failed to do so. Moreover, after identifying solution-relevant areas of a graph used for this problem and highlighting it in a follow-up test, they were able to dramatically improve success rates, demonstrating that fixations can be used to draw conclusions on cognition processes and improve solution rates. [53] Lai et al, in a literature review, identified 113 studies in total using eye tracking in some form to gain insights into human learning, also covering the topic of instructional strategies and processing of information. [54] Regarding the instructional strategies, they e.g. present primary work that investigated possible effects of "cued" animations of the human cardiovascular system, i.e. animations where various subparts are visually highlighted in a predefined sequence. Using eye tracking, the authors of this mentioned study found that cueing itself may help to focus on important areas but does not necessarily improve understanding. [55]

In summary, this demonstrates that eye tracking has been and is used extensively to gain insights into processes of the human mind. In one sentence, it backs the eye-mind hypothesis, which, simplified, states that visual fixations normally reflect the cognitive focus. [48] Importantly, eye tracking research has also looked into differences between experts in a field and novices, an aspect of particular relevance to this thesis. [56], [57] Using this approach, differences in attention allocation were identified: It was shown that experts in a field, in this case, chess players, fixate on relevant areas of a task more often using fewer fixations overall, which led to faster problem-solving, in comparison to less experienced players.

[58] Practice plays an important role in this as shown by Haider et al: Over time and with increasing experience, fixation of task-relevant areas gradually improves. In the authors' conclusion, information that is not relevant to a task is not perceived anymore by skilled individuals, but left out already on a "[…] perceptual […] level of processing." [59]

2.2.4 Eye tracking in medical simulation

Eye tracking technology has been used in various settings throughout the medical field, including evaluation of differences in expert versus novice visual behavior and identification of error-prone processes. [6], [60] Of particular importance to this thesis are its' applications in medical simulation settings, where video feedback has been used for quite some time, but more from a third-person perspective. [61] It was shown that eye tracking technology was ranked as highly effective by simulation participants in aiding learning success and the post-simulation feedback process. [62] Henneman et al. described Eye tracking as a feasible approach to gain additional information for debriefing after simulation, resulting in improved performance of patient safety-related tasks. [60] The same group also sees potential in Eye tracking to improve simulation design by analyzing expert behavior, objectively documenting participants' errors, or adherence to e.g. standard operating procedures. [63]

In terms of inter-group differences, Browning et al found evidence that nursing and paramedic students vary significantly in their visual behavior during a simulated medical emergency, with the nursing students looking significantly more often at an assistant who was provided to guide them through the scenario. This might reflect differences in emergency experience, i.e. expertise. [64] Additionally, Tanoubi et al. showed that medical residents, who managed to complete an emergency simulation scenario, needed significantly less time to identify the area of interest (AOI) containing the necessary equipment to stabilize the patient and subsequently dwelled significantly less on this AOI, implying that they were ready to use the provided tools. [65]

In total, there is sufficient evidence nowadays to accept Eye tracking as a valuable approach in medical learning and simulation training.

2.3 Decision theories for medical education

As the first two parts of this section provided an introduction to simulation and Eye tracking, the following part shall now cover some of the psychological aspects of this thesis, which were in focus of investigation using the techniques mentioned in the method section below. This mainly regards potential differences in decision-making between experts and novices as well as the amount of resources they have to dedicate to processing medical scenarios.

2.3.1 Experts' vs. Novices' decisions

Research has shown clear differences in problem-solving processes between experts in a field and novices. [66] Klein and Hoffmann emphasized the approach of seeing expert thinking as a perceptual-cognitive skill rather than just the collection of sheer knowledge and working hours, with some key aspects [67]:

1. The capability of experts to register not explicitly visible cues/patterns or even their absence when they would be expected – i.e. experts perceive the world around them differently – which represents expertise as a perceptual-cognitive skill. [67]
2. The ability to simulate a situation mentally and therefore evaluate possible consequences of available options and to imagine what has happened prior to the present scenario. [67]
3. The usage of an expert's knowledge base to implement top-down thinking strategies, such as deductive reasoning (i.e. a form of thinking where a specific conclusion is drawn from general ideas) [68]

This skill is then helpful in making decisions under complex, real-world circumstances, including shortness of time and overload of information as well as the management of uncertainties, a process that is also referred to as "Macrocognition". [69] In a stressful medical situation, such as an Emergency Department, experts tend to be better at macrocognitive problem solution – i.e. seeing the "bigger picture" of a problem and dynamically adapting their solution strategies as the problem evolves – than novices, which tend to rely more on previously collected objective data and ignore information that does not fit their pattern of solution. [3]

2.3.2 Clinical decision-making in emergency situations

Clinical decision-making can be seen as the competence to make decisions on diagnoses and treatments for patients based on provided information, such as symptoms (reported by the patient), signs (registered by the examiner), or past medical history. [70] It has nowadays been recognized as a crucial medical skill, developing over time and with exposure to clinical cases, being emphasized also in medical curricula. [71] One form of drawing conclusions in context with clinical decision-making is reflected in the usage of the hypothetico-deductive model, where professionals in a linear manner first generate a (working) hypothesis, then move on to its evaluation, refinement and finally – although this step is often omitted in emergency situations due to, time pressure – verification. [72] Also, the concept of “dual process theory” has gained increasing attention in the context of clinical reasoning. [73]

Made popular by Kahneman in his book “*Thinking, fast and slow*”, the theory basically states that the brain uses two “systems” to solve problems: “System 1” relies on fast, intuitive thinking that uses little effort but is also prone to mistakes. “System 2” on the other hand can be described as a more conscious, analytical process of thinking, requiring more energy and working a lot slower. [74] Congdon et al laid out its implementation in the clinical setting based on earlier research, stating that “System 1” is mainly used when dealing with known, frequent symptom patterns, and “System 2” is active when a clinician is facing a complex, less common case. [71] At present, the most popular approach to the two systems in clinical thinking is that they work simultaneously. Ten Cate describes the process of generating a differential diagnosis as an initial attempt of pattern recognition (System 1) and, if this attempt fails, a fallback on “causal reasoning”, i.e. trying to generate a hypothesis through a process of deliberate thought (System 2). [75]

On the other hand, the idea of naturalistic decision-making, also found its place in the discussion of decision-making in the medical field, especially when time for structured linear thinking is scarce, resembling often the situation in emergency departments, as pointed out in [3]. Naturalistic decision-making theory aims to account for factors that influence the process of how decisions are made in real-life scenarios, which include e.g. the lack of time or rapidly changing circumstances, putting a lot of focus on how especially experts act under those conditions. [76], [77]

Al-Azri even proposed a specific mental model for decisions during a medical emergency: It, especially in the first moments of a medical emergency, emphasizes that expert physicians must rely on pattern recognition for early life-saving treatment decisions, again mentioning that novices tend to struggle with this kind of approach. [78] The author then describes a transition to more linear decision approaches, such as the already mentioned hypothetico-deductive model, as the patient somewhat stabilizes, and time now allows a thorough investigation. He states that this kind of dynamically adapting decision strategy is "presented informally or insufficiently in emergency medicine training or education". [78]

To summarize, the process of decision-making in a clinical setting is complex, and not yet fully understood with multiple concurring theories, interacting with each other. Again, Eye tracking has been used in an attempt to gain insights in this area. [79]

2.4 Cognitive Load and its influence on learning

A short definition of cognitive load has been provided earlier, with working memory being one of its key features. Like a computer, the human brain can address resources to tasks that are performed at a given moment, but those resources are limited. [80] In 1956, George Miller showed that the human brain can process no more than seven items of <u>new</u> information (± 2) simultaneously. [81] Based on earlier research in this field, including e.g. the human capacity to distinguish the pitch of tones, he lays out that there is a limit to the amount of information that can be processed by the human brain, when participants need to decide between presented alternatives. [81], [82] Miller calls this "channel capacity" and expresses the capacity of those channels in bits, similar to computers: One bit, in this context, is defined as the amount of information involved in a decision between two equally likely alternatives. [81]

In the abovementioned experiment using pitches of tone, the channel capacity was found to be 2.5 bits or roughly 6 tones, that could be distinguished by the participants correctly. Everything greater than this led to an increasing number of errors. [82] Later, cognitive load theory was developed and refined based on such findings. [9] In an article specifically targeting the implications of cognitive load theory on educational technologies, Sweller, who contributed substantially to the development of the theory, lies out its framework, dividing human knowledge into two forms, originally introduced by Geary [83], [84]:

a) "*Generic-cognitive*", also termed as *biologically primary*. This form of skills is generally inherited and develops naturally, meaning also, that they cannot be improved by teaching. An example would be the ability to listen. [83], [84]
b) "*Domain-specific*" skills. These skills, in contrast, can be taught and at the same time need explicit advising. [83], [84]

Cognitive load theory (CLT) focuses on the latter form of knowledge, and how the amount of cognitive load that a learning system with limited resources faces may influence learning success. [85]

It emphasizes the fact that the working memory part of the human brain is limited and serves as a bottleneck for storing knowledge in long-term memory, which was Termed by Sweller as "The narrow limits of change principle". [84] Every new piece of information needs to be processed by working memory before it can be stored in long-term memory, organized in a schema-like manner. [84] Keeping this fact in mind, CLT aims to provide a framework that intends to support presenters of information in preparing the to-be-taught content in a way that does not exceed the working memory of learners and thereby optimizes the learning effect. [85] Also, in a paper from 1991, Sweller et al. provide some evidence on how CLT has influenced the generation of new experiments and how it supported the concept of schema theory, which states that novices in a field must rely on means-end analysis for problem-solving, whereas experts solve problems much faster and more accurate using the already mentioned schemata, previously acquired through experience. [86], [87]

Before the focus is put on how cognitive load is measured, which has also been highly relevant for this work, it is important to split cognitive load itself into its three main components:

- *Intrinsic cognitive load* refers to the intrinsic complexity of the material that needs to be internalized. This depends strongly on how elements within the material are interwoven with each other, meaning if they can be learned independently or must be understood simultaneously, a concept that is also called element interactivity. [88]
- *Extraneous cognitive load* refers to the amount of cognitive load that is additionally superimposed on the learners by the quality of teaching methods. It should be minimized to free resources for dealing with intrinsic complexity. [85]
- *Germane cognitive load*, which addresses the effort to transform the presented knowledge into schemas, which can then be used for problem-solving. This can only be successful if there is some working memory capacity left that allows schema acquisition. [85]

Briefly summarized, in an optimal setting, extraneous cognitive load in a learning environment is reduced to a minimum to allow resources to be allocated firstly to the processing of the problem's inherent complexity, i.e. to deal with intrinsic cognitive load and, secondly, to the usage of techniques to transform this knowledge into schemas in long term memory, which represents germane cognitive load. [85]

As it is of relevance to this thesis, it shall also be mentioned that there is evidence that high cognitive load may not only hinder learning, as outlined above, but also clinical reasoning and decision-making, rendering cognitive load an interesting target for this setting as well. [89]

2.4.1 Measurement of cognitive load

Substantial effort has been made so far to quantify the amount of cognitive load subjects experience during learning. To achieve this, both subjective as well as objective approaches have been chosen. [90] Those with relevance to this thesis shall be presented here. In principle, approaches to quantify cognitive load can be divided into subjective approaches, where participants answer questionnaires defined to measure cognitive demands, and objective approaches, trying to estimate cognitive load using biosignals. [90], [91], [92]

2.4.1.1 Subjective measurement of cognitive load using questionnaires

Two frequently used approaches are the Paas cognitive load questionnaire as well as the National Aeronautics and Space Association's (NASA) Task Load Index (TLX).

2.4.1.1.1 Paas questionnaire

PAAS cognitive load scale, named after the author of the initial research article, was developed, and published, for the first time in 1992. [93] In the original paper, Paas used it to test different forms of learning materials in teenage students, which again were developed based on the already mentioned cognitive load theory by Sweller. [86] The study showed that those students who used learning materials that contained already solved problems, which the students could use as examples for correct solutions, performed better in a subsequent test covering statistical basics. [93]

The scale itself consists of only 1 item, rated on a nine-point Likert scale, based on a scale for measuring the amount of difficulty experienced by participants when solving a task, originally developed by Bratfisch et al. [94] Study participants had to rate their perceived mental effort from "very, very low mental effort", corresponding to 1, to "very, very high mental effort", corresponding to 9. [93]

Due to its subjective nature, Paas' approach has been in the focus of discussion and criticism ever since, e.g. because of the assumption that participants understand the concept of “invested effort” and that every participant can sufficiently reflect the perceived effort during solving a task. [92] Also, the scale has been used inconsistently regarding wording, sometimes evaluating “task difficulty”, and sometimes assessing “perceived mental effort”. The former was shown to more accurately reflect intrinsic cognitive load, the latter germane cognitive load, as they were defined above. [95]

Nonetheless, as summarized by Paas himself in 2003, subjective measurement of cognitive load using rating scales has been used extensively. [91]

2.4.1.1.2 NASA TLX

NASA’s TLX is another approach to evaluating cognitive load connected to task solving. It was developed in the 1980s for aviation and defined by one of the developers, Hart, as a “multi-dimensional scale designed to obtain workload estimates from one or more operators while they are performing a task or immediately afterwards”, which has since been used in hundreds of studies. [96] In brief, an overall workload is calculated as the sum of various subjective sub-workloads in the categories of mental demand, physical demand, temporal demand, performance, effort, and frustration. [97] Each of the respective results is weighted, as users might find different subcategories to have a different impact on total workload. The weights for this process are acquired a priori by letting users rate their subjective importance of every subcategory regarding total workload. Summing the weighted results eventually returns the total weighted workload for a task. [97]

It has to be mentioned that the TLX and Paas scale differ in a way that the Paas questionnaire exclusively aims to measure cognitive effort, whereas the TLX takes a more holistic approach in attempting to measure the combined concept of workload. [93], [97]

Additionally, as mentioned by Hoonakker et al in [98], who looked into workload of intensive care unit nurses, different definitions of this sometimes hard to exactly grasp framework exist. One of them was provided by Vredenburgh et al., defining workload as “[…] a construct used to describe the extent to which an operator has engaged the cognitive and physical resources required for task performance”, based on the work of Backs et al. [99], [100]

It can be argued that the TLX is nowadays the most popular tool used for assessing task workload. [101] Its use and popularity have outgrown all other available subjective workload questionnaires by far, with search results of TLX-related papers lying in the thousands, while other questionnaire options produced no results at all. [102]

Yet, TLX has also faced criticism: McKendrick et al., for example, questioned its ability to correctly reflect workload and, especially, pointed out that the results of subscales might also be influenced by how participants interpret their respective meanings. Also, they raised doubts that subscales adequately reflect effects that they claim to measure, e.g. task performance assessed by the performance subscale. [103] De Winter et al., who performed the TLX search analysis mentioned above, have furthermore attributed the TLX's success partly to the "Matthew effect", meaning that the already huge popularity of the TLX even more strongly amplifies its usage and thereby leads to even more popularity. [102], [104]

Nonetheless, TLX has been shown to be a valid approach of measuring workload multiple times, also in a healthcare setting. [98] Coming back to the work of McKendrick et al., even in this work it was stated that the subscale of mental demand adequately reflects the objective mental demand placed on subjects. [103]

Summing up the available evidence, the Paas Questionnaire and NASA TLX represent two popular subjective approaches to measuring cognitive and overall workload. The adapted Paas Questionnaire, integrated into the "Debriefing Questionnaire", as well as the NASA TLX Questionnaire, are provided in the appendix

2.4.1.2 Objective measurements of cognitive load

2.4.1.2.1 Galvanic skin response

With the abovementioned problems connected to the subjective measurement of cognitive load, naturally, an effort has been made to develop reliable objective techniques for its quantification. These range from the measurement of pupil dilation, functional magnetic resonance imaging techniques, heart rate, and heart rate variability as well as GSR. [12], [105], [106], [107] As GSR was of importance to this thesis, it shall be explained in more detail here:

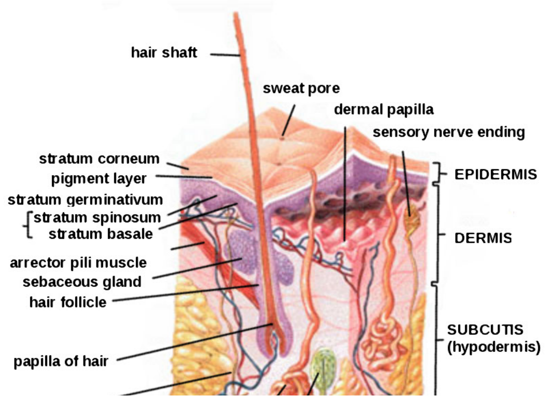

Figure 6 – Layers of the human skin with associated appendages. A sweat gland body is depicted in the middle layer of the skin, the "dermis", with a duct reaching the epidermis. Image by Daniel Souza, licensed under CC BY-SA 3.0 Deed

To begin with a brief anatomic summary, the human skin is comprised of various layers, the *epidermis*, *dermis*, and *hypodermis*, with the epidermis representing the outermost layer (Figure 6). Throughout the whole skin surface, embedded sweat glands can be found with a maximum concentration at the palms of the hands, soles of the feet, the forehead, and armpits. The majority of sweat glands belong to the so-called “eccrine” type, producing a watery solution. [108] Besides their crucial function in thermoregulation, some of these glands also respond to psychological stimuli, and increase their activity due to triggers such as fear or cognitive stress. Most of those psychologically “active” glands are located again at the palms and soles. [109]

Human sweat represents an electrolyte solution, mainly consisting of water and sodium chloride as its major salt component. [110] It influences the skin’s ability to conduct electrical currents. [111] Simplified, as the body is exposed to a stimulus and produces sweat, electric skin resistance decreases, and conductivity increases. Cacioppo describes this network of glands more detailed as a “set of resistors”, that allow electrical currents to flow with less resistance, as sweat rises in the duct of sweat glands. [109]

This phenomenon can be measured using various methods. The most popular method used today is the “exosomatic approach”: A small electrical current, harmless to the participant, is applied between two skin electrodes with constant voltage, and changes in conductivity as well as resistance due to psychological activation are then measured. [112]

The physical rationale behind this is that, like in other electrical current phenomena, Ohm’s Law can be applied, describing the relationship between Voltage (V), Resistance (R), and electric current (I) as $I = V/R$. [113] As the voltage applied is held constant, as mentioned above, via measuring the resulting electric current between two skin electrodes, one can then calculate skin resistance, expressed in the unit of Ohm (Ω) and its reciprocal, skin conductance, which is expressed in the unit of Siemens. In terms of skin conductance, the values lie in the range of microSiemens (μS), and for resistance are located in the range of kΩ to MΩ. [109]

Using this method, a conductivity and resistance signal is then acquired and plotted over time, which can be analyzed. The signal mainly consists of a tonic component, representing a “baseline” of psychovegetative activation, also called skin conductance level (SCL) with a phasic component, that resembles spikes or peaks, stacked on top of it, called skin conductance responses (SCR, Figure 7). [111], [112] Whenever such a peak in the GSR signal occurs in proximity to a stimulus, e.g. a picture, one can use this as a surrogate for cognitive activation/load. This is then also referred to as “event-related” SCR. [111] As pointed out by Stern et al. this feature is not exclusive to GSR signals, but to psychophysiological signals in general, also including heart rate for example. [114]

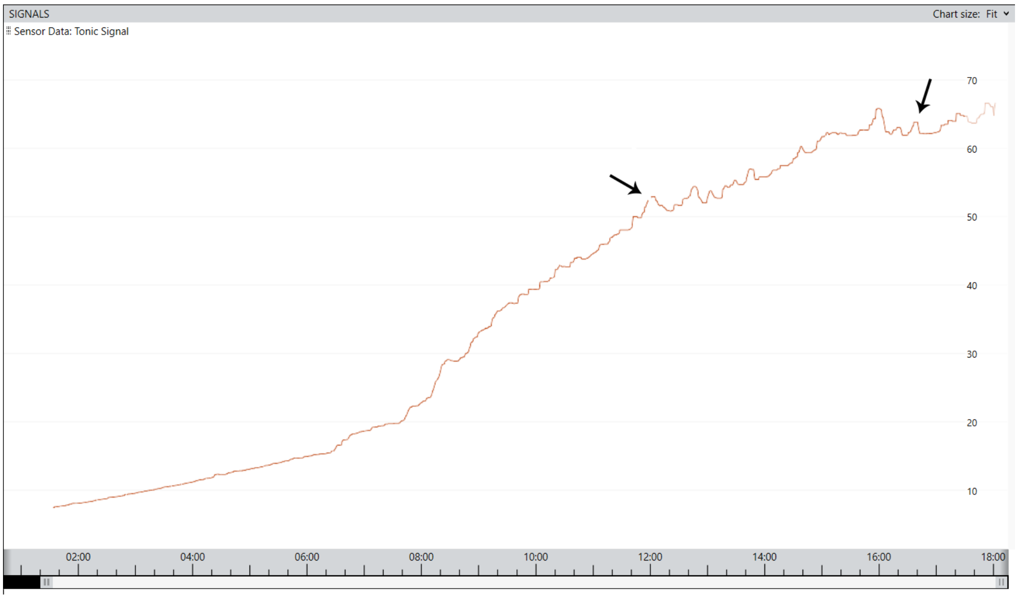

Figure 7 – The GSR signal and its components. Tonic GSR signal (line) with intermittent peaks stacked on top (black arrows). Self-created figure using iMotions 10.

In summary, galvanic skin response is a frequently used approach to objectively measure cognitive load by measuring changes in electrical skin properties. [11], [12], [18]

2.4.1.2.2 Heart rate – Heart rate variability

The second biomarker set that was of importance to this thesis in conjunction with cognitive load is formed by heart rate (HR), i.e. the number of heartbeats per unit of time, generally reported in beats per minute, and heart rate variability (HRV), which can be defined as the variation of heart rate between two consecutive heartbeats. [115] Like heart rate, it can be influenced by the autonomic nervous system, the part of the human body's nervous system that is generally concerned with the regulation of bodily functions. [116] External stimuli, such as psychological and physical stressors, can trigger autonomic responses that lead to changes in body functions, such as an elevation of heart rate or a change in HRV, that allow coping with forthcoming external, including psychological, challenges. [117]

HRV variability is calculated using electrocardiogram (ECG) signals, as shown in Figure 8. The distance between two consecutive beats is determined using the major spike of an ECG, also called R wave, and reported in milliseconds. The heart is not beating at a perfectly constant rate even under stable conditions and thus, time between two beats, i.e. between two R waves, varies. [116] This variation in time is measured and reported as heart rate variability. Both too little and too high HRV may reflect pathological conditions. [117]

In this context, HR and HRV have been evaluated as possible markers for cognitive load and stress, in general learning but also a medical decision and learning context. [89], [105], [118] Results are sometimes conflicting, with HR and HRV being reported as less sensitive in comparison to subjective quantification of cognitive load, e.g. in a reflection by Paas et al. on their work in the field, but also as valid cognitive load markers by others. [89], [119] In a meta-analysis performed by Hughes et al., cardiac biosignals such as the ones just outlined, were, in summary, reported as being a valid approach to objectively quantify cognitive load, as an influence of load on them has been demonstrated. [120]

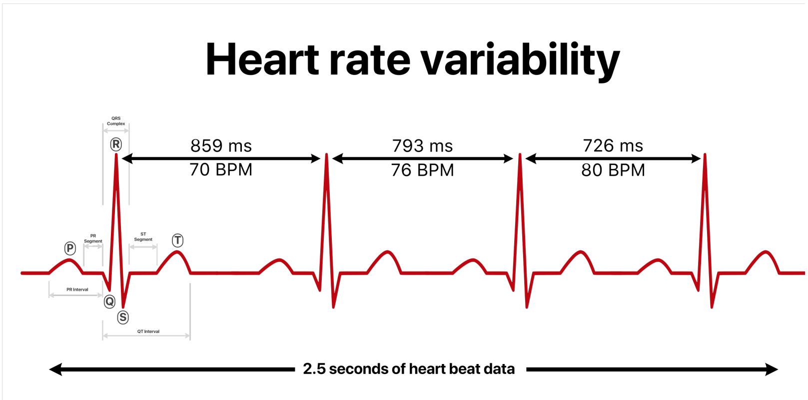

Figure 8 – Example of heart rate variability extraction from an exemplary electrocardiogram (ECG). The distance between two consecutive “R” waves of an ECG is measured in milliseconds, with subsequent calculation of the variation of this distance from beat to beat. BPM: Beats per minute, Ms: Milliseconds. Image by YitzhakNat, used under CC BY-SA 4.0 / Cropped from original

2.4.1.2.3 Eye tracking

Regarding eye tracking, higher fixation duration, higher saccade velocity and number of saccades, as well as shorter saccade duration and amplitude, have been reported to reflect higher cognitive load. [121], [122], [123], [124] Nevertheless, it must be mentioned that regarding saccades, reported correlations with cognitive load are somewhat contradictory, with a position paper from Zagermann et al, e.g. stating, that longer saccades, in terms of covered distance, indicate higher cognitive load, whereas Keskin et al. e.g., as well as Chen et al., report shorter saccades with increasing load. [122], [123]

2.4.2 Cognitive load in medical education

To conclude the section on cognitive load, a few examples of how it has been used in medical simulation shall be provided, as the cognitive load approach has also gained momentum in medical teaching. [125]

Even more specifically and of importance to this thesis, developers of medical simulation have asked the question of how to incorporate frameworks such as cognitive load into the design of medical simulation scenarios to maximize learning. [8] For example, Fraser et. al reflect on implications of CLT specifically for medical simulation, pointing out that there might be an optimal load for learning, which must not be either too low or too high. They also carry on and give specific recommendations for the respective subphases of simulation, like the debriefing process and how to manage the different categories of cognitive load, which have been defined above. [8]

Importantly, in the same work, limitations of CLT in a medical simulation environment are also discussed, including the fact that CLT was developed for environments a lot more controllable and expectable, such as classrooms. While under these circumstances cognitive load seems a lot more adaptable, there are a lot of “uncontrollable variables” during medical simulation, as the authors point out, such as equipment malfunction, which might influence load. Also, emotions play a substantial role during simulation training and might reflect a part of cognitive load that deserves active “load management”. [8]

Importantly, research has also been looking into the validity of cognitive load measurement approaches during medical simulation: A work on cognitive load in respirology and internal medicine residents, for example, showed that the overall NASA – TLX and the Paas questionnaires might not be used interchangeably for measurement of total cognitive load, as the two approaches did not correlate with each other in terms of overall load score. Only, as the authors reported, the subcomponent of intrinsic cognitive load was equally assessed by both. [126] Other work from the same group investigated cognitive load assessment in simulation in general, including aviation and medicine. The study included subjective as well as objective approaches to cognitive load measurement. Interestingly, in this analysis correlation between cognitive load and learning ranged from positive to negative, with low validity evidence for cognitive load assessment, suggesting the need for further development and validation of cognitive load quantification tools. [127]

Nonetheless, the CLT framework has been used in contemporary simulation-associated research, also in combination with emerging technologies, including virtual reality (VR), task-evoked pupil response, and automated adaption of medical simulation difficulty via real-time cognitive load assessment using galvanic skin response, heart rate variability, and machine learning. [18], [128] Interestingly, some research on VR used for surgical training showed, that an overwhelming immersive VR environment may negatively influence learning success, as it might unnecessarily increase task complexity for beginners, pointing out limitations of the human mind which need to be considered. [129]

2.5 Summary – Connecting the dots

During this background section, information essential for further understanding of the key concepts underlying this thesis, its research questions, and possible ways to answer it, were provided. A statement that can now connect these fragments and draw a bigger picture of fundamental ideas important for this work could be as follows:

Simulation training plays a key role in medical education. [21], [130] As the setting of simulation often is one of a medical emergency, theories of how decisions under pressure are reached come into play. [78] Continuing, as simulation aims to improve the skills of participants, learning theory, especially in a connection to how experts are capable of making life-saving decisions promptly and under great pressure, must get involved. [70] As experts obviously acquired some sort of skill set and experience that enables them to do so, this unavoidably leads to the question of how to gain information on these skills and how to teach them to novices. [79] Because emerging technologies, such as eye tracking allow unprecedented insights into expert behavior, they are used in analyzing it. [58], [79] Now, given the fact that novices are active learners, one needs to ask the question of how to best present information about experts, gained using mentioned technologies. Here, last but not least, the piece of cognitive load theory falls into place, aiming at presenting information in a way that optimizes learning outcomes. [8]

If the incorporation of all these frameworks, techniques, and approaches to learning leads to new insights into how to improve simulation training, patient outcomes could be further improved. This mainly represents the backbone idea of this work and its research questions. The methods used to address them will be presented in the following section.

3 Requirements / Methods

The broader underpinning of this work, as already defined briefly in the introduction, is the fact that evidence regarding the combination of various assessment tools to examine human behavior in novices versus experts in the field during medical simulation training is scarce. Yet, such work could allow improvement of future medical curricula as well as developments in simulation technology. The following section is dedicated to laying out the research plan providing rationale for the chosen methods.

The Personal competence of the author of this thesis in planning the to-be-presented project comes from years of experience in medicine, more precisely in cardiology and emergency/intensive care medicine as a resident in the field, as well as strong personal engagement in educational projects for medical students / young doctors at the author’s workplace. The scientific backbone was provided in the background section, most relevant work will be once again mentioned here when deemed necessary.

3.1 Methodology

Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist, as provided under [131], was used to ensure comprehensive reporting of necessary information

The workflow of this thesis was divided in three stages:

1. Review of existing literature in the field
2. Conduction of a pilot-trial
3. Analysis and interpretation of acquired data

3.1.1 Literature review

First, the existing literature was screened for work relevant to the previously defined research question. The background section above generally shows the result of this stage, as it summarizes the existing evidence and puts it into context. As the goal of the review was more the acquisition of research regarding simulation training and measurement of various body parameters in general rather than the search for a change in outcome, a P(Population), I(Intervention), C(Control), O(Outcome) strategy seemed unfitting.

3.1.2 Development and conduction of a pilot trial

In this second, largest, step, an observational pilot trial was designed and conducted. The development, including the rationale for the chosen methods, shall be outlined here.

3.1.2.1 Ethics

As no intervention was planned in this study, it was deemed exempt from requiring approval by the Upper Austrian Ethics Board.

3.1.2.2 Selection of participants and simulation equipment

At first, the design of the pilot trial was to be determined. As the aforementioned research question was directed towards potential differences between novices and experts regarding visual behavior and cognitive load, two groups were defined, one expert group and one with less experienced learners.

The expert group comprised certified senior physicians in the field of anesthesiology and internal intensive medicine, or residents in their last year of training who already passed their certification exams. Both subspecialties run a dedicated intensive care unit at the study site. They were therefore deemed fitting to have gained experience throughout their career to deal with an internal medicine emergency (see below) in an “expert way”. For the participants from the internal medicine department, to assess regular exposure to emergency medicine patients, at least two years of working experience at an intensive care unit was mandatory additionally.

As for the emergency medicine novices, residents in anesthesiology or internal medicine, certified general practitioners, or general practitioners in training were included. Certain criteria to guarantee at least a basic knowledge in some crucial emergency medicine skills, such as intubation – i.e. securing a human’s airway by inserting a tube that allows controlled ventilation of the lung – had to be fulfilled for this group and were evaluated during the inclusion process. This was introduced to avoid bias in the results by exposing a totally untrained individual to a very likely extremely overwhelming situation, resulting in an exceedance of working memory and failure to complete the scenario. In contrast, as some participants, potentially eligible for the novice group, also might have gained experience by working as paramedics, etc., this was evaluated during the inclusion process and participants were excluded if they regularly had worked as paramedics or emergency physicians in the past.

Other relevant inclusion and exclusion criteria, e.g. regarding vision, were defined as provided in the appendix.

Next, the equipment used had to be defined. As mentioned above, evidence for the efficacy of high-fidelity simulation equipment is ambiguous, yet the ERC still vowed for its use when resources are readily available. [20] Thus, a Realiti 360 simulation unit (iSimulate Pty. Ltd, Fyshwick, Australia) in combination with a simulation mannequin, fully controllable by a wirelessly connected tablet (iPad, Apple Inc., Cupertino, USA), which is used for teaching in the hospital where the study was conducted, was also the solution of choice for this project. Regarding airway equipment and medication, the standardized emergency bags provided by the clinic where the study was conducted were used.

3.1.2.3 Definition and preparation of study site

As the author of this thesis was working at Salzkammergut Hospital Voecklabruck, Austria, during the conductance of this thesis work, this hospital was also chosen as study site. A location at the emergency department, providing even light conditions important for eye tracking equipment to work properly, was chosen as in house site for the first round of the study. Because of unplanned construction work at the emergency department due to damage unrelated to this project, the second round was then conducted at a momentarily closed ward, also providing proper illumination. The mannequin and monitoring equipment were positioned in a manner so they would be easily accessible and allow tracking of all important actions. After each round, the room and equipment were prepared for the next round by the study team and the nurses.

3.1.2.4 Dates of data collection

Data was collected on two days between January and March 2024.

3.1.2.5 Definition and preparation of the medical team

The medical team for the simulation scenario consisted of three people comprising one medical doctor, i.e. the study participant, and two experienced intensive care nurses providing assistance during the scenario, representing also the standard team for an emergency patient at the cardiac care unit (CCU) of the hospital where the study was performed. As for communication with the patient, the simulation coach provided answers “as the patient”, when the participants addressed the simulation mannequin with questions. If deemed necessary by the team, simulated contact with the hospital's CPR team as well as a cardiologist could be performed, with the simulation trainer again acting as the respective contact.

Before the start of the scenario, the nurses were trained on the simulation equipment and allowed to position the provided equipment for securing the airway, as well as emergency medication, in a way that would serve them best. An intravenous line was already placed before the start of the scenario.

3.1.2.6 Definition of the simulation scenario.

A general cardiovascular emergency scenario was chosen for this study, as it was deemed likely to be common enough to be recognized by all doctors, regardless of specialty. [132] More precisely, cardiogenic shock related to ST-segment elevation myocardial infarction (MI) was defined as the main presenting pathology. Cardiogenic shock may be described in simple terms as a reduction in the heart's ability to pump blood, leading to low blood pressure, and hypoperfusion of vital organs, such as the brain, and therefore resulting in a life-threatening state. [133] It may be caused by myocardial infarction, which can be explained as acute myocardial (= the muscle cells of the heart) injury due to reduced oxygen supply, a state also known as *ischemia*. [133], [134] The term ST-segment elevation refers to typical alterations of an ECG that may be present with MI and other conditions. [135]

Symptoms and signs of the patient were severe chest pain, low blood pressure, tachycardia, and clammy forehead skin. A detailed list of the patient's signs, symptoms, vital parameters, and past medical history is provided in the appendix.

Regarding the scenario timeline, a five-minute initial phase was defined, with the simulated patient still conscious, to allow participants to assess the patient presenting with severe chest pain and perform anamnesis, initial diagnostics, including ECG, as well as monitoring. After 5 minutes, regardless of prior actions, the patient would always go into ventricular fibrillation (VF) related cardiac arrest, triggering the resuscitation phase. The latter lasted for 4 two-minute cycles as defined by the ERC, i.e. two minutes of chest compressions and ventilations, interrupted by analyses of heart rhythm. [136] Study participants could then decide freely if they wanted to deliver a shock, using the provided defibrillator, to the patient. The simulation trainer guided the participants by informing them when a cycle was over. Despite this, the simulation trainer did not provide any input on correct patient treatment etc. Around minute 10 of the scenario, a distraction in form of a garbage can being kicked over was introduced. If the participant was performing intubation at this time, the distraction was delayed until one minute after the intubation procedure was finished.

Study participants could administer all available medication ad libitum. When they saw fit, they could intubate the patient, i.e. secure the airway with an endotracheal tube, whenever they found the point in time to be adequate. After the 3rd cycle, if defibrillation was delivered, the patient would return to a heart rhythm compatible with proper blood circulation. At the last analysis, participants could then recognize the state of terminated VF. As soon as this happened and they called for further steps, the scenario was terminated. A visual representation of the study timeline is provided in Figure 9.

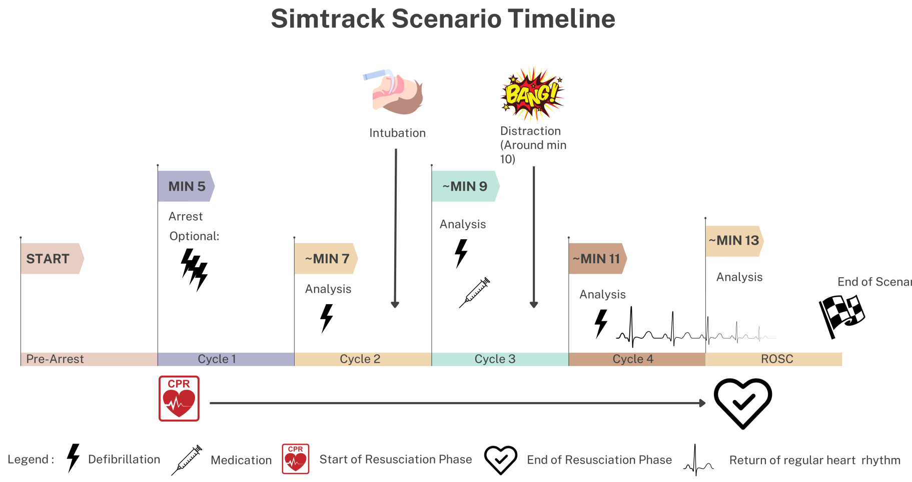

Figure 9 - Visual representation of simulation scenario timeline. Analysis refers to the points in time where the patient's heart rhythm was assessed and the decision made if an external electrical shock (defibrillation) should be delivered as treatment. CPR: Cardiopulmonary resuscitation, ROSC: Return of spontaneous circulation

3.1.2.7 Definition of sample size

The sample size definition of this project was performed based on previous work. Looking at the existing evidence, there are some well-conducted trials, for example by Wagner et al., that analyzed visual behavior during emergency medical simulation. [5] Also, other work has evaluated the usage of eye tracking during e.g. heart ultrasound (echocardiography) or simulated anesthesiologic procedures. [57], [137] GSR was assessed for example in one trial in the operating theatre. [138] Number of participants included ranged from 13 to over 100. Keeping in mind the available number of staff in the hospital where the study was conducted, the timeline and scope of this thesis, as well as the pilot-like approach, a number of N = 12 participants was defined to be fitting, with an allocation rate between experts and novice group of 1:1

3.1.2.8 Allocation of participants to groups

Potential participants were personally addressed by the author in the clinic regarding their willingness to participate prior to the study. On the study day, participants were allocated to their respective groups by the author of this thesis, based on the answers of the used register questionnaire (see below).

3.1.3 Equipment for biosignal measurement

This section provides insights on the used equipment. For more precise information on how equipment was fitted onto participants, see section 4.1.4

3.1.3.1 Eye Tracking

VPS 19 eye tracking glasses (Viewpointsystem (VPS), Vienna, Austria) were used for this work. Study participants were equipped with the eye tracking system by an in-house specialist from VPS. The system used for this study leverages a precise dark-pupil approach, as explained above. Frame rate for the cameras was 60 fps for the eye cameras and 30 fps for the front camera capturing the scene.

3.1.3.2 Galvanic Skin Response

As described above, Galvanic Skin Response may provide insights into an individual’s level of sympathetic activation and also serve as a surrogate for cognitive load. [12]

To measure GSR, a Shimmer3 GSR+ unit (Shimmer Research Ltd., Dublin, Ireland), provided by St.Pölten University of Applied Sciences (UAS), was chosen to be a fitting sensor for this project. This sensor measures electrodermal activity using an exosomatic approach, where a small current is delivered between two skin electrodes, as explained above. [112] An internal analog-to-digital converter then converts the measured value to a 12-bit number representing internal skin resistance. Calculation of resistance and conductance is performed automatically by the device and reported in the units also stated above [109] The device auto-calibrates to the range of measurement that fits best to the current participant. [139] The final result of a GSR measurement is then a multi-component signal, comprising a tonic and phasic component, as described in the background section.

Recommended electrode positions for GSR measurement are palms, fingers, and the sole of the feet, where the highest density of eccrine sweat glands can be found. [112], [114] However, these positions were expected to be subject to strong artifacts by excess movements and were therefore deemed to be unfitting for the planned approach. Reasons were as follows:

- Asking participants to completely refrain from manual work would have led to a drastic reduction of simulation validity. Also, the usage of gloves, which is demanded during medical contact, was expected to be made impossible by wearing a GSR device.
- Medical emergency scenarios naturally tend to demand caregivers to move around the patient, which would also have severely influenced a signal collected at a participant’s foot.

Thus, during a setup scenario, alternative electrode positions were evaluated. After testing and excluding the “classic” positions for the abovementioned reasons, alternative positions were then tested. In a comprehensive work, van Dooren et al. tested several alternative measurement positions spread throughout the whole body and reported the wrist as a valid position for GSR measurements. [140] For the work presented here, the wrist also represented a suitable position and was therefore chosen to be the location of choice.

3.1.3.3 Heart rate sensor

For heart rate acquisition, a Polar H10 ECG chest belt (Polar Electro, Kempele, Finland) was used, which was also provided by St.Pölten UAS.

3.1.4 Selection of cognitive load assessment tools

As described in 2.4.1.1., with the NASA TLX and the PAAS cognitive load scale being valid tools for subjective assessment of cognitive load, they were also chosen for this study. [91], [98]

3.1.5 Development of used questionnaires

To assess participants for eligibility, collect demographic data, as well as feedback regarding experience with sensors and the scenario itself, digital questionnaires were developed using the “LimeSurvey” Online Tool (LimeSurvey GmbH, Hamburg, Germany). Data were stored anonymously on a General Data Protection Regulation (GDPR) conformant server within the European Union. Questionnaires used are provided in the appendix.

Regarding question selection, for the feasibility part, questions were adapted from two previously published questionnaires by Wagner et al. and Blum et. al. [4], [5] To allow reflection on emotions and thoughts during the scenario, a “recapitulation” questionnaire was also developed where participants could reflect on certain points during the scenario using free text as well as the already mentioned PAAS cognitive load scale. For this part, three times were defined:

- The point in time when the patient went into cardiac arrest
- The point in time when a distraction (loud noise, see below) occurred
- The point in time when the participants performed intubation

3.1.6 Participant path on study day

At the beginning, participants were assessed for potential exclusion criteria by the author, as mentioned. Eligible participants then signed an informed consent and GDPR information sheet and subsequently completed the digital registration questionnaire, which also covered eligibility criteria. Each participant was assigned a unique ID consisting of a single-digit number and a letter.

Following assignment to the respective study group, expert or novice, participants then underwent sensor fitting. First, the chest belt was attached to the participants’ chest, followed by the GSR unit. For the latter, two gel-electrodes (Kendall H98LG, Cardinal Health, Dublin, Ohio, USA) were attached to the participants’ wrist on the non-dominant hand, on the outer and inner side of the forearm respectively. Lastly, participants were equipped with VPS 19 glasses. A display of a participant fully equipped with the used biosensors is provided in Figure 10. As the chest belt used for heart rate acquisition was worn directly on the chest, it was covered by the participant’s clothes.

Following gaze calibration by the in-house technician from VPS, data stream to a study computer was then started using a secure WIFI connection. The same procedure was performed for the chest belt and the GSR unit. Data were recorded using iMotions Lab (iMotions A/S, Copenhagen, Denmark). As soon as stable connections were established, recording was started, and participants were then asked to sit still on a chair for 1 minute to allow collection of baseline signals for heart rate and GSR. After this interval, they were handed a simulation briefing sheet displaying information about the simulation equipment, as well as a triage sheet similar to the ones used in the emergency department of the study site, providing the most important information, about the patient they were to encounter (see appendix). Once all questions that may still have arisen from the participant’s side were cleared, the scenario was then started. After completion of the scenario participants were allowed to recover for a few minutes, and, once the equipment was detached, were asked to complete the feasibility and recapitulation questionnaires described above. Lastly, the NASA TLX was performed on an iPad mini (Apple Inc., Cupertino, USA) using the freely available NASA TLX application. Once this was completed, participants were discharged from the study site if they had no further questions. In the last step, biosensor equipment was disinfected, and functionality rechecked before the next participant was allowed to enter the study room. The simulation site and equipment were also prepared for the next round. For a visual representation of this process, refer to Figure 11.

Figure 10 - Fully equipped participant wearing eye-tracking glasses, the related smart unit stored in a bag (arrow), as well as a GSR sensor attached to the forearm of the non-dominant hand (circle). Note that the bag was brought to the front for visibility reasons. ECG chest belt covered by participant's clothes. Image used with participant's permission. © author of this thesis, all rights reserved.

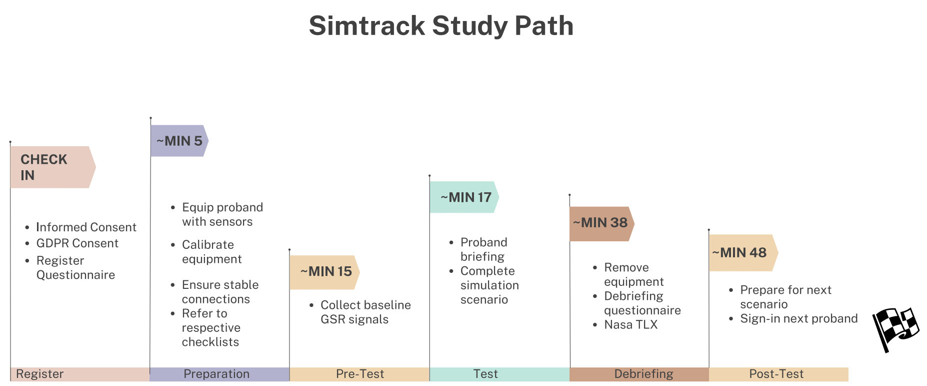

Figure 11 – Pathway for study execution. GDPR: General data protection regulation, GSR: galvanic skin response, NASA TLX: national aeronautics and space association task load index.

3.1.7 Selection of variables

3.1.7.1 Eye Tracking

To address the research questions defined above in a more systematic manner, the monitoring equipment was defined as the AOI for this study. Besides analysis of AOI statistics during the overall scenario, the following times of interest (TOIs) were defined for sub-analyses:

- Time of first medical contact
- Time of cardiac arrest
- Time of intubation
- Time of intentional distraction by a loud noise (Plastic container dropped on the floor)

A more precise definition of TOIs, including their exact time points during the scenario, is provided in the appendix.

Table 1 provides an overview of the eye tracking variables selected for further analyses, including those selected for inferential statistical testing, as they were expected to provide the most information gain regarding the study questions.

Table 1 - Recorded eye-tracking metrics

Fixation based metrics	Saccade based metrics	Blink metrics
Fixation count = How often participants fixated the AOI	Saccade count = How many saccades were registered	Blink count = How many times participants blinked
Time to first fixation (TTFF) (sec) = Time until participants fixated AOI for the first time	Average saccade amplitude (deg) * = Average displacement of the eyeball caused by a saccade in degrees of visual field	Blink rate = Blinks/min
Fixation time (min) = Total time the respondent spent fixating at the AOI (i.e. the sum of all fixation durations inside this AOI).	Average peak velocity of saccade (deg/sec) *	
Fixation time in percent of the time the AOI was active *	Saccade rate * = Saccades/min	
Fixation rate = Fixations/min		
Duration of average fixation (msec) * = How long did an average fixation on the AOI last		

* = variables chosen for inferential statistical testing

3.1.7.2 GSR

3.1.7.3 Heart rate and HRV

Multiple variables have been reported to reflect the effect of cognitive load and stress on HRV. Amongst the most popular is RMSSD, the root mean square of successive differences between normal heartbeats. [120] Thus, RMSSD was chosen to be analyzed during this setting.

3.1.8 Analysis of acquired data

Analysis of acquired data was performed in a multi-step approach. Eye tracking analysis was performed for each participant separately, using iMotions. The predefined AOI was drawn around the monitoring equipment and adjusted manually throughout the scenario. AOI data were then calculated automatically by iMotions. For GSR analysis, a peak amplitude threshold of 0.005 µSiemens, peak onset threshold of 0.01 µSiemens, and a minimum peak duration of 500 ms was set and the automated GSR analysis, embedded in iMotions, was then performed. Sampling frequency was set at 128 Hz. Heart rate data were left as is, with no automatic analysis.

All acquired data were then exported to a CSV file and processed using automated Python scripts written in Jupiter Notebooks by the author himself. This includes calculation of HRV metrics. Statistics were calculated using SciPy library for Python. Graphics were developed using Matplotlib and Seaborn. A GitHub repository was created for version control. Open AI's (San Francisco, USA) GPT 4o was used partially for writing, debugging, and refining code. For quantitative variables, where appropriate, conclusive statistical testing was performed using Mann-Whitney test for independent samples. Friedman test with post-hoc Dunn-Bonferroni or Wilcoxon signed rank test were used for related samples. As sample sizes were small, exact Mann-Whitney method was chosen, as recommended. [141] Analyzed eye tracking variables are marked in Table 1.

Regarding GSR, the number of peaks during the overall scenario and during defined TOIs were selected for analysis.

Additionally, from the acquired heart rate data, RMSSD was selected to be analyzed for differences between subgroups.

From the questionnaires, for this thesis, the overall NASA TLX score as well as sub score for mental effort were also included inferential statistical analysis. Respective p values are reported in the results sections with the selected variables. Due to the already small sample size, all participants were included in the analysis, even when data was missing. Where necessary, e.g. because of a premature data stream loss, participants were excluded for analysis of certain absolute values, and only relative values, e.g. proportions of available scenario recordings, were used (see also below in results). The sample size calculation, the fixation time of equipment in relation to overall scenario duration, average saccade amplitude, as well as the NASA TLX weighted mental sub-rating, were chosen. Calculation was performed using G*Power 3.1 open-source software, as published in [142].

3.1.9 Post hoc analysis

As a pattern emerged, showing intragroup differences in some variables throughout the scenario, the following variables were selected for post hoc analysis:

- RMSSD at defined TOIs in comparison to baseline
- Peak velocity of average saccade and average fixation duration during distraction TOI in comparison to remaining TOIs.

4 Results

4.1 Study collective

In total, 30 physicians potentially eligible for this work were contacted by the author. 14 were not willing to participate because of personal reasons and 4 were not able to participate due to staff shortage on scheduled study days. 12 participants were eventually included in this study, 7 in the novice group and 5, due to mentioned inclusion difficulties, in the expert group (Figure 12). Regarding gender, 3 participants overall identified as female, 9 as male, and 0 as nonbinary. Mean age of all included participants was 35 years (± 6). Mean age per subgroup was 32 (± 4) years for novices and 38 (± 7) years for experts respectively. Overall confidence in the individual participant's ability to handle internal medicine situations on a five-point scale, where 1 = not confident and 5 = very confident, was 3.4 (± 0.8) for the overall group, 3.3 (± 0.8) for novices and 3.6 (± 0.9) for experts respectively. Regarding specialty of medicine three participants were affiliated with the field of internal medicine (25% of participants), three were classified as general

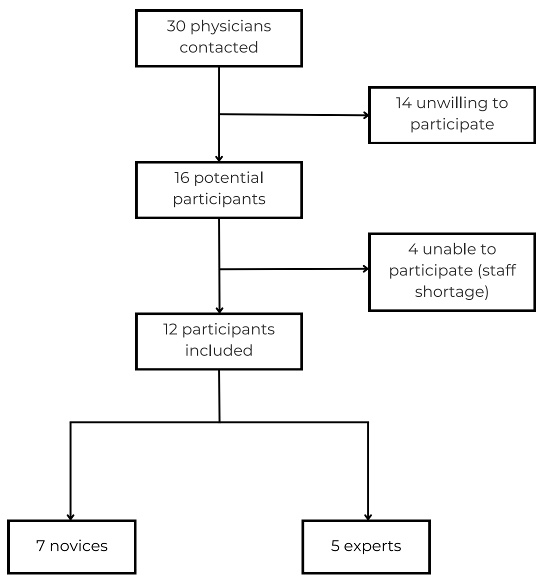

Figure 12 - Inclusion flow chart.

practitioners (25%), and six were associated with the anesthesiology department (50%). Collected baseline measurements for HR showed a baseline mean HR of 78 (± 8) bpm in the overall group, with 80 (± 9) bpm in novices and 76 (± 9) bpm in experts respectively. As for HRV, the mean baseline RMSSD in the overall group was 37 (± 21) ms in the overall group, 37 (± 26) ms in novices, and 37 (± 14) in experts.

4.2 Scenario and biosensor measurements

All 12 participants completed the scenario. However, as for one participant, the video stream ended prematurely and the last minutes of scenario recording were lost, this participant was excluded from analysis for scenario length. Mean scenario duration in the remaining 11 participants was 13.7 min (± 0.4) for the overall group, 13.7 min (± 0.4) in novices, and 13.8 min (± 0.5) in the expert group respectively. Completely identical scenario durations were not expected, despite of a standardized medical emergency used, because of decision-making latencies, etc.

4.2.1 Eye Tracking, Heart Rate and HRV

4.2.1.1 Overall scenario

For analysis of overall scenario metrics, two participants from the novice group were excluded for analysis of the following variables due to missing data caused by loss of the data stream: Mean fixation count + fixation rate, absolute fixation time, average fixation duration, average saccade count + saccade rate, average saccade amplitude and peak velocity of the average saccade (therefore n = 10 in overall and n = 5 in novice subgroup respectively). As data was available for time to first fixation and fixation time in relation to the timespan that the AOI was active, those participants were nonetheless included in this analysis (N = 12 overall and n = 7 in the novice subgroup respectively).

The overall study group fixated the AOI for a mean of 387 (± 79) times, accounting for a total fixation time of 2.3 (± 0.9) min and a fixation rate of 28 (± 6) fixations/min. Relative fixation time for the overall group was 16 (± 7) % of the whole time interval that the AOI was active. (extracted from overall N = 12 group). An average fixation lasted 357 (± 112) ms. Mean time to first fixation was 39 (± 30) seconds (extracted from the overall N = 12 group). The average saccade count for the overall study population was 573 (± 120), or 42 (± 9) saccades/min with an average saccade amplitude of 4.7 (± 1.0) degrees (deg) and an average peak velocity of 106.2 (±

24.3) degrees per second (deg/sec). A comprehensive display of analyzed variables for the overall study group is provided in supplementary table 1.

As for heart rate and heart rate variability, the average heart rate of the 12 participants during the complete scenario was 98 (± 12) bpm, with a mean RMSSD for the overall study group of 26 ms (± 11). In comparison to the baseline phase, this translates to an increase in heart rate of 26 (± 13) % and a decrease of RMSSD of 20 (± 34) %.

Regarding subgroup differences, mean fixation count was 389 (± 88) in novices and 384 (± 80) in experts, translating to 28 (± 6) fixations per minute in both groups. Novices took on average 37 (± 27) seconds (extracted from complete novice n = 7 group) to fixate the monitoring equipment for the first time, whereas mean time to first fixation in the expert group was 41 (± 37) seconds.

While the average fixation duration in novices was detected at 295 (± 30) ms, experts took on average 419 (± 133) for a fixation (p = 0.0556).

In total, novices spent a mean of 1.9 (± 0.4) minutes and experts 2.7 (± 1.1) minutes fixating the AOI. In relation to the timespan that the AOI was active, novices spent 13 (± 4) % (extracted complete novice n = 7 group) of this interval fixating the monitoring equipment, in contrast to the expert group, which fixated the equipment for 20 (± 8) % of the time (p = 0.1061, Figure 13)

A mean of 619 (± 68) saccades were registered in the novice group, equaling 45 (± 5) saccades/min, whereas experts showed on average 527 (± 150) saccades during the whole scenario, translating to 38 (± 11) saccades/min (p = 0.4206 for saccade rate) The mean amplitude of average saccade was 5.3 (± 1.1) degrees in novices and 4.1 (± 0.4) deg in experts (p = 0.0952), with an average peak velocity of 113.7 (± 28.1) and 98.7 (± 20.1) deg/sec respectively (p = 0.4206, Figure 14). A summary of measured values per subgroup is provided in supplementary table 1.

Regarding blink analysis, participant A4 was excluded from the analysis of the overall scenario due to recurring loss of the eye tracking stream and therefore missing data. Within the 11 other participants, mean blink count during the scenario was 329 (± 139) blinks, with a blink rate of on average 24 (± 11) blinks per minute. Split into subgroups, a mean of 327 (± 136) blinks was detected in the novice group (n = 6), translating to a blink rate of on average 25 (± 11) blinks per minute. In contrast, experts showed on average 331 (± 159) blinks throughout the scenario, with a mean blink rate of 24 (± 11).

Average heart rate in novices during the complete scenario was 97 (± 13) bpm, an increase of 22 (± 15) % in comparison to baseline. Mean RMSSD was 29 (± 12) ms, reflecting a 4 (± 36) % decrease. In experts, the average heart rate was 100 (± 11), equaling a 32 (± 7) % increase from baseline, with an average RMSSD of 21 (± 7) ms (p = 0.2020, for comparison to novice group), which is a decrease of 43 (± 9) % compared to baseline.

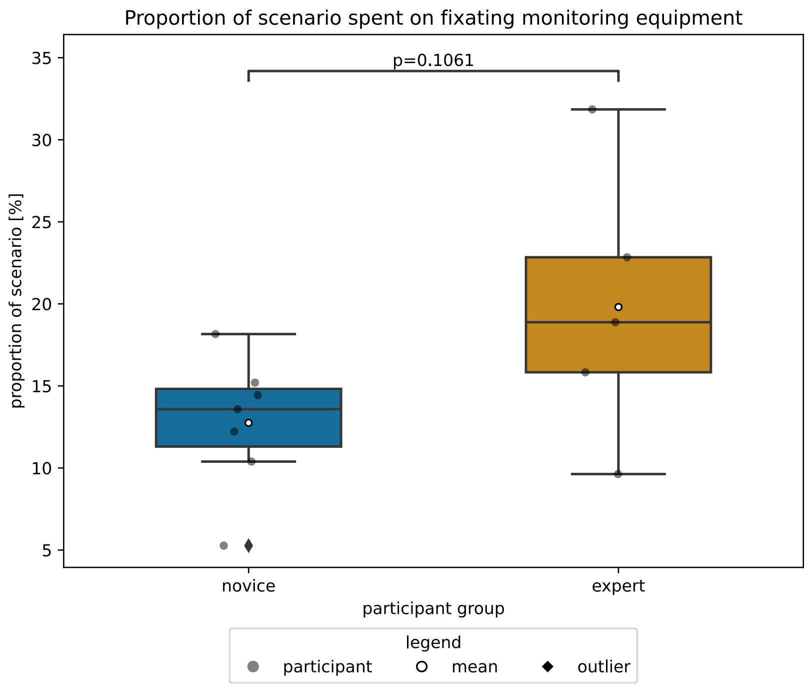

Figure 13 - Proportion of overall scenario both groups spent on fixating the monitoring equipment. A trend towards the expert group spending a larger amount of time during the scenario fixating on the monitoring equipment is visible.

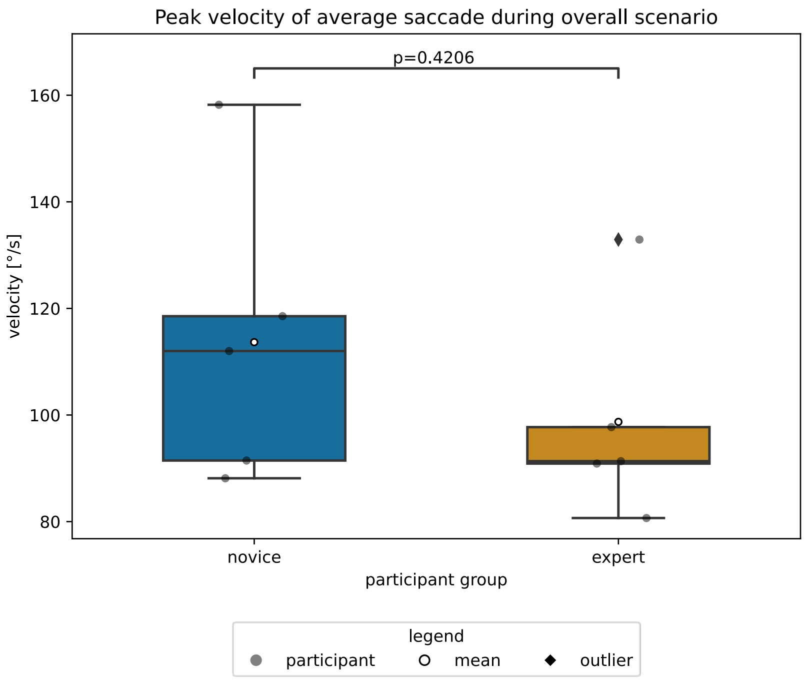

Figure 14 - Peak velocity of average saccade during overall scenario for study subgroups. Velocity reported in degrees of visual field per second.

4.2.1.2 Sub analyses per TOI

4.2.1.2.1 First medical contact

During the first medical contact TOI, representing the first 4.5 minutes of the scenario, participants exhibited a mean fixation count of 147 (± 44) times on the AOI, with a total fixation duration of 0.8 (± 0.3) minutes, corresponding to 18 (± 7) % of the first contact TOI and 33 (± 10) fixations per minute. The mean time to initial fixation remained unchanged at 39 (± 30) seconds.

Across the study population, the average saccade count stood at 216 (± 72) – 48 (±16) saccades/min – accompanied by an average saccade amplitude of 4.6 (± 1.2) deg and an average peak velocity of 111.5 (± 35.4) deg/sec. A mean total of 130 (± 64) blinks, with a mean blink rate of 29 (± 14) blinks per minute, was recorded.

Within the novice's subgroup, fixation occurrences on the AOI averaged 147 (± 40) times, equal to 33 (± 9) fixations per minute, with a cumulative fixation duration of 0.7 (± 0.2) minutes, representing approximately 15 (± 5) % of the first contact TOI.

Novices typically achieved their initial fixation within an average time of 37 seconds (± 27) and an average fixation lasted for 276 (± 53) ms.

This subgroup demonstrated an average saccade count of 225 (± 49) – with a saccade rate of 50 (± 11) per minute – and an average saccade amplitude of 5.2 (± 1.0) deg. Mean peak velocity was measured at 123.1 (± 43.0) deg/sec.

Mean number of blinks and mean blink rate in novices was 128 (± 58) and 29 (± 13) blinks per minute respectively.

The expert subgroup exhibited an average fixation count of 147 (± 53) instances in total and 33 (± 12) instances per minute on the AOI, totaling 1.0 (± 0.4) minutes, which accounted for approximately 22 (± 9) % of the first contact TOI, a proportion similar to novices (p = 0.3434). Their mean time to initial fixation was 41 (± 37) seconds, mean fixation duration was 404 (± 90) ms, noticeably longer than in novices (p = 0.0101, Figure 15).

Experts demonstrated an average saccade count of 204 (± 101) – reaching 45 (± 23) saccades per minute (p = 1.000 for comparison to novice group) – and an average saccade amplitude of 3.8 deg (± 0.9), noticeably smaller than in comparison to novices (p = 0.0303, Figure 16) with a mean peak velocity of 95 (± 11) deg/sec (p = 0.2020 for comparison to novice group).

As for blinks, experts blinked for a mean number of 132 (± 79) times during first medical contact, with a blink rate of 29 (± 18) blinks per minute.

Concerning heart rate, the overall group showed a 24 (± 10) % increase from baseline, with a mean of 97 (± 11) bpm. The novices' heart rate rose to 97 (± 12) bpm, equaling a relative increment of 21 (± 11) %. In experts, mean heart rate was detected at 98 (± 11) bpm, representing an increase of 29 (± 4) % in comparison to baseline.

Mean RMSSD in the overall group during first contact was 29 (± 11) ms, with a value of 32 (± 12) ms for novices and 23 (± 6) ms for experts (p = 0.2677, Figure 18). In the overall population, RMSSD decreased for a mean of 10 (± 35) % in comparison to baseline. In the novice subgroup, due to a strong outlier showing a

71% increase, relative mean RMSSD change was positive with an increase of 7 (± 38) %, despite a reduction in the absolute mean RMSSD of 4 (± 15) ms. Experts, in contrast, showed a 34 (± 11) % decrease. A tabular overview of results is also provided in supplementary table 2.

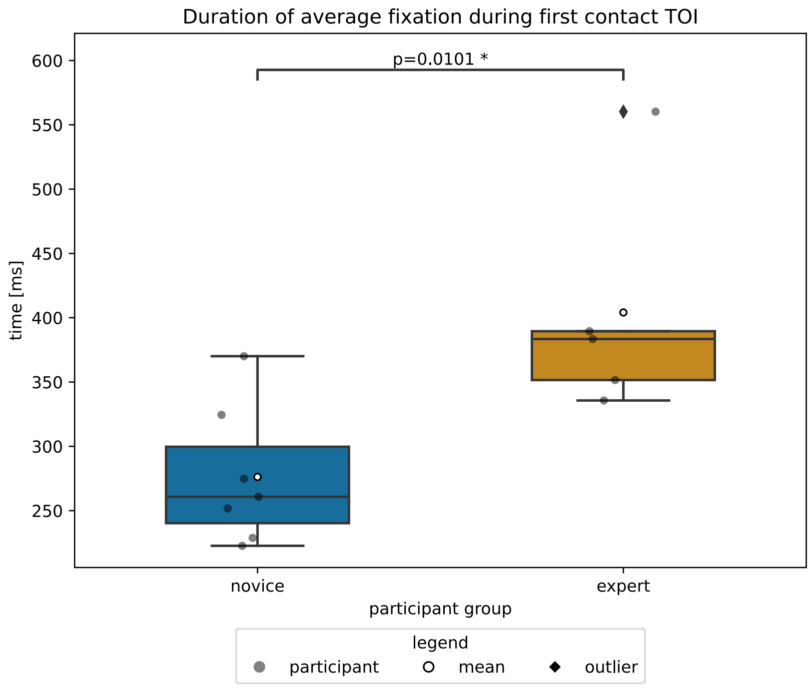

Figure 15 - Average duration of fixation during first contact time of interest for study subgroups. Experts showing longer fixation durations, potentially reflecting a more attentive process. Ms: Milliseconds, TOI: Time of interest

4.2.1.2.2 Cardiac arrest

For the cardiac arrest TOI, please note that fixation duration in this context is reported in seconds. Also, as the cardiac arrest TOI had a predefined timespan of 1 minute, counts translate to rate/min. The overall study group exhibited an average fixation count of 41 (± 16) during arrest phase, corresponding to an absolute average fixation time of 15 (± 7) seconds. Mean relative fixation time therefore was 24 (± 11) %. Mean time to first fixation during cardiac arrest was 1.3 (± 2.4) seconds. Saccades averaged to a mean of 63 (± 28), while average saccade amplitude was 5.1 (± 2.1) deg, with a mean peak velocity of 115.2 (± 48.2) deg/sec. As for subgroups, in novices, on average 46 (± 15**)** fixations of the AOI were detected. Average fixation time was 14 (± 5) seconds, translating to a relative fixation time of 23 (± 9) % of the cardiac arrest TOI. Mean time to first fixation of the monitoring equipment during arrest phase was 1.2 (± 2.8) seconds.

An average fixation lasted for 306 (± 85) ms in novices and 434 (± 141) ms in experts (p = 0.0732)

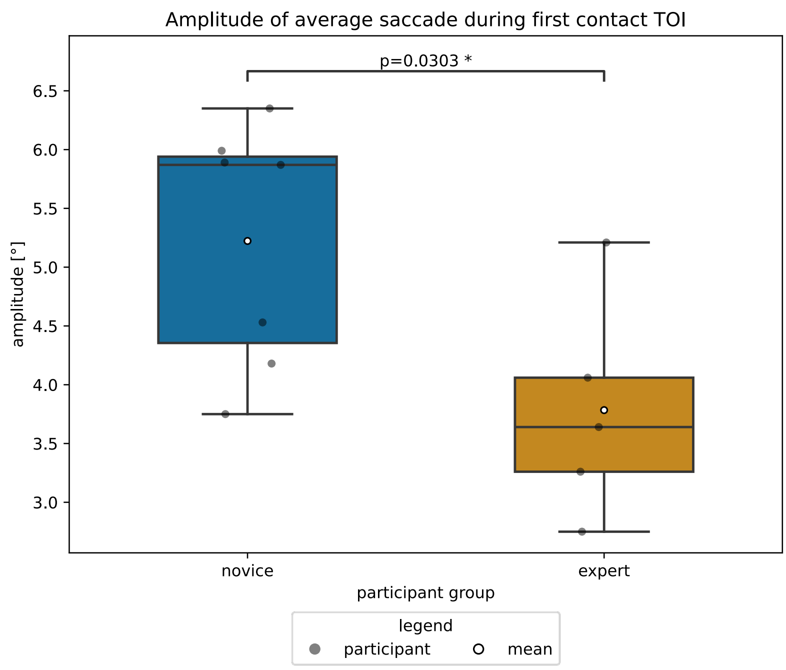

Figure 16 - Amplitude of average saccade during first contact time of interest for study subgroups. Amplitude reported in degrees of visual field. TOI: Time of interest

In novices, an average of 73 (± 23) saccades, with a mean amplitude of the average saccade of 6.0 (± 1.7) deg and a mean peak velocity of 130.3 (± 58.5) deg/sec, were registered.

In contrast, experts fixated the monitoring equipment a mean number of 34 (± 15) times, summing up to an average fixation time of 15 (± 9) seconds. This group spent 26 (± 15) % of the TOI duration on fixating the AOI (p = 1.0000 for comparison to novices), with a mean time to first fixation of 1.4 (± 1.8) seconds. A mean of 50 (± 30) saccades was performed by the experts (p = 0.0732 in comparison to novices' saccade rate) with an average saccade amplitude of 3.9 (± 2.0) deg (p = 0.1061 for comparison to novices) and an average peak velocity of 94.0 (± 17.1) deg/sec (p = 0.3434).

The overall study group showed a mean blink count of 22 (± 10) blinks during this TOI, with a subgroup mean blink count of 23 (± 10) blinks in novices and 21 (± 11) blinks in experts respectively.

Focusing on heart rate analysis, the cardiac arrest phase led to a 28 (± 16) % increase in heart rate in the overall group, with a mean of 100 (± 15) bpm. Novices' HR rose to 100 (± 16) bpm on average, equaling a 25 (± 20) % increase. In contrast, experts showed a stronger relative increase of 33 (± 7) %, with a mean HR of 101 (± 15) bpm.

Looking at HRV, the overall group showed an RMSSD of 20 (± 9) ms, with 22 (± 9) ms in novices and 18 (± 8) ms in experts (p = 0.4318, Figure 18). This reflects a relative RMSSD change of -36 (± 33) % in the overall group, -23 (± 38) % in novices, and -54 (± 8) % in experts respectively.

For a tabular overview of results refer to supplementary table 3.

4.2.1.2.3 Intubation

During intubation, neither novices nor experts visited the monitor AOI, as they were focused on the airway of the patient. Thus, relative fixation time was not included in conclusive statistical analysis for this TOI. The intubation process took on average 24 (± 8) seconds in the overall group. The evaluated timespan was defined as the moment in time when a participant picked up the laryngoscope until the point in time when the endotracheal tube was successfully inserted, and the laryngoscope removed.

Fixations on any point in the field of view reached an average of 33 (± 17), corresponding to 83 (± 27) fixations per minute in the overall group. Total time spent fixating a point was 11 (± 6) seconds for all participants, corresponding to 50 (± 25) % of the intubation process.

An average of 80 (± 57) saccades, equaling 195 (± 95) saccades/min, were detected, with an average saccade amplitude of 5.9 (± 2.3) deg and an average peak velocity of 110.8 (± 45.1) deg/sec. Concerning blinks, the number of blinks was drastically reduced in comparison to the overall scenario or other TOIs, with an average blink number of 2 (± 3) and a blink rate of 6 (± 6) blinks per minute in the overall group.

Novices needed a mean of 26 (± 8) seconds to intubate the patient (Figure 17). They showed an average fixation count of 41 (± 19), with a mean fixation rate of 93 (± 20) per minute, and spent 13 (± 5) seconds fixating a point in the field of view, equaling 54 (± 22) % of the intubation procedure.

Average fixation duration in the novices was 360 (± 151) ms, similar to experts, which showed a mean duration of 385 (± 276) ms (p = 1.0000)

A mean of 89 (± 66) saccades, with a saccade rate of 193 (± 91) saccades per minute on average and an average saccade amplitude of 6.6 (± 2.8) deg were recorded. Average peak velocity was measured at 114.7 (± 56.8) deg/sec. This group blinked a mean time of 3 (± 4) times with a blink rate of 6 (± 7) blinks per minute during intubation, showing a drastic reduction of blinks during this TOI.

In comparison, it took the experts 23 (± 8) seconds on average to perform intubation (Figure 17). Experts' average blink count was 2 (± 2) blinks, also with a blink rate of 6 (± 6) blinks per minute. Regarding fixations and saccades, the expert group showed on average 23 (± 2) fixations and 68 (± 45) saccades, with respective average rates of 69 (± 31) per minute for fixations and 198 (± 111) per minute for saccades ($p = 0.7551$ in comparison to novice group saccade rate). Average saccade amplitude and peak velocity in this group were 4.9 (± 0.7) deg ($p = 0.4318$ for comparison to novices) and 105.3 (± 26.1) deg/sec ($p = 0.8763$ for comparison to novice group).

Total time spent fixating in the expert group was 9 (± 6) seconds, translating to 43 (± 29) % of the whole duration of the intubation procedure ($p = 0.6389$ in comparison to novices)

As for HR and HRV, the overall groups' average heart rate showed the strongest increase of all TOIs in comparison to baseline, with 104 (± 14) bpm, a relative increase of 33 (± 17) %. Novices' mean heart rate during this phase was 101 (± 14) bpm, increasing by 26 (± 14) %. Experts HR was slightly higher on average, with 108 (± 15) bpm, reflecting a relative increase of 42 (± 18) %

Mean RMSSD of the overall group was 15 (± 15) ms, with a relative decrease of 59 (± 22) % in comparison to baseline. Novices showed a mean RMSSD of 19 (± 19) ms and a relative mean decrease of 48 (± 23) %, whereas experts' RMSSD fell for a mean of 74 (± 7) %, to 10 (± 6) ms ($p = 0.5303$ for comparison to novices, Figure 18). Supplementary table 4 provides an overview of results.

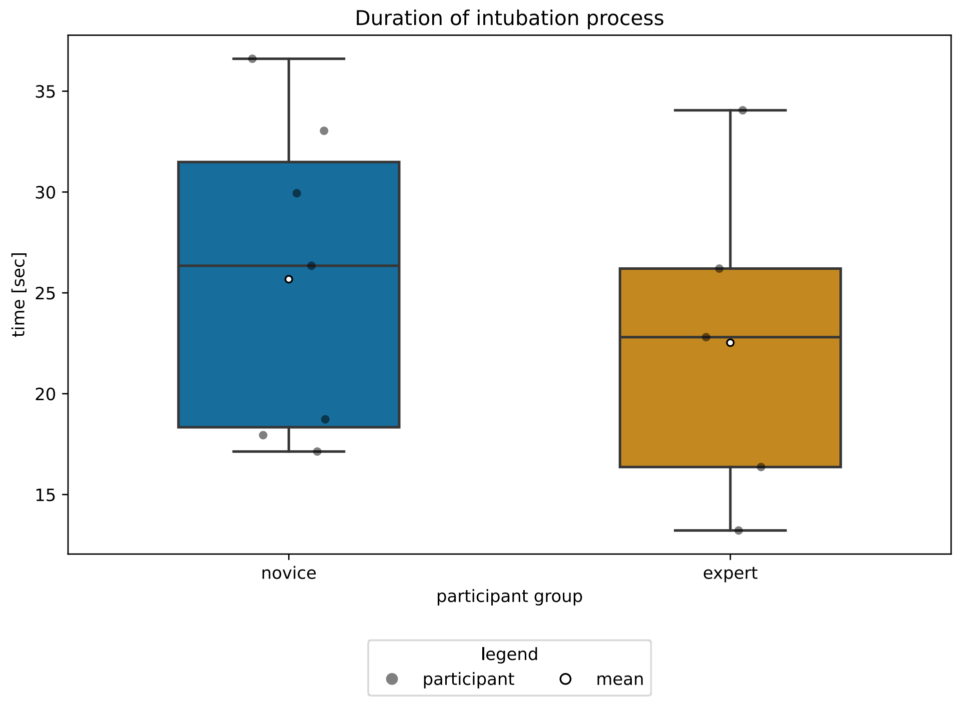

Figure 17 - Intubation duration in novice and expert groups. Sec = seconds.

4.2.1.2.4 Distraction

For the ten-second distraction phase, initiated by a trash can falling over to produce an acoustic distraction, the overall group showed an average of 19 (± 8) fixations, translating to 116 (± 47) fixations per minute, or 4 (± 2) seconds, which accounts for 41 (± 16) percent of the distraction TOI. Saccades were detected for an average amount of 39 (± 8), with an average saccade rate of 233 (± 48) / min, the highest saccade rate of all TOIs in this study. Average saccade amplitude was detected at 7.9 (± 1.6) degrees, average peak velocity at 165.5 (± 43.9) deg/sec. For blinks, the overall group showed an average blink count of 4 (± 2), translating to 22 (±12) blinks per minute.

Divided into subgroups, novices fixated any point in the field of view for 22 (± 8) times, corresponding to 134 (± 47) fixations per minute, equally accounting for 4 (± 1) seconds or 42 (± 14) % of the TOI duration. Mean fixation duration was detected at 194 (± 59) ms.

Saccade wise, a mean number of 40 (± 9), or 242 (± 54) saccades per minute, were detected, with an average saccade amplitude of 7.2 (± 1.3) deg and a mean peak velocity of 157.0 (± 33.4) deg/sec. Regarding blinks, the novices blinked on average 4 (± 2) times during distraction, whereas mean blink rate was 25 (± 14) blinks per minute.

In contrast, experts showed a mean of 15 (± 6) fixations during distraction TOI, translating to a fixation rate of 91 (± 36) per minute, accounting for 39 (± 20) % of the TOI duration ($p = 0.7551$ for comparison to novices), equaling 4 (± 2) seconds. Mean fixation duration was slightly higher at 241 (± 88) ms in comparison to novices ($p = 0.2677$).

On average, 37 (± 7) saccades, or a rate of 221 (± 40) saccades per minute ($p = 0.5303$ for comparison to novice saccade rate), were detected. Average saccade amplitude was 9.0 (± 1.4) deg ($p = 0.0732$ for comparison to novices), with a mean peak velocity of 177.5 (± 57.6) deg/sec ($p = 0.6389$ for comparison to novices). The expert subgroup blinked a mean of 3 (± 2) times, with a blink rate of 18 (± 10) per minute.

Concluding again with HR and HRV, mean HR was 98 (± 14) bpm in the overall group, representing an increase of 25 (± 17) %, during distraction TOI. Split into subgroups, novices had an average HR of 95 (± 17) and experts of 102 (± 9) bpm. This represents a relative increase of 19 (± 20) % and 34 (± 4) % in subgroups respectively.

Average RMSSD in the overall study group was 23 (± 17) ms, in subgroups 28 (± 21) ms and 15 (± 6) ms for novices and experts respectively ($p = 0.5303$, Figure 18). Relative mean changes from baseline were -23 (± 62) % in the overall group, and -54 (± 20) % in experts. Throughout the 7 novices, again due to strong outliers in both directions, the mean of RMSSD changes was only 0.3 (± 74) % despite an absolute change of a mean of 8 (± 31) ms.

A tabular overview of results can be found in supplementary table 5.

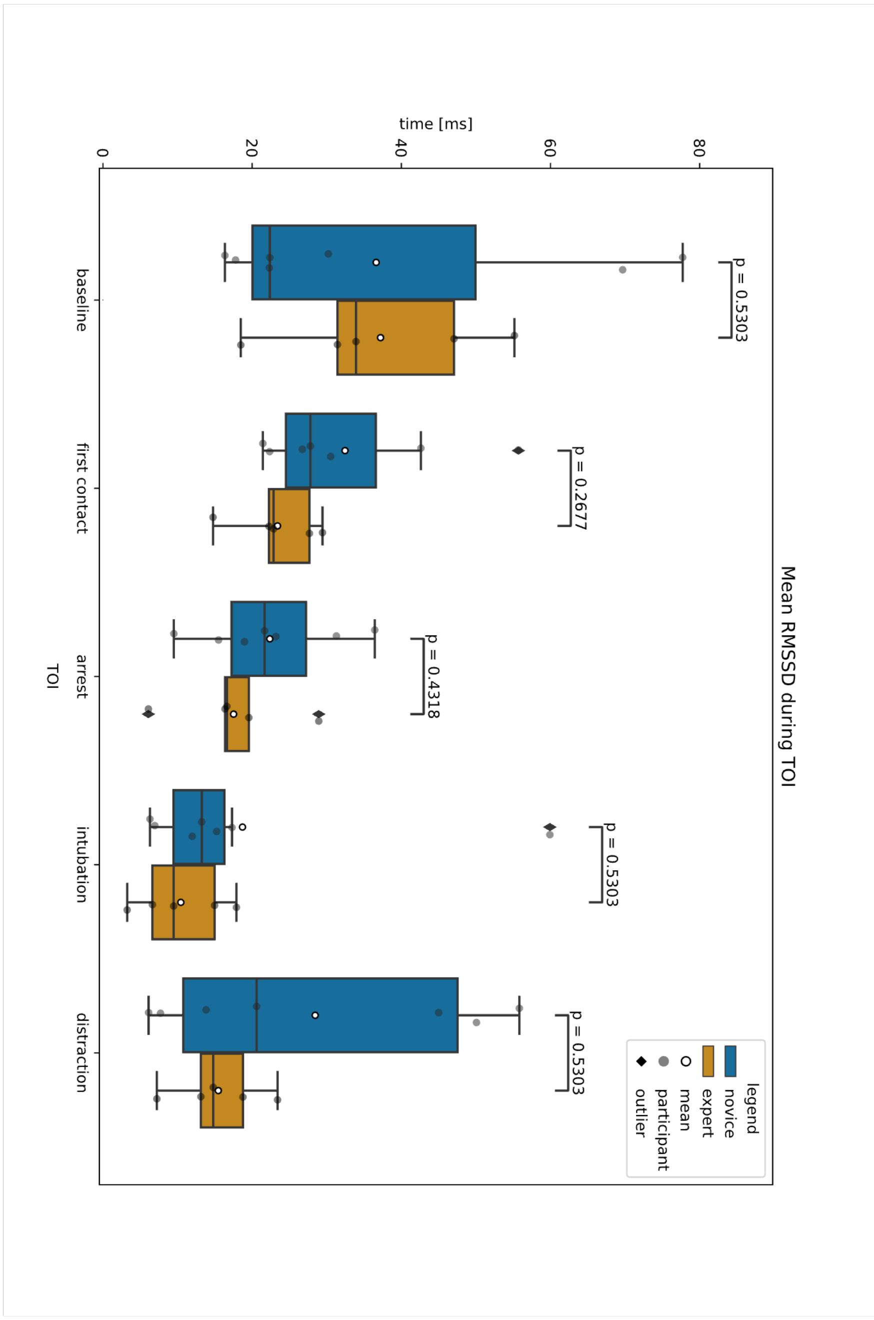

Figure 18 - Mean RMSSD during TOIs for study subgroups. Ms: milliseconds, RMSSD: Root mean square of successive differences, TOI: Time of interest

4.2.2 GSR

The initial intention of GSR data collection was to generate another potential source of objective data that might provide insight into participants' cognitive load. Unfortunately, after study execution, the data was found to be highly compromised by motion artifacts, rendering a valid analysis with the available algorithm, integrated in the used analysis software, impossible, as massive over-detection of peaks occurred due to motion (Figure 19). Altering peak detection threshold also did not lead to relevant improvement in peak detection. As development of a proprietary algorithm to analyze GSR data was beyond the scope of this thesis, GSR data are not reported here.

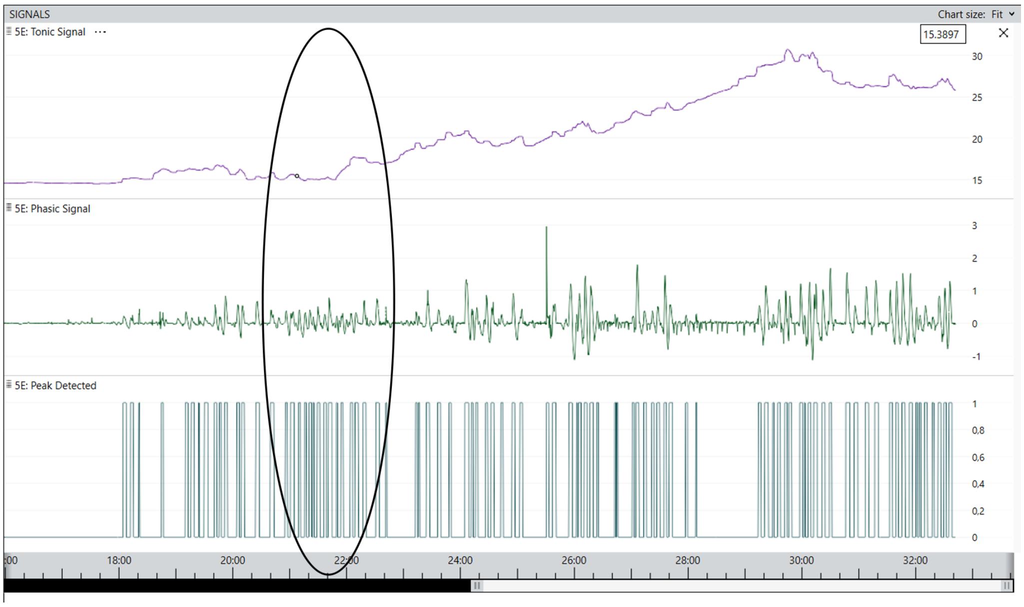

Figure 19 - GSR Signal with over-detection of phasic peaks. Strong over-detection of phasic peaks occurring due to motion artifacts. While no peaks are visible in the tonic component, the phasic signal shows strong alterations during motions, falsely detected by the software as phasic peaks (marked in ellipse). GSR: galvanic skin response

4.3 Subjective ratings of cognitive load

Regarding results of the NASA TLX evaluation, the total weighted rating for the executed simulation scenario was 56 points (± 15) in the overall study group, 59 (± 11) for novices, and 52 (± 20) for experts (p = 0.5303), respectively. Mental demand received a weighted mean score of 272 (± 132) overall, 352 (± 85) in novices and 161 (± 102) in experts (p = 0.0145, Figure 20)

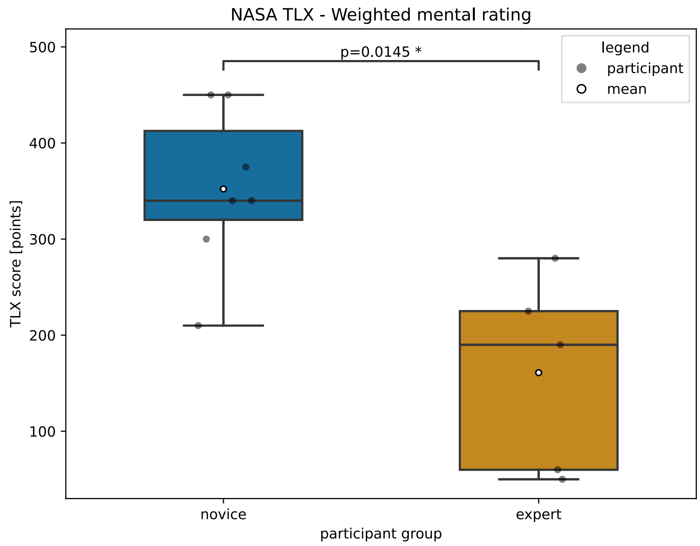

Figure 20 - NASA TLX weighted mental rating for study subgroups. NASA: National Aeronautics and Space Administration, TLX: task load index

For the prespecified TOIs, participants' mean PAAS cognitive load responses shall be provided:

- Time of Cardiac arrest: A mean rating of 5.2 (± 1.4) in the overall group, 5.3 (± 1.7) in novices, and 5.2 (± 0.8) in experts was documented.
- Time of acoustic distraction: Overall group 3.8 (± 1.9), novice group 4.4 (± 2.1), and expert group 3.0 (± 1.2).
- Time of intubation: Overall group 4.2 (± 2.1), novice group 4.9 (± 1.9), expert group 3.2 (± 2.2)

Subjective ranking of cognitive load, with cardiac arrest rated as the most challenging part, followed by intubation and then distraction, hence was different to objective parameters, such as e.g. RMSSD or peak saccade velocity (Figure 21).

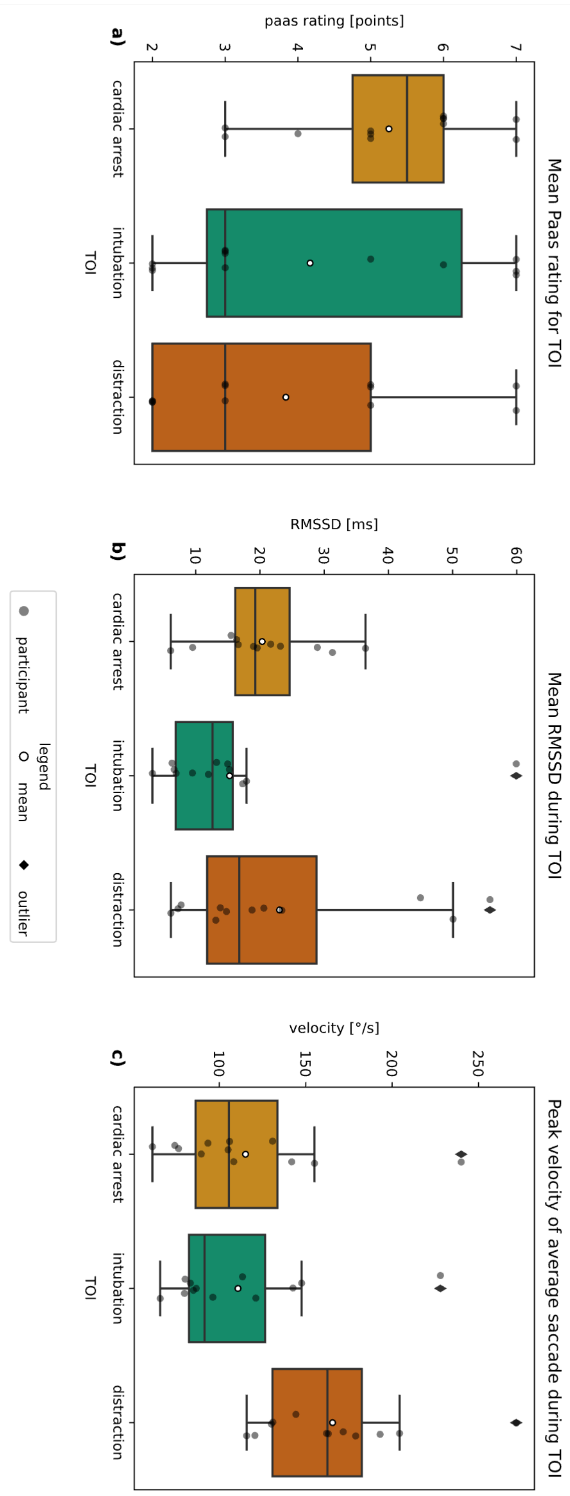

Figure 21 - Comparison of different parameters for cognitive load during TOIs. a) Paas rating b) RMSSD c) Peak velocity of average saccade. Ms: Milliseconds, RMSSD: root mean square of successive differences, TOI: time of interest. Velocity reported in degrees of visual field per second.

4.4 Intragroup differences

In contrast to the largely similar numbers between groups, more accentuated differences were detected in various parameters in the respective subgroups between TOIs. For example, duration of average fixation increased steadily from first contact (329 ± 94 ms) to cardiac arrest (359 ± 125 ms) and intubation (371 ± 201 ms) time of interest and then showed a sharp decline during distraction phase (213 ± 73 ms) (Figure 22).

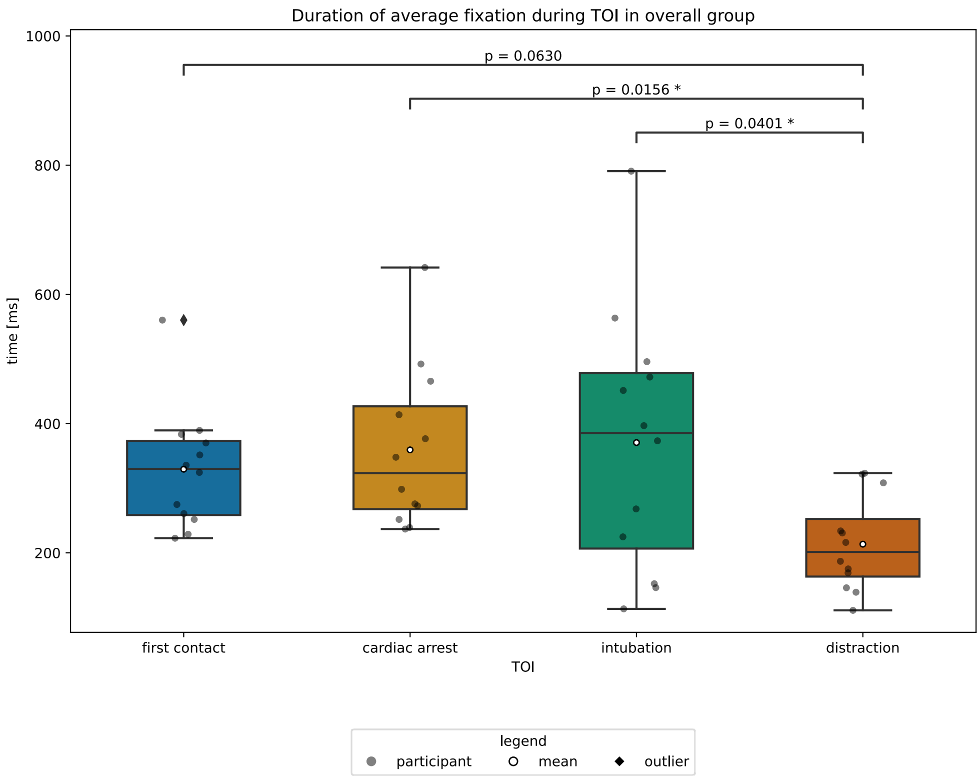

Figure 22 - Duration of average fixation during predefined times of interest (TOIs). Ms: milliseconds

Similarly highest average saccadic peak velocity was again detected during distraction (166 ± 44 deg/s), followed by cardiac arrest (115 ± 48 deg/s), first medical contact (112 ± 35 deg/sec), and lastly intubation (111 ± 45) (Figure 23).

As another example, comparing the relative time spent on fixating the monitoring equipment in the overall group between first contact and cardiac arrest TOIs, a 6% increase occurred: Probands spent an average of 18 (±7) % of first contact TOI and 24 (± 11) % of cardiac arrest TOI fixating monitoring equipment.

Analysis of RMSSD also draws a similar picture, with a noticeable decrease during all TOIs in comparison to baseline (37 ± 21 ms), especially accentuated for cardiac arrest (20 ± 9 ms) and intubation (15 ± 15 ms) TOIs (Figure 24).

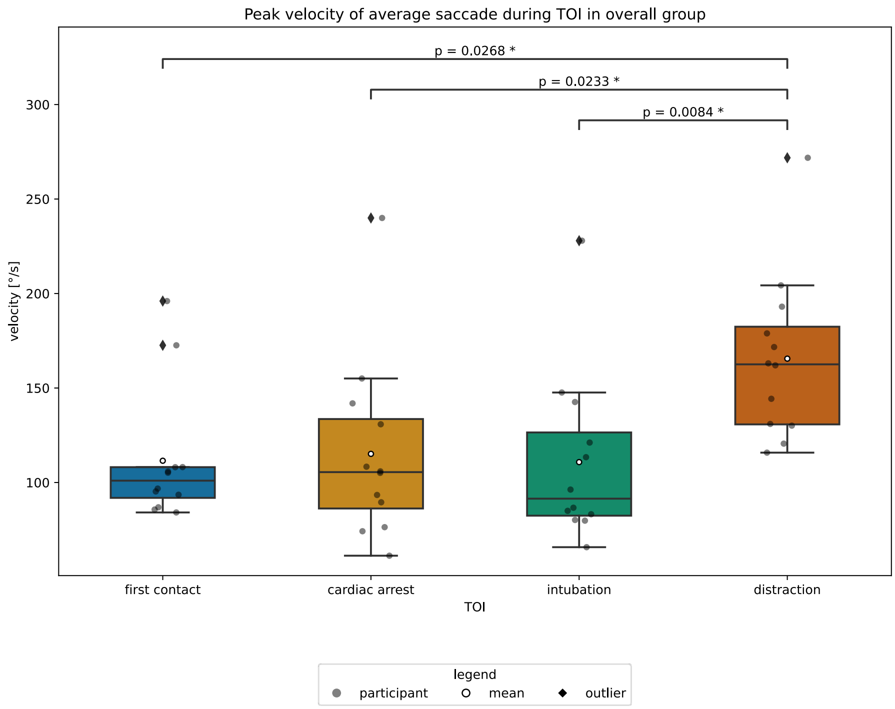

Figure 23 – Peak velocity of average saccade during times of interest in the overall study group. TOI: time of interest. Velocity reported in degrees of visual field per second.

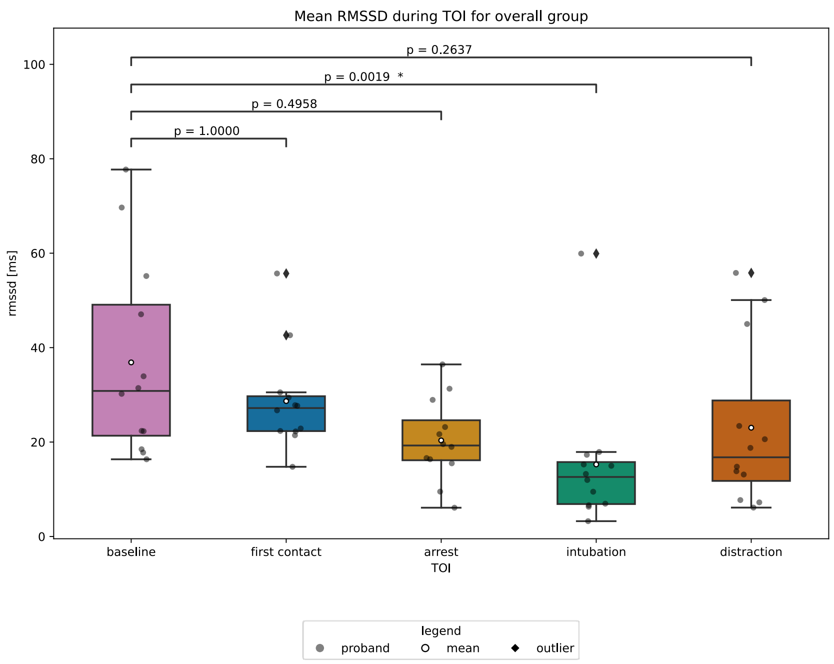

Figure 24 – Mean RMSSD during TOIs for overall study group in comparison to baseline. A trend to reduction in RMSSD is visible. Ms: Milliseconds, RMSSD: root mean square of successive differences, TOI: time of interest

4.5 Sample size calculation

For all following sample size calculations, two-tailed level of significance was set at $\alpha = 0.05$. All calculations were performed to detect significant differences with a power of $1 - \beta = 0.8$. Method of choice, as Mann-Whitney-U tests were used, was sample size calculation for non-parametric tests as published by Noether using probability of superiority (P(X>Y)) as effect size. [143] According to Grissom, equivalents for small, middle, and large effects are a P(X>Y) of 0.56, 0.64, and 0.71 respectively. [144] An openly available online calculator was used for performing calculations, accessible under [145].

4.5.1 Proportion of scenario spent fixating monitoring equipment

Using the acquired data reported above, with novices tending to spend less time on the monitoring equipment throughout the scenario, effect size calculation revealed a P(X>Y) of 0.2262, resulting in a total sample size of N = 34 participants (17 in each group) necessary to detect a statistically significant difference between groups regarding time spent on monitoring equipment.

4.5.2 Saccade amplitude

As a difference in average saccade amplitude between novices and experts was visible during the first contact TOI, those results were used for sample size calculation to detect a significant difference between groups for this potential objective marker of cognitive load. Calculation of effect size revealed a $P(X>Y)$ of 0.8489, resulting in a sample size of $N = 22$ (11 in each group) necessary to detect a statistically significant difference.

4.5.3 NASA TLX

Lastly, to also assess the sample size necessary for evaluating a subjective cognitive load parameter, results of the NASA TLX mental load subset was used for effect and sample size calculation, revealing a $P(X>Y)$ of 0.9253 and a necessary sample size of $N = 16$ (8 in each group).

4.6 Feasibility

4.6.1 Eye tracking Equipment

Eye tracking equipment was generally well accepted with the majority of participants agreeing quite strongly that they forgot that they were wearing the equipment during the simulation scenario. Mean value for this questionnaire item was 1.8 (± 1.4) on a 5-point Likert scale, where a value of 1 would correspond to “fully agree”, whereas a value of 5 represented “fully disagree”. With a mean value of 1.9 (±1.4), participants also appeared quite willing to wear the equipment during a real emergency medicine scenario.

As for comfort, participants mostly disagreed with statements asking them if the Eye tracker e.g. distracted them from medical tasks (e.g. mean 4.8 (± 0.4) for the question if the Eye tracker was distractive during intubation) or that the device felt uncomfortable for eyes (mean 4.5 (± 0.9)), ears (mean 4.7 (± 0.7)) or nose (mean 4.8 (± 0.4)). Figure 25 provides an overview of eye tracking feasibility results.

4.6.2 GSR Sensor

Similar to the eye tracking device, the GSR sensor was received well by participants, with a strong agreement to the question if they forgot that they were wearing the sensor (mean 1.3 (± 0.5)). Probands strongly disagreed when asked if they behaved differently than normal (mean question score 4.6 (±1.2)) or if they experienced any pain (mean 4.8 (± 0.4)).

4.6.3 Heart rate sensor

Feasibility assessment of heart rate measurements using a chest strap was not performed, as it is nowadays a well-accepted approach in sports, science, etc. [146]

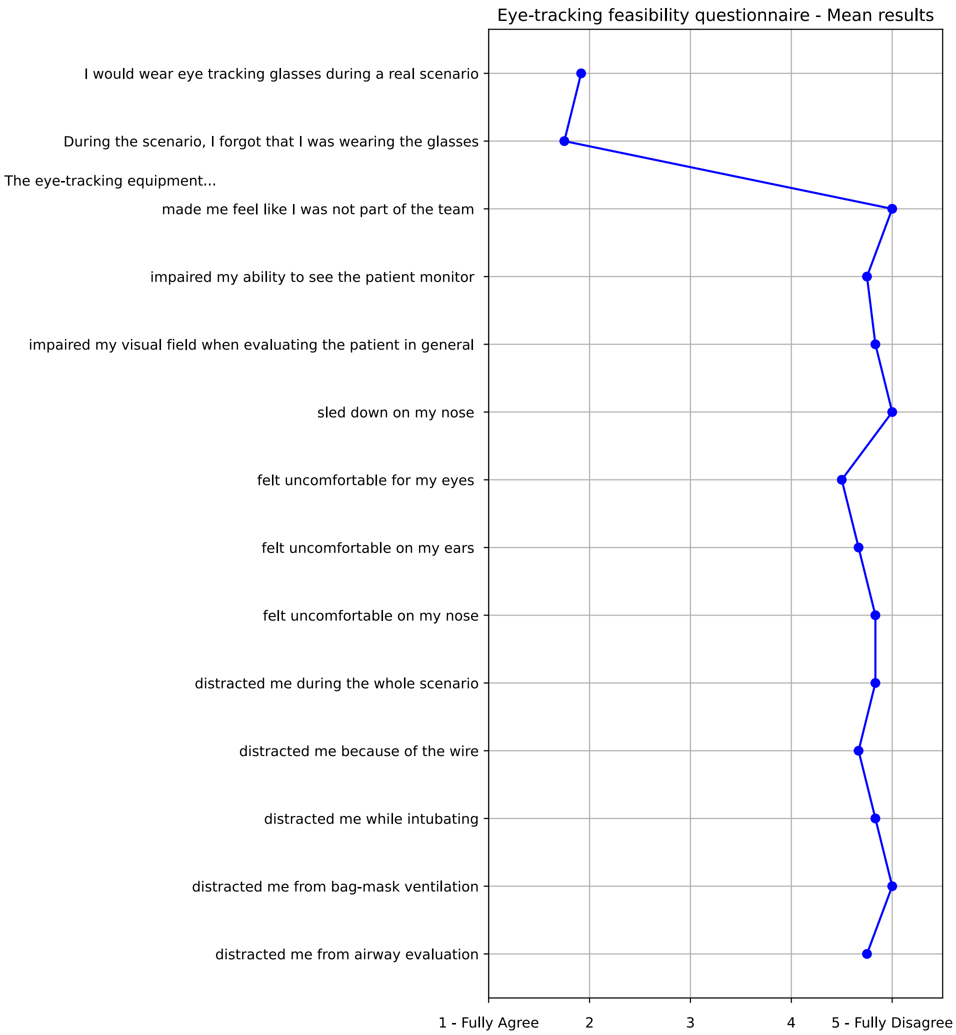

Figure 25 - Mean results of eye tracking feasibility evaluation for n = 12 participants. Equipment usage appears highly feasible.

5 Discussion

As discussing all the results acquired during this study would be beyond the scope of this document, the above-formulated research questions shall be addressed in systematically in this section with a discussion of relevant data. Also, as this was a pilot project, statistical results must be interpreted with uttermost caution, and properly sized studies are necessary to further investigate the current results.

5.1 Question 1 - Do gaze patterns and areas of visual focus differ between medical experts and novices during a simulated training scenario?

As stated above, differences between expert and novice decision-making have been reported, with novices tending to rely more on objective data and experts seeming more capable of subconscious pattern recognition. [3], [78] It was therefore hypothesized that experts might spend less time looking at the provided monitoring equipment, as they might extract relevant information faster and focus more on the patient.

As marked in the results section, this is not visible in the current study population, as experts showed a tendency towards spending even more time looking at the equipment. This could be due to various reasons. First, experts might focus more on the area that gives the most relevant information, as also reported in the previously mentioned work with chess players. [58] Second, a simulation mannequin cannot (yet) present with all the signs and symptoms of a patient in a life-threatening condition, thereby creating a bias, as necessary information for case handling, that would otherwise be extracted by examining a patient, is collected by looking at auxiliary equipment.

Third, as this was a small sample size project, differences could always be due to chance or due to data quality. With a mean of 87.4% valid data in experts, and 84.8 % in novices, this seems less likely though. A larger study would be necessary to evaluate if this trend would reach significance.

Looking at fixation and saccade duration, it has been reported that the human mind may switch between two states, local and global visual attention. [147] While the latter somewhat represents the activity of identifying the position of relevant information in a visual stimulus, the former describes the process of exploring the area marked as relevant in more detail. [147] This has also been referred to as “What” and “Where” pathways of visual attention. The former was described to be associated with longer fixations and shorter saccades in comparison to the latter. [148]

Focusing on fixations, a trend of a longer duration of fixations in experts was visible, during the overall scenario and certain times of interest. Especially during the first contact phase, experts’ fixation duration is well above the duration of fixations during “preattentive” states, with modal values of around 220 ms, as reported by Velichovsky et al. The authors state though, that in their work, the mean values of fixation durations were not significantly different between multiple variations of a simulated driving scenario, but it was the modal values that showed significant differences. [52] Due to the small sample size of this work, still, mean values rather than modal values are reported.

This longer mean fixation time in experts could represent an attentive process with a focus on relevant information. Interestingly, novices’ average fixation duration during first contact was located more in the abovementioned modal range reported above, possibly reflecting a “search for information” (i.e. the “where” pathway). If this difference is due to the small sample size or represents a real difference in attention allocation remains to be elucidated. During the more challenging task of intubation, for example, fixation duration again showed more similar values, potentially indicating similar focus.

Regarding saccades in general, during the overall scenario, saccade rate, duration, and amplitude appeared similar between the two groups, with a trend in experts performing shorter and smaller saccades. Absolute saccade count was not compared, as scenario duration slightly varied from participant to participant. Solely in the first contact phase, experts appear to have a larger saccade amplitude in this work. The possible association between saccade amplitude and cognitive load is discussed below. If an appropriately sized and powered study could confirm this trend remains subject to further investigation.

Worth mentioning, work in monkeys has also shown that saccade velocity and amplitude reflect performance in saccadic choice tasks, with higher velocity and amplitude values for correct saccadic choices. [149] Interestingly, in their work, Seideman et al. report much higher saccadic velocities than seen in this study, maybe reflecting again the more pronounced process of analyzing information already located, i.e. on the monitoring equipment, than searching for it. [149]

To conclude this section with the discussion of blink metrics, both groups showed a drastic reduction in blinks during the intubation phase in contrast to the rest of the scenario, with similarly low numbers of blinks, possibly because of the mentally and motorically difficult task of securing an airway.

Summing up the acquired data, it can be stated that in this trial, differences in visual behavior between experts and novices were noticeable in some parts of the scenario, but not to the degree and in the form that was expected. Appropriately sized and powered studies are necessary in the future to further investigate this.

5.2 Question 2 - Does cognitive load potentially differ between experts and novices during a simulated medical scenario?

Section 2.4.1.3 outlined that eye tracking may be used for analysis of cognitive load including both fixation and saccade metrics.

As already mentioned, saccade metrics appeared similar during the overall scenario. This also was found in TOI comparison, except for the average saccade amplitude during first medical contact which appeared smaller in experts. As outlined above, this result could be interpreted in various manners, possibly indicating a higher load in the expert group.

This would not be necessarily contradictory. Indeed, e.g. in some work regarding cognitive load during geographic map analysis, it has been demonstrated that experts show a higher cognitive load during a task they well know, represented by shorter saccade amplitude and higher velocity: The authors conclude that experts might not experience less cognitive load during a specific task but still achieve higher success rates due to a, as in this case, more efficient way of scanning and retrieving data from a map, as represented by shorter saccades. The novices' longer saccades, in contrast, were interpreted as a more “chaotic search pattern”. [121]

As already mentioned above, mean fixation duration appeared, at the first contact phase, different between groups. This could on the one hand indicate a higher cognitive load, in this case in the expert group. On the other hand, this could be interpreted more as an effect of expertise, where highly skilled and experienced individuals know where to allocate their attention to gain the most information, in this case, the monitoring equipment.

As for RMSSD, again the intragroup differences between different TOIs were more striking than differences between experts and novices, as mean RMSSD values between those two groups appeared similar. Especially during intubation and cardiac arrest, RMSSD indicated the highest cognitive load.

Looking at the subjective evaluation of cognitive load, the overall scenario load using NASA TLX was quite similar for experts and novices, as described above. TLX covers multiple aspects, e.g. also physical workload, which was rated quite low in both groups as participants were asked in advance to refrain from physically exhausting activities, such as delivering chest compressions, to minimize movement artifacts on the GSR sensor. Also, it was expected that the performance sub rating would be similar, given the fact that most likely everybody would have wanted to save the patient. The sub-rating of cognitive load, which appeared higher in novices, might be connected to not yet perfectly internalized CPR algorithm steps, which could lead to higher subjective concentration efforts. This is not completely in line with the assessed objective parameters of cognitive load:

Using the defined questionnaires, the overall group – as well as the subgroups – ranked the tasks assessed with the PAAS cognitive load scale with the cardiac arrest task being the most mentally challenging, followed by the intubation and then distraction part. In contrast to this, RMSSD values and saccadic velocity, e.g., draw a different picture. Interpreting a smaller RMSSD as a higher load imposed, the smallest RMSSD was detected during intubation, followed by arrest and distraction TOIs

Saccadic velocity was highest during distraction, followed by arrest and, lastly, intubation. This might lead to the conclusion that, albeit not consciously registered by participants, the loud noise did cause a distraction that altered visual behavior and induced a search pattern, driving attention away from the patient. If this although is equal to an automatically elevated cognitive load remains subject to debate.

Aiming to find a (preliminary) answer to question 2, it could be stated that differences in cognitive load appeared to lie less on an inter-, but more on an intragroup level during TOIs. Despite very few metrics for some time points, objective load parameters seemed mostly similar between experts and novices. The scenario itself, though, imposed a decent amount of stress on both groups, with the experienced stress levels at distinct time points not always matching the results of objective data analyses.

5.3 Question 3 - What would be an appropriate sample size to further investigate differences in visual behavior and cognitive load

Necessary sample sizes for detecting statistically significant differences between experts and novices varied quite strongly with used measurement results, as expected. With a minimal sample size of 34 to evaluate differences in the proportion that each group spends on fixating monitoring equipment, this study size would also be sufficiently powered to detect differences in the selected objective and subjective parameters of cognitive load. Including 34 participants though would render the execution of such a study likely impossible in a medium-sized hospital like the one chosen for this pilot project. Recruiting participants from multiple hospitals, or, e.g., a large university hospital, would thus be necessary.

5.4 Question 4 – Is the combined usage of eye tracking and cognitive load equipment during medical simulation feasible?

The used equipment was equally well accepted by the expert as well as novice groups. Probands were only marginally or not at all influenced by the eye tracking equipment, with no participant stating that they were distracting during the scenario in a relevant manner. During this study, participants even stated that they almost forgot the were wearing eye tracking equipment and quite open to using such technology during a real-world emergency scenario.

An equal result was achieved for the GSR equipment, albeit it was not evaluated if participants would be willing to use it during a real emergency. An issue that has to be mentioned nonetheless, is signal quality. Although the electrode position used in this study has been reported to deliver valid GSR signals, the acquired data was highly compromised by motion and muscle artifacts, rendering an analysis with the existing “out of the box” analysis algorithm of the used software impossible. For future research alternative electrode positions and a tighter fixation of electrode cables should be considered.

As heart rate sensors have been used for decades and are also widely used and well-accepted by the general population, no specific questionnaire was designed to evaluate usage during this simulation study. No negative feedback regarding the chest belt was communicated from participants.

Summing up the collected evidence, the multi-sensor approach designed for this scenario may be deemed highly feasible, with eye tracking equipment even showing potential usage during real-world medical scenarios. Feedback not mirrored in the questionnaires but communicated orally by the participants after the scenario showed the most “concern” with the chest bag that was used for storing the eye tracking smart unit as well as the cable connecting glasses and smart unit. A more tailored approach concerning these parts, e.g. a more fitting storage solution for the smart unit, could even increase acceptance. GSR signal collection needs further refinement to compensate for motion and muscle artifacts.

5.5 Limitations

As any pilot work has weaknesses, so does this study. First, the sample size of 7 novices and 5 participants clearly is small. Results more reflect trends in the selected populations than actual statistically significant differences. Given the time available for planning and the limited resources regarding participant recruitment with only one study site and an obvious shortage of healthcare personnel, a study with a larger sample size would require more time and potentially recruitment in multiple healthcare facilities. Also, outliers identified in the dataset were still included in statistical analysis, as their exclusion would have shrunk the available data even more. Additionally, identifying an outlier reliably in such a small dataset is a very complex task per se.

Secondly, novice and expert groups were somewhat heterogeneous. A stricter recruitment strategy, e.g. potentially demanding at least a decade of experience in a true “expert” group, would maybe have shown more accentuated differences in groups, but again was hardly realizable. As the pool of participants with this amount of experience was extremely limited for this work and would have led to an even smaller sample size in the expert group, a somewhat more blurred line, defining expertise as having passed the certification exam for the respective medical specialty, had to be introduced. If profound theoretical knowledge truly defines an expert already is at least debatable. For future work, emphasis has to be put on a clearer distinction between experts and novices, potentially resulting in more clear outcomes.

Third, due to unplanned construction work, the study site had to be altered. During the second study day, the simulation scenario was performed in a more open location, with non-involved personnel passing by the whole day, potentially leading to distractions that were not part of the study plan. If this introduced bias into the results remains unknown. Also, the monitoring equipment had to be positioned at a different location in relation to the simulation mannequin, which could also potentially have influenced participants' visual behavior.

Fourth, data quality, using high-end technology, always causes some kind of headache. With eye tracking data lost at some points for two participants, the available pool of data was reduced even more in an already small sample size. Nonetheless, given the complexity of this work using 3 sensors and 2 questionnaires, a stable amount of data quality was reached.

Fifth, the attempt to collect GSR signals itself was successful, but using the acquired data to gain extra insight on stress and cognitive load by running them through "out-of-the-box" algorithms provided by the software used for biosignal analysis failed. Motions artifacts rendered automatic analysis impossible in the end. Nonetheless, some work has been published where GSR signals were successfully collected during simulation training and subsequently analyzed using machine learning. [18] A future study using the collected data in this trial could therefore aim at creating a similar solution to filter out motion artifacts and allow for a valid analysis of the acquired signals.

Lastly, it must be mentioned that medical simulation has been gaining attention at the study sight during the last few years, with especially younger doctors training a lot with simulation equipment. If this may have led to a certain amount of "reversal" in expertise, with the younger staff more accustomed to this kind of training setting, requires further investigation.

6 Conclusion

This work indicates that a complex multi-sensor analysis approach during simulation scenarios is plannable in a relatively small amount of time, highly feasible, and well accepted by participants. It may even suggest the usage of sensors during real-world medical scenarios, albeit ethical concerns, e.g. if a piece of a piece equipment leads to a delay in the provision of medical aid, remain standing. Additionally, the usage of GSR sensors needs further refinement to produce reliable data, with potential optimization sources lying in sensor position, fixation, and analysis strategies.

Regarding visual behavior, experts and novices showed quite similar results, with differences being more found between the times of interest than between groups. Other reasons possibly explaining this, despite the small sample size of this work, have already been outlined above. Nonetheless, this work brought to light highly interesting insights into visual behavior that also can be found in non-expected similarities, e.g. the amount of time both groups invested in looking at monitoring equipment.

Also, noticeable alterations of eye tracking as well as heart rate metrics due to imposed stress on participants were detected, being in line with former work on this topic, rendering the usage of this approach promising for the future.

Regarding the subjective evaluation of cognitive load, it was shown that subjective experiences and objective sensor data do not necessarily align, supporting an integrated approach, with each data source providing a valuable puzzle piece, aiding in finding the “ground truth” of behavior and cognitive load.

To outline possible upcoming work, a repetition of this study with a higher sample size, a clearer distinction between experts and novices and, – in the case of GSR – more robust sensor fitting, and analysis solutions seem promising.

References

[1] M. Panda and N. A. Desbiens, “An ‘Education for Life’ Requirement to Promote Lifelong Learning in an Internal Medicine Residency Program,” *Journal of Graduate Medical Education*, vol. 2, no. 4, pp. 562–565, Dec. 2010, doi: 10.4300/JGME-D-09-00068.1.

[2] D. Barsuk *et al.*, “Using Advanced Simulation for Recognition and Correction of Gaps in Airway and Breathing Management Skills in Prehospital Trauma Care,” *Anesthesia & Analgesia*, vol. 100, no. 3, p. 803, Mar. 2005, doi: 10.1213/01.ANE.0000143390.11746.CF.

[3] C. C. Schubert, T. K. Denmark, B. Crandall, A. Grome, and J. Pappas, “Characterizing Novice-Expert Differences in Macrocognition: An Exploratory Study of Cognitive Work in the Emergency Department,” *Annals of Emergency Medicine*, vol. 61, no. 1, pp. 96–109, Jan. 2013, doi: 10.1016/j.annemergmed.2012.08.034.

[4] S. Blum, R. Emkes, F. Minow, J. Anlauff, A. Finke, and S. Debener, “Flex-printed forehead EEG sensors (fEEGrid) for long-term EEG acquisition,” *J. Neural Eng.*, vol. 17, no. 3, p. 034003, Jun. 2020, doi: 10.1088/1741-2552/ab914c.

[5] M. Wagner, P. Gröpel, K. Bibl, M. Olischar, M. A. Auerbach, and I. T. Gross, “Eye-tracking during simulation-based neonatal airway management,” *Pediatr Res*, vol. 87, no. 3, Art. no. 3, Feb. 2020, doi: 10.1038/s41390-019-0571-9.

[6] H. Ashraf, M. H. Sodergren, N. Merali, G. Mylonas, H. Singh, and A. Darzi, “Eye-tracking technology in medical education: A systematic review,” *Medical Teacher*, vol. 40, no. 1, pp. 62–69, Jan. 2018, doi: 10.1080/0142159X.2017.1391373.

[7] R. A. Block, P. A. Hancock, and D. Zakay, “How cognitive load affects duration judgments: A meta-analytic review,” *Acta Psychologica*, vol. 134, no. 3, pp. 330–343, Jul. 2010, doi: 10.1016/j.actpsy.2010.03.006.

[8] K. L. Fraser, P. Ayres, and J. Sweller, “Cognitive Load Theory for the Design of Medical Simulations,” *Simulation in Healthcare*, vol. 10, no. 5, p. 295, Oct. 2015, doi: 10.1097/SIH.0000000000000097.

[9] J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory*, 1st ed. New York, NY: Springer, 2011. doi: 10.1007/978-1-4419-8126-4.

[10] “APA Dictionary of Psychology.” Accessed: Dec. 28, 2023. [Online]. Available: https://dictionary.apa.org/galvanic-skin-response

[11] N. Nourbakhsh, Y. Wang, F. Chen, and R. A. Calvo, “Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks,” in *Proceedings of the 24th Australian Computer-Human Interaction Conference*, in OzCHI ’12. New York, NY, USA: Association for Computing Machinery, Nov. 2012, pp. 420–423. doi: 10.1145/2414536.2414602.

[12] N. Nourbakhsh, F. Chen, Y. Wang, and R. A. Calvo, “Detecting Users’ Cognitive Load by Galvanic Skin Response with Affective Interference,” *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 3, p. 12:1-12:20, Sep. 2017, doi: 10.1145/2960413.

[13] "Cambridge Business English Dictionary." Accessed: Dec. 30, 2023. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/eye-tracking
[14] N. J. Maran and R. J. Glavin, "Low- to high-fidelity simulation – a continuum of medical education?," *Medical Education*, vol. 37, no. s1, pp. 22–28, 2003, doi: 10.1046/j.1365-2923.37.s1.9.x.
[15] K. R. Rosen, "The history of medical simulation," *Journal of Critical Care*, vol. 23, no. 2, pp. 157–166, Jun. 2008, doi: 10.1016/j.jcrc.2007.12.004.
[16] "History of CPR," cpr.heart.org. Accessed: Jan. 02, 2024. [Online]. Available: https://cpr.heart.org/en/resources/history-of-cpr
[17] S. Toohey *et al.*, "Comparing the Psychological Effects of Manikin-Based and Augmented Reality–Based Simulation Training: Within-Subjects Crossover Study," *JMIR Med Educ*, vol. 8, no. 3, p. e36447, Aug. 2022, doi: 10.2196/36447.
[18] K. Ross *et al.*, "Toward Dynamically Adaptive Simulation: Multimodal Classification of User Expertise Using Wearable Devices," *Sensors (Basel)*, vol. 19, no. 19, p. 4270, Oct. 2019, doi: 10.3390/s19194270.
[19] "Simulationszentrum Wien | Wiener Gesundheitsverbund," Simulationszentrum. Accessed: Jan. 02, 2024. [Online]. Available: https://simulation.gesundheitsverbund.at/
[20] R. Greif *et al.*, "European Resuscitation Council Guidelines 2021: Education for resuscitation," *Resuscitation*, vol. 161, pp. 388–407, Apr. 2021, doi: 10.1016/j.resuscitation.2021.02.016.
[21] W. C. Mundell, C. C. Kennedy, J. H. Szostek, and D. A. Cook, "Simulation technology for resuscitation training: A systematic review and meta-analysis," *Resuscitation*, vol. 84, no. 9, pp. 1174–1183, Sep. 2013, doi: 10.1016/j.resuscitation.2013.04.016.
[22] B. M. Lo *et al.*, "Comparison of traditional versus high-fidelity simulation in the retention of ACLS knowledge," *Resuscitation*, vol. 82, no. 11, pp. 1440–1443, Nov. 2011, doi: 10.1016/j.resuscitation.2011.06.017.
[23] C. Massoth *et al.*, "High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students," *BMC Medical Education*, vol. 19, no. 1, p. 29, Jan. 2019, doi: 10.1186/s12909-019-1464-7.
[24] K. Couper *et al.*, "An evaluation of three methods of in-hospital cardiac arrest educational debriefing: The cardiopulmonary resuscitation debriefing study," *Resuscitation*, vol. 105, pp. 130–137, Aug. 2016, doi: 10.1016/j.resuscitation.2016.05.005.
[25] D. P. Edelson *et al.*, "Improving In-Hospital Cardiac Arrest Process and Outcomes With Performance Debriefing," *Archives of Internal Medicine*, vol. 168, no. 10, pp. 1063–1069, May 2008, doi: 10.1001/archinte.168.10.1063.
[26] J. Lee *et al.*, "Debriefing methods and learning outcomes in simulation nursing education: A systematic review and meta-analysis," *Nurse Education Today*, vol. 87, p. 104345, Apr. 2020, doi: 10.1016/j.nedt.2020.104345.
[27] K. T. Dreifuerst, "Using Debriefing for Meaningful Learning to Foster Development of Clinical Reasoning in Simulation," *Journal of Nursing Education*, vol. 51, no. 6, pp. 326–333, Jun. 2012, doi: 10.3928/01484834-20120409-02.
[28] F. Ciceron *et al.*, "Individual versus collective debriefing after interprofessional training course simulation: The randomised DEBRIEF-SIM trial," *Anaesthesia Critical Care & Pain Medicine*, vol. 40, no. 2, p. 100828, Apr. 2021, doi: 10.1016/j.accpm.2021.100828.

[29] C. W. Oyster, *The Human Eye: Structure and Function*, 1st ed. Sunderland, MA: Sinauer Associates, 1999.
[30] D. H. Hubel, *Eye, brain, and vision*. New York, NY, US: Scientific American Library/Scientific American Books, 1995, pp. viii, 242.
[31] M. Xiang, H. Zhou, and J. Nathans, “Molecular biology of retinal ganglion cells.,” *Proc Natl Acad Sci USA*, vol. 93, no. 2, pp. 596–601, Jan. 1996.
[32] Thomas F. Freddo and Edward Chaum, *Anatomy of the Eye and Orbit : The Clinical Essentials*, 1st ed. Philadelphia: Wolters Kluwer Health, 2017.
[33] X. R. Gao, H. Huang, and H. Kim, “Genome-wide association analyses identify 139 loci associated with macular thickness in the UK Biobank cohort,” *Human Molecular Genetics*, vol. 28, no. 7, pp. 1162–1172, Apr. 2019, doi: 10.1093/hmg/ddy422.
[34] A. Bringmann *et al.*, “The primate fovea: Structure, function and development,” *Progress in Retinal and Eye Research*, vol. 66, pp. 49–84, Sep. 2018, doi: 10.1016/j.preteyeres.2018.03.006.
[35] M. M. Schira, C. W. Tyler, M. Breakspear, and B. Spehar, “The Foveal Confluence in Human Visual Cortex,” *J. Neurosci.*, vol. 29, no. 28, pp. 9050–9058, Jul. 2009, doi: 10.1523/JNEUROSCI.1760-09.2009.
[36] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer, *Eye Tracking: A comprehensive guide to methods and measures*, 1st ed. OUP Oxford, 2011.
[37] F. Tong, “Primary visual cortex and visual awareness,” *Nat Rev Neurosci*, vol. 4, no. 3, pp. 219–229, Mar. 2003, doi: 10.1038/nrn1055.
[38] M. Schindler and A. J. Lilienthal, “Domain-specific interpretation of eye tracking data: towards a refined use of the eye-mind hypothesis for the field of geometry,” *Educ Stud Math*, vol. 101, no. 1, pp. 123–139, May 2019, doi: 10.1007/s10649-019-9878-z.
[39] J. E. Hoffman and B. Subramaniam, “The role of visual attention in saccadic eye movements,” *Percept Psychophys*, vol. 57, no. 6, pp. 787–795, Aug. 1995, doi: 10.3758/bf03206794.
[40] A. K. Mishra, Y. Aloimonos, L. F. Cheong, and A. Kassim, “Active Visual Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 639–653, Apr. 2012, doi: 10.1109/TPAMI.2011.171.
[41] “Eye tracking products for research and integrations.” Accessed: Apr. 01, 2024. [Online]. Available: https://www.tobii.com/products/
[42] “VPS 19 – die einfachsten SMART GLASSES der Welt,” VIEWPOINTSYSTEM. Accessed: Apr. 07, 2024. [Online]. Available: https://viewpointsystem.com/produkte/
[43] J. Sigut and S.-A. Sidha, “Iris Center Corneal Reflection Method for Gaze Tracking Using Visible Light,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 411–419, Feb. 2011, doi: 10.1109/TBME.2010.2087330.
[44] “Eye Tracking: The Complete Pocket Guide - iMotions.” Accessed: Apr. 01, 2024. [Online]. Available: https://imotions.com/blog/learning/best-practice/eye-tracking/
[45] S. A. Mosquera, S. Verma, and C. McAlinden, “Centration axis in refractive surgery,” *Eye and Vision*, vol. 2, no. 1, p. 4, Feb. 2015, doi: 10.1186/s40662-015-0014-6.
[46] C. Jian-nan, Z. Peng-yi, Z. Si-yi, Z. Chuang, and H. Ying, “Key Techniques of Eye Gaze Tracking Based on Pupil Corneal Reflection,” in *2009 WRI Global*

Congress on Intelligent Systems, Xiamen, China: IEEE, 2009, pp. 133–138. doi: 10.1109/GCIS.2009.338.

[47] K. Rayner, “Eye movements in reading and information processing: 20 years of research,” *Psychological Bulletin*, vol. 124, no. 3, pp. 372–422, 1998, doi: 10.1037/0033-2909.124.3.372.

[48] M. A. Just and P. A. Carpenter, “A theory of reading: From eye fixations to comprehension,” *Psychological Review*, vol. 87, no. 4, pp. 329–354, 1980, doi: 10.1037/0033-295X.87.4.329.

[49] M. K. Eckstein, B. Guerra-Carrillo, A. T. Miller Singley, and S. A. Bunge, “Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?,” *Developmental Cognitive Neuroscience*, vol. 25, pp. 69–91, Jun. 2017, doi: 10.1016/j.dcn.2016.11.001.

[50] A. L. Yarbus, “Eye Movements During Perception of Complex Objects,” in *Eye Movements and Vision*, A. L. Yarbus, Ed., Boston, MA: Springer US, 1967, pp. 171–211. doi: 10.1007/978-1-4899-5379-7_8.

[51] B. W. Tatler, N. J. Wade, H. Kwan, J. M. Findlay, and B. M. Velichkovsky, “Yarbus, eye movements, and vision,” *Iperception*, vol. 1, no. 1, pp. 7–27, Jul. 2010, doi: 10.1068/i0382.

[52] B. M. Velichkovsky, S. M. Dornhoefer, S. Pannasch, and P. J. A. Unema, “Visual fixations and level of attentional processing,” in *Proceedings of the 2000 symposium on Eye tracking research & applications*, in ETRA ’00. New York, NY, USA: Association for Computing Machinery, Nov. 2000, pp. 79–85. doi: 10.1145/355017.355029.

[53] E. R. Grant and M. J. Spivey, “Eye Movements and Problem Solving: Guiding Attention Guides Thought,” *Psychol Sci*, vol. 14, no. 5, pp. 462–466, Sep. 2003, doi: 10.1111/1467-9280.02454.

[54] M.-L. Lai *et al.*, “A review of using eye-tracking technology in exploring learning from 2000 to 2012,” *Educational Research Review*, vol. 10, pp. 90–115, Dec. 2013, doi: 10.1016/j.edurev.2013.10.001.

[55] B. B. de Koning, H. K. Tabbers, R. M. J. P. Rikers, and F. Paas, “Attention guidance in learning from a complex animation: Seeing is understanding?,” *Learning and Instruction*, vol. 20, no. 2, pp. 111–122, Apr. 2010, doi: 10.1016/j.learninstruc.2009.02.010.

[56] Y. Tomizawa, H. Aoki, S. Suzuki, T. Matayoshi, and R. Yozu, “Eye-tracking analysis of skilled performance in clinical extracorporeal circulation,” *J Artif Organs*, vol. 15, no. 2, pp. 146–157, Jun. 2012, doi: 10.1007/s10047-012-0630-z.

[57] E. Capogna, F. Salvi, L. Delvino, A. Di Giacinto, and M. Velardo, “Novice and Expert Anesthesiologists’ Eye-Tracking Metrics During Simulated Epidural Block: A Preliminary, Brief Observational Report,” *Local Reg Anesth*, vol. 13, pp. 105–109, Aug. 2020, doi: 10.2147/LRA.S267879.

[58] N. Charness, E. M. Reingold, M. Pomplun, and D. M. Stampe, “The perceptual aspect of skilled performance in chess: Evidence from eye movements,” *Memory & Cognition*, vol. 29, no. 8, pp. 1146–1152, Dec. 2001, doi: 10.3758/BF03206384.

[59] H. Haider and P. A. Frensch, “Eye movement during skill acquisition: More evidence for the information-reduction hypothesis,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 25, no. 1, pp. 172–190, 1999, doi: 10.1037/0278-7393.25.1.172.

[60] E. A. Henneman *et al.*, “Eye Tracking as a Debriefing Mechanism in the Simulated Setting Improves Patient Safety Practices,” *Dimensions of Critical*

Care Nursing, vol. 33, no. 3, p. 129, Jun. 2014, doi: 10.1097/DCC.0000000000000041.
[61] J. S. Grant, J. Moss, C. Epps, and P. Watts, "Using Video-Facilitated Feedback to Improve Student Performance Following High-Fidelity Simulation," *Clinical Simulation in Nursing*, vol. 6, no. 5, pp. e177–e184, Sep. 2010, doi: 10.1016/j.ecns.2009.09.001.
[62] P. O'Meara *et al.*, "Developing situation awareness amongst nursing and paramedicine students utilizing eye tracking technology and video debriefing techniques: A proof of concept paper," *International Emergency Nursing*, vol. 23, no. 2, pp. 94–99, Apr. 2015, doi: 10.1016/j.ienj.2014.11.001.
[63] E. A. Henneman, J. L. Marquard, D. L. Fisher, and A. Gawlinski, "Eye Tracking: A Novel Approach for Evaluating and Improving the Safety of Healthcare Processes in the Simulated Setting," *Simulation in Healthcare*, vol. 12, no. 1, p. 51, Feb. 2017, doi: 10.1097/SIH.0000000000000192.
[64] M. Browning *et al.*, "The use and limits of eye-tracking in high-fidelity clinical scenarios: A pilot study," *International Emergency Nursing*, vol. 25, pp. 43–47, Mar. 2016, doi: 10.1016/j.ienj.2015.08.002.
[65] I. Tanoubi *et al.*, "Comparing the Visual Perception According to the Performance Using the Eye-Tracking Technology in High-Fidelity Simulation Settings," *Behavioral Sciences*, vol. 11, no. 3, Art. no. 3, Mar. 2021, doi: 10.3390/bs11030031.
[66] A. M. Persky and J. D. Robinson, "Moving from Novice to Expertise and Its Implications for Instruction," *Am J Pharm Educ*, vol. 81, no. 9, p. 6065, Nov. 2017, doi: 10.5688/ajpe6065.
[67] G. A. Klein and R. R. Hoffman, "Seeing the invisible: Perceptual-cognitive aspects of expertise," in *Cognitive science foundations of instruction*, Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, 1993, pp. 203–226.
[68] A. M. Colman, "deductive reasoning," in *A Dictionary of Psychology*, 3rd ed., Oxford University Press, 2009, p. 109. Accessed: Jan. 04, 2024. [Online]. Available: https://www.oxfordreference.com/display/10.1093/acref/9780199534067.001.0001/acref-9780199534067-e-2110
[69] G. Klein, K. G. Ross, B. M. Moon, D. E. Klein, R. R. Hoffman, and E. Hollnagel, "Macrocognition," *IEEE Intelligent Systems*, vol. 18, no. 3, pp. 81–85, May 2003, doi: 10.1109/MIS.2003.1200735.
[70] M. L. Gamborg, M. Mehlsen, C. Paltved, S. S. Vetter, and P. Musaeus, "Clinical decision-making and adaptive expertise in residency: a think-aloud study," *BMC Med Educ*, vol. 23, p. 22, Jan. 2023, doi: 10.1186/s12909-022-03990-8.
[71] M. Congdon *et al.*, "Diagnostic Reasoning of Resident Physicians in the Age of Clinical Pathways," *J Grad Med Educ*, vol. 14, no. 4, pp. 466–474, Aug. 2022, doi: 10.4300/JGME-D-21-01032.1.
[72] G. Kovacs and P. Croskerry, "Clinical Decision Making: An Emergency Medicine Perspective," *Academic Emergency Medicine*, vol. 6, no. 9, pp. 947–952, 1999, doi: 10.1111/j.1553-2712.1999.tb01246.x.
[73] P. Croskerry, D. A. Petrie, J. B. Reilly, and G. Tait, "Deciding About Fast and Slow Decisions," *Academic Medicine*, vol. 89, no. 2, p. 197, Feb. 2014, doi: 10.1097/ACM.0000000000000121.
[74] D. Kahneman, *Thinking, fast and slow*. in Thinking, fast and slow. New York, NY, US: Farrar, Straus and Giroux, 2011, p. 499.

[75] O. Ten Cate, E. J. F. M. Custers, and S. J. Durning, Eds., “After Medical Problem Solving (1978): A Role Left for Teaching Clinical Reasoning?,” in *Principles and Practice of Case-based Clinical Reasoning Education*, vol. 15, in Innovation and Change in Professional Education, vol. 15. , Cham, SUI: Springer International Publishing, 2018, pp. 28–29. doi: 10.1007/978-3-319-64828-6.
[76] R. Lipshitz, G. Klein, J. Orasanu, and E. Salas, “Taking stock of naturalistic decision making,” *Journal of Behavioral Decision Making*, vol. 14, no. 5, pp. 331–352, 2001, doi: 10.1002/bdm.381.
[77] P. Meso, M. D. Troutt, and J. Rudnicka, “A review of naturalistic decision making research with some implications for knowledge management,” *Journal of Knowledge Management*, vol. 6, no. 1, pp. 63–73, Jan. 2002, doi: 10.1108/13673270210417709.
[78] N. H. Al-Azri, “How to think like an emergency care provider: a conceptual mental model for decision making in emergency care,” *Int J Emerg Med*, vol. 13, no. 1, p. 17, Apr. 2020, doi: 10.1186/s12245-020-00274-0.
[79] M. R. White *et al.*, “Getting Inside the Expert’s Head: An Analysis of Physician Cognitive Processes During Trauma Resuscitations,” *Annals of Emergency Medicine*, vol. 72, no. 3, pp. 289–298, Sep. 2018, doi: 10.1016/j.annemergmed.2018.03.005.
[80] F. Paas, T. van Gog, and J. Sweller, “Cognitive Load Theory: New Conceptualizations, Specifications, and Integrated Research Perspectives,” *Educ Psychol Rev*, vol. 22, no. 2, pp. 115–121, Jun. 2010, doi: 10.1007/s10648-010-9133-8.
[81] G. A. Miller, “The magical number seven, plus or minus two: Some limits on our capacity for processing information,” *Psychological Review*, vol. 63, no. 2, pp. 81–97, 1956, doi: 10.1037/h0043158.
[82] I. Pollack, “The Information of Elementary Auditory Displays,” *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 745–749, Nov. 1952, doi: 10.1121/1.1906969.
[83] D. C. Geary, “Principles of evolutionary educational psychology,” *Learning and Individual Differences*, vol. 12, no. 4, pp. 317–345, 2000, doi: 10.1016/S1041-6080(02)00046-8.
[84] J. Sweller, “Cognitive load theory and educational technology,” *Education Tech Research Dev*, vol. 68, no. 1, pp. 1–16, Feb. 2020, doi: 10.1007/s11423-019-09701-3.
[85] J. Sweller, J. J. G. van Merrienboer, and F. G. W. C. Paas, “Cognitive Architecture and Instructional Design,” *Educational Psychology Review*, vol. 10, no. 3, pp. 251–296, Sep. 1998, doi: 10.1023/A:1022193728205.
[86] J. Sweller and P. Chandler, “Evidence for Cognitive Load Theory,” *Cognition and Instruction*, vol. 8, no. 4, pp. 351–362, Dec. 1991, doi: 10.1207/s1532690xci0804_5.
[87] F. G. W. C. Paas and J. J. G. Van Merriënboer, “Instructional control of cognitive load in the training of complex cognitive tasks,” *Educ Psychol Rev*, vol. 6, no. 4, pp. 351–371, Dec. 1994, doi: 10.1007/BF02213420.
[88] J. Sweller, “Cognitive load theory, learning difficulty, and instructional design,” *Learning and Instruction*, vol. 4, no. 4, pp. 295–312, Jan. 1994, doi: 10.1016/0959-4752(94)90003-5.
[89] S. Solhjoo *et al.*, “Heart Rate and Heart Rate Variability Correlate with Clinical Reasoning Performance and Self-Reported Measures of Cognitive Load,” *Sci Rep*, vol. 9, no. 1, p. 14668, Oct. 2019, doi: 10.1038/s41598-019-50280-3.

[90] M. Klepsch, F. Schmitz, and T. Seufert, “Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load,” *Front Psychol*, vol. 8, p. 1997, Nov. 2017, doi: 10.3389/fpsyg.2017.01997.

[91] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, “Cognitive Load Measurement as a Means to Advance Cognitive Load Theory,” *Educational Psychologist*, vol. 38, no. 1, pp. 63–71, Jan. 2003, doi: 10.1207/S15326985EP3801_8.

[92] K. Ouwehand, A. van der Kroef, J. Wong, and F. Paas, “Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales?,” *Front. Educ.*, vol. 6, Sep. 2021, doi: 10.3389/feduc.2021.702616.

[93] F. G. W. C. Paas, “Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach,” *Journal of Educational Psychology*, vol. 84, no. 4, pp. 429–434, 1992, doi: 10.1037/0022-0663.84.4.429.

[94] O. Bratfisch, G. Borg, and S. Dornič, “Perceived Item-Difficulty in Three Tests of Intellectual Performance Capacity,” Institute of Applied Psychology, Stockholm, Sweden, 29, 1972. Accessed: Mar. 13, 2024. [Online]. Available: https://eric.ed.gov/?id=ED080552

[95] K. E. DeLeeuw and R. E. Mayer, “A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load,” *Journal of Educational Psychology*, vol. 100, no. 1, pp. 223–234, 2008, doi: 10.1037/0022-0663.100.1.223.

[96] S. G. Hart, “Nasa-Task Load Index (NASA-TLX); 20 Years Later,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 9, pp. 904–908, Oct. 2006, doi: 10.1177/154193120605000909.

[97] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research,” in *Advances in Psychology*, vol. 52, P. A. Hancock and N. Meshkati, Eds., in Human Mental Workload, vol. 52. , North-Holland, 1988, pp. 139–183. doi: 10.1016/S0166-4115(08)62386-9.

[98] P. Hoonakker *et al.*, “Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX),” *IIE Transactions on Healthcare Systems Engineering*, vol. 1, no. 2, pp. 131–143, Apr. 2011, doi: 10.1080/19488300.2011.609524.

[99] A. G. Vredenburgh, M. B. Weinger, K. J. Williams, M. J. Kalsher, and A. Macario, “Developing a Technique to Measure Anesthesiologists’ Real-Time Workload,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 44, no. 26, pp. 241–244, Jul. 2000, doi: 10.1177/154193120004402627.

[100] R. W. Backs, A. M. Ryan, and G. F. Wilson, “Psychophysiological Measures of Workload during Continuous Manual Performance,” *Hum Factors*, vol. 36, no. 3, pp. 514–531, Sep. 1994, doi: 10.1177/001872089403600308.

[101] S. Rubio-Valdehita, M. I. López-Núñez, R. López-Higes, and E. M. Díaz-Ramiro, “Development of the CarMen-Q Questionnaire for mental workload assessment,” *Psicothema*, vol. 29, no. 4, pp. 570–576, Nov. 2017, doi: 10.7334/psicothema2017.151.

[102] J. C. F. de Winter, “Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective,” *Cogn Tech Work*, vol. 16, no. 3, pp. 289–297, Aug. 2014, doi: 10.1007/s10111-014-0275-1.

[103] R. D. McKendrick and E. Cherry, “A Deeper Look at the NASA TLX and Where It Falls Short,” *Proceedings of the Human Factors and Ergonomics*

Society Annual Meeting, vol. 62, no. 1, pp. 44–48, Sep. 2018, doi: 10.1177/1541931218621010.
[104]R. K. Merton, “The Matthew effect in science. The reward and communication systems of science are considered,” *Science*, vol. 159, no. 3810, pp. 56–63, Jan. 1968.
[105]F. G. W. C. Paas and J. J. G. Van Merriënboer, “Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach,” *Journal of Educational Psychology*, vol. 86, no. 1, pp. 122–133, 1994, doi: 10.1037/0022-0663.86.1.122.
[106]P. W. M. Van Gerven, F. Paas, J. J. G. Van Merriënboer, and H. G. Schmidt, “Memory load and the cognitive pupillary response in aging,” *Psychophysiology*, vol. 41, no. 2, pp. 167–174, 2004, doi: 10.1111/j.1469-8986.2003.00148.x.
[107]R. R. Whelan, “Neuroimaging of cognitive load in instructional multimedia,” *Educational Research Review*, vol. 2, no. 1, pp. 1–12, Jan. 2007, doi: 10.1016/j.edurev.2006.11.001.
[108]J. Kanitakis, “Anatomy, histology and immunohistochemistry of normal human skin,” *Eur J Dermatol*, vol. 12, no. 4, pp. 390-399;, 2002.
[109]J. T. Cacioppo, L. G. Tassinary, and G. Berntson, *Handbook of Psychophysiology*, 3rd ed. Cambridge University Press, 2007. [Online]. Available: https://books.google.at/books?id=E7hRKwVBXb4C
[110]L. B. Baker and A. S. Wolfe, “Physiological mechanisms determining eccrine sweat composition,” *Eur J Appl Physiol*, vol. 120, no. 4, pp. 719–752, Apr. 2020, doi: 10.1007/s00421-020-04323-7.
[111]H. F. Posada-Quintero and K. H. Chon, “Innovations in Electrodermal Activity Data Collection and Signal Processing: A Systematic Review,” *Sensors*, vol. 20, no. 2, Art. no. 2, 2020, doi: 10.3390/s20020479.
[112]W. Boucsein *et al.*, “Publication recommendations for electrodermal measurements,” *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, Aug. 2012, doi: 10.1111/j.1469-8986.2012.01384.x.
[113]M. S. Gupta, “Georg Simon Ohm and Ohm’s Law,” *IEEE Transactions on Education*, vol. 23, no. 3, pp. 156–162, Aug. 1980, doi: 10.1109/TE.1980.4321401.
[114]R. M. Stern, W. J. Ray, and K. S. Quigley, *Psychophysiological Recording*, 2nd ed. Oxford University Press, 2000. doi: 10.1093/acprof:oso/9780195113594.001.0001.
[115]G. Billman, “Heart Rate Variability – A Historical Perspective,” *Frontiers in Physiology*, vol. 2, 2011, doi: 10.3389/fphys.2011.00086.
[116]F. Shaffer, R. McCraty, and C. L. Zerr, “A healthy heart is not a metronome: an integrative review of the heart’s anatomy and heart rate variability,” *Front. Psychol.*, vol. 5, Sep. 2014, doi: 10.3389/fpsyg.2014.01040.
[117]F. Shaffer and J. Ginsberg, “An Overview of Heart Rate Variability Metrics and Norms,” *Front. Public Health*, vol. 5, 2017, doi: 10.3389/fpubh.2017.00258.
[118]J. E. Peabody *et al.*, “A Systematic Review of Heart Rate Variability as a Measure of Stress in Medical Professionals,” *Cureus*, vol. 15, no. 1, Jan. 2023, doi: 10.7759/cureus.34345.
[119]F. G. W. C. Paas, J. J. G. van Merriënboer, and J. J. Adam, “Measurement of Cognitive Load in Instructional Research,” *Percept Mot Skills*, vol. 79, no. 1, pp. 419–430, Aug. 1994, doi: 10.2466/pms.1994.79.1.419.

[120] A. M. Hughes, G. M. Hancock, S. L. Marlow, K. Stowers, and E. Salas, "Cardiac Measures of Cognitive Workload: A Meta-Analysis," *Hum Factors*, vol. 61, no. 3, pp. 393–414, May 2019, doi: 10.1177/0018720819830553.
[121] M. Keskin, K. Ooms, A. O. Dogru, and P. De Maeyer, "Exploring the Cognitive Load of Expert and Novice Map Users Using EEG and Eye Tracking," *ISPRS International Journal of Geo-Information*, vol. 9, no. 7, Art. no. 7, Jul. 2020, doi: 10.3390/ijgi9070429.
[122] J. Zagermann, U. Pfeil, and H. Reiterer, "Measuring Cognitive Load using Eye Tracking Technology in Visual Computing," in *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, in BELIV '16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 78–85. doi: 10.1145/2993901.2993908.
[123] S. Chen, J. Epps, N. Ruiz, and F. Chen, "Eye activity as a measure of human mental effort in HCI," in *Proceedings of the 16th international conference on Intelligent user interfaces*, in IUI '11. New York, NY, USA: Association for Computing Machinery, Feb. 2011, pp. 315–318. doi: 10.1145/1943403.1943454.
[124] M. Behroozi, A. Lui, I. Moore, D. Ford, and C. Parnin, "Dazed: measuring the cognitive load of solving technical interview problems at the whiteboard," in *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, in ICSE-NIER '18. New York, NY, USA: Association for Computing Machinery, May 2018, pp. 93–96. doi: 10.1145/3183399.3183415.
[125] J. J. G. Van Merriënboer and J. Sweller, "Cognitive load theory in health professional education: design principles and strategies," *Medical Education*, vol. 44, no. 1, pp. 85–93, 2010, doi: 10.1111/j.1365-2923.2009.03498.x.
[126] L. M. Naismith, J. J. H. Cheung, C. Ringsted, and R. B. Cavalcanti, "Limitations of subjective cognitive load measures in simulation-based procedural training," *Medical Education*, vol. 49, no. 8, pp. 805–814, 2015, doi: 10.1111/medu.12732.
[127] L. M. Naismith and R. B. Cavalcanti, "Validity of Cognitive Load Measures in Simulation-Based Training: A Systematic Review," *Acad Med*, vol. 90, no. 11 Suppl, pp. S24-35, Nov. 2015, doi: 10.1097/ACM.0000000000000893.
[128] J. Y. Lee, N. de Jong, J. Donkers, H. Jarodzka, and J. J. G. van Merriënboer, "Measuring Cognitive Load in Virtual Reality Training via Pupillometry," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 704–710, 2024, doi: 10.1109/TLT.2023.3326473.
[129] J. G. Frederiksen *et al.*, "Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: a randomized trial," *Surg Endosc*, vol. 34, no. 3, pp. 1244–1252, Mar. 2020, doi: 10.1007/s00464-019-06887-8.
[130] N. McInerney, D. Nally, M. F. Khan, H. Heneghan, and R. A. Cahill, "Performance effects of simulation training for medical students – a systematic review," *GMS J Med Educ*, vol. 39, no. 5, p. Doc51, Nov. 2022, doi: 10.3205/zma001572.
[131] E. von Elm *et al.*, "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies," *Ann Intern Med*, vol. 147, no. 8, pp. 573–577, Oct. 2007, doi: 10.7326/0003-4819-147-8-200710160-00010.
[132] R. A. Byrne *et al.*, "2023 ESC Guidelines for the management of acute coronary syndromes: Developed by the task force on the management of

acute coronary syndromes of the European Society of Cardiology (ESC)," *European Heart Journal*, vol. 44, no. 38, pp. 3720–3826, Oct. 2023, doi: 10.1093/eurheartj/ehad191.

[133] C. Vahdatpour, D. Collins, and S. Goldberg, "Cardiogenic Shock," *Journal of the American Heart Association*, vol. 8, no. 8, Apr. 2019, doi: 10.1161/JAHA.119.011991.

[134] K. Thygesen *et al.*, "Fourth Universal Definition of Myocardial Infarction (2018)," *Circulation*, vol. 138, no. 20, pp. e618–e651, Nov. 2018, doi: 10.1161/CIR.0000000000000617.

[135] F. Sousa-Nunes, A. Dias, J. Ribeiro, and R. Fontes-Carvalho, "A Pleasant Surprise in the Face of ST-Segment Elevation," *JACC: Case Reports*, vol. 3, no. 11, pp. 1384–1386, Sep. 2021, doi: 10.1016/j.jaccas.2021.06.035.

[136] G. D. Perkins *et al.*, "European Resuscitation Council Guidelines 2021: Executive summary," *Resuscitation*, vol. 161, pp. 1–60, Apr. 2021, doi: 10.1016/j.resuscitation.2021.02.003.

[137] C. Hafner *et al.*, "Eye-tracking during simulation-based echocardiography: a feasibility study," *BMC Medical Education*, vol. 23, no. 1, p. 490, Jul. 2023, doi: 10.1186/s12909-023-04458-z.

[138] R. Phitayakorn, R. D. Minehart, M. C. M. Pian-Smith, M. W. Hemingway, and E. R. Petrusa, "Practicality of using galvanic skin response to measure intraoperative physiologic autonomic activation in operating room team members," *Surgery*, vol. 158, no. 5, pp. 1415–1420, Nov. 2015, doi: 10.1016/j.surg.2015.04.009.

[139] "GSR+ User Guide Revision 1.13." Accessed: Mar. 02, 2024. [Online]. Available: https://shimmersensing.com/wp-content/docs/support/documentation/GSR_User_Guide_rev1.13.pdf

[140] M. van Dooren, J. J. G. (Gert-J. de Vries, and J. H. Janssen, "Emotional sweating across the body: Comparing 16 different skin conductance measurement locations," *Physiology & Behavior*, vol. 106, no. 2, pp. 298–304, May 2012, doi: 10.1016/j.physbeh.2012.01.020.

[141] "Mann-Whitney-U-Test." Accessed: Aug. 22, 2024. [Online]. Available: https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/unterschiede/zentral/mann.html

[142] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses," *Behavior Research Methods*, vol. 41, no. 4, pp. 1149–1160, Nov. 2009, doi: 10.3758/BRM.41.4.1149.

[143] G. E. Noether, "Sample Size Determination for Some Common Nonparametric Tests," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 645–647, Jun. 1987, doi: 10.1080/01621459.1987.10478478.

[144] R. J. Grissom, "Probability of the superior outcome of one treatment over another," *Journal of Applied Psychology*, vol. 79, no. 2, pp. 314–316, 1994, doi: 10.1037/0021-9010.79.2.314.

[145] R. Ristl, "Sample size calculator." Accessed: Sep. 02, 2024. [Online]. Available: https://homepage.univie.ac.at/robin.ristl/samplesize.php?test=wilcoxon

[146] R. Gajda, J. Gajda, M. Czuba, B. Knechtle, and W. Drygas, "Sports Heart Monitors as Reliable Diagnostic Tools for Training Control and Detecting Arrhythmias in Professional and Leisure-Time Endurance Athletes: An Expert

Consensus Statement,” *Sports Med*, vol. 54, no. 1, pp. 1–21, 2024, doi: 10.1007/s40279-023-01948-4.

[147] J. Liechty, R. Pieters, and M. Wedel, “Global and local covert visual attention: Evidence from a bayesian hidden markov model,” *Psychometrika*, vol. 68, no. 4, pp. 519–541, Dec. 2003, doi: 10.1007/BF02295608.

[148] M. Wedel, R. Pieters, and R. van der Lans, “Modeling Eye Movements During Decision Making: A Review,” *Psychometrika*, vol. 88, no. 2, pp. 697–729, 2023, doi: 10.1007/s11336-022-09876-4.

[149] J. A. Seideman, T. R. Stanford, and E. Salinas, “Saccade metrics reflect decision-making dynamics during urgent choices,” *Nat Commun*, vol. 9, no. 1, p. 2907, Jul. 2018, doi: 10.1038/s41467-018-05319-w.

Abbreviations

AOI: Area of interest

CCU: Cardiac care unit

CLT: Cognitive load theory

CPR: Cardiopulmonary resuscitation

DML: Debriefing for meaningful learning

ECG: Electrocardiogram

ERC: European resuscitation council

FPGA: Field programmable gate array

GDPR: General data protection regulation

GSR: Galvanic skin response

HR: Heart rate

HRV: Heart rate variability

ILCOR: International liaison committee on resuscitation

IR: Infrared

MI: Myocardial infarction

RMSSD: Root mean square of successive differences

SCL: Skin conductance level

SCR: Skin conductance response

SCTP: Stream control transmission protocol

STROBE: Strengthening the reporting of observational studies in epidemiology

TOI: Time of interest

UDP: User datagram protocol

USB: Universal serial bus

VF: Ventricular fibrillation

VPS: Viewpointsystem

VR: Virtual reality

List of Figures

List of Tables

Appendix

Supplementary tables

Supplementary Table 1 - Biosensor results for overall scenario

	Novice	Expert	Overall
Fixations			
Fixation count (n, mean (SD))	389 (± 88)	384 (± 80)	387 (± 79)
Fixation time abs. (min, mean (SD))	1.9 (± 0.4)	2.7 (± 1.1)	2.3 (± 0.9)
Fixation time rel. (% of TOI, mean (SD))	13 (± 4)¥	20 (± 8)	16 (± 7)*
Fixation rate (n/min, mean (SD))	28 (± 6)	28 (± 6)	28 (± 6)
Mean fixation duration (ms, mean(SD))	295 (± 30)	419 (± 133)	357 (± 112)
TTFF (sec, mean (SD))	37 (± 27)¥	41 (± 37)	39 (± 30)*
Saccades			
Saccade count (n, mean (SD))	619 (± 68)	527 (± 150)	573 (± 120)
Saccade rate (n/min, mean (SD))	45 (± 5)	38 (± 11)	42 (± 9)
Saccade amplitude (deg, mean (SD))	5.3 (± 1.1)	4.1 (± 0.4)	4.7 (± 1.0)
Saccade peak velocity (deg/sec, mean (SD))	113.7 (± 28.1)	98.7 (± 20.1)	106.2 (± 24.3)
Blinks			
Blink count (n, mean (SD))	327 (± 136) †	331 (± 159)	329 (± 139) ∆
Blink rate (n/min, mean (SD))	25 (± 11) †	24 (± 11)	24 (± 11) ∆
Heart Rate			
HR (bpm, mean (SD))	97 (± 13)	100 (± 11)	98 (± 12)
RMSSD (ms, mean (SD))	29 (± 12)	21 (± 7)	26 (± 11)

Legend: ¥: n = 7, *: n = 12, †: n = 6, ∆: n = 11. If not otherwise marked n for novice = 5. Inclusion of participants depended on data availability.

Supplementary Table 2 - Biosensor results for first contact TOI

	Novice	**Expert**	**Overall**
Fixations			
Fixation count (n, mean (SD))	147 (± 40)	147 (± 53)	147 (± 44)
Fixation time abs. (min, mean (SD))	0.7 (± 0.2)	1.0 (± 0.4)	0.8 (± 0.3)
Fixation time rel. (% of TOI, mean (SD))	15 (± 5)	22 (± 9)	18 (± 7)
Fixation rate (n/min, mean (SD))	33 (± 9)	33 (± 12)	33 (± 10)
Mean fixation duration (ms, mean(SD))	276 (± 53)	404 (± 90)	329 (± 94)
TTFF (sec, mean (SD))	37 (± 27)	41 (± 37)	39 (± 30)
Saccades			
Saccade count (n, mean (SD))	225 (± 49)	204 (± 101)	216 (± 72)
Saccade rate (n/min, mean (SD))	50 (± 11)	45 (± 23)	48 (±16)
Saccade amplitude (deg, mean (SD))	5.2 (± 1.0)	3.8 (± 0.9)	4.6 (± 1.2)
Saccade peak velocity (deg/sec, mean (SD))	123.1 (± 43)	95.3 (± 11)	111.5 (± 35.4)
Blinks			
Blink count (n, mean (SD))	128 (± 58)	132 (± 79)	130 (± 64)
Blink rate (n/min, mean (SD))	29 (± 13)	29 (± 18)	29 (± 14)
Heart Rate			
HR (bpm, mean (SD))	97 (± 12)	98 (± 11)	97 (± 11)
RMSSD (ms, mean (SD))	32 (± 12)	23 (± 6)	29 (± 11)

Supplementary Table 3 - Biosensor results for cardiac arrest TOI

	Novice	**Expert**	**Overall**
Fixations			
Fixation count (n, mean (SD))	46 (± 15)	34 (± 15)	41 (± 16)
Fixation time abs. (sec, mean (SD))	14 (± 5)	15 (± 9)	15 (± 7)
Fixation time rel. (% of TOI, mean (SD))	23 (± 9)	26 (± 15)	24 (± 11)
Fixation rate (n/min, mean (SD))	46 (± 15)	34 (± 15)	41 (± 16)
Mean fixation duration (ms, mean(SD))	306 (± 85)	434 (± 141)	359 (± 125)
TTFF (sec, mean (SD))	1.2 (± 2.8)	1.4 (± 1.8)	1.3 (± 2.4)
Saccades			
Saccade count (n, mean (SD))	73 (± 23)	50 (± 30)	63 (± 28)
Saccade rate (n/min, mean (SD))	73 (± 23)	50 (± 30)	63 (± 28)
Saccade amplitude (deg, mean (SD))	6.0 (± 1.7)	3.9 (± 2.0)	5.1 (± 2.1)
Saccade peak velocity (deg/sec, mean (SD))	130.3 (± 58.5)	94.0 (± 17.1)	115.2 (± 48.2)
Blinks			
Blink count (n, mean (SD))	23 (± 10)	21 (± 11)	22 (± 10)
Blink rate (n/min, mean (SD))	23 (± 10)	21 (± 11)	22 (± 10)
Heart Rate			
HR (bpm, mean (SD))	100 (± 16)	101 (± 15)	100 (± 15)
RMSSD (ms, mean (SD))	22 (± 9)	18 (± 8)	20 (± 9)

Supplementary Table 4 - Biosensor results for intubation TOI

	Novice	Expert	Overall
Fixations			
Fixation count (n, mean (SD))	41 (± 19)	23 (± 2)	33 (± 17)
Fixation time abs. (sec, mean (SD))	13 (± 5)	9 (± 6)	11 (± 6)
Fixation time rel. (% of TOI, mean (SD))	54 (± 22)	43 (± 29)	50 (± 25)
Fixation rate (n/min, mean (SD))	93 (± 20)	69 (±31)	83 (± 27)
Mean fixation duration (ms, mean(SD))	360 (± 151)	385 (± 276)	371 (± 201)
TTFF (sec, mean (SD))	na	na	na
Saccades			
Saccade count (n, mean (SD))	89 (± 66)	68 (± 45)	80 (± 57)
Saccade rate (n/min, mean (SD))	193 (± 91)	198 (± 111)	195 (± 95)
Saccade amplitude (deg, mean (SD))	6.6 (± 2.8)	4.9 (± 0.7)	5.9 (± 2.3)
Saccade peak velocity (deg/sec, mean (SD))	114.7 (± 56.8)	105.3 (± 26.1)	110.8 (± 45.1)
Blinks			
Blink count (n, mean (SD))	3 (± 4)	2 (± 2)	2 (± 3)
Blink rate (n/min, mean (SD))	6 (± 7)	6 (± 6)	6 (± 6)
Heart Rate			
HR (bpm, mean (SD))	101 (± 14)	108 (± 15)	104 (± 14)
RMSSD (ms, mean (SD))	19 (± 19)	10 (± 6)	15 (± 15)

Supplementary Table 5 - Biosensor results for distraction TOI

	Novice	**Expert**	**Overall**
Fixations			
Fixation count (n, mean (SD))	22 (± 8)	15 (± 6)	19 (± 8)
Fixation time abs. (sec, mean (SD))	4 (± 1)	4 (± 2)	4 (± 2)
Fixation time rel. (% of TOI, mean (SD))	42 (± 14)	39 (± 20)	41 (± 16)
Fixation rate (n/min, mean (SD))	134 (± 47)	91 (± 36)	116 (± 47)
Mean fixation duration (ms, mean (SD))	194 (± 59)	241 (± 88)	213 (± 73)
TTFF (sec, mean (SD))	na	na	na
Saccades			
Saccade count (n, mean (SD))	40 (± 9)	37 (± 7)	39 (± 8),
Saccade rate (n/min, mean (SD))	242 (± 54)	221 (± 40)	233 (± 48)
Saccade amplitude (deg, mean (SD))	7.2 (± 1.3)	9.0 (± 1.4)	7.9 (± 1.6)
Saccade peak velocity (deg/sec, mean (SD))	157.0 (± 33.4)	177.5 (± 57.6)	165.5 (± 43.9)
Blinks			
Blink count (n, mean (SD))	4 (± 2)	3 (± 2)	4 (± 2)
Blink rate (n/min, mean (SD))	25 (± 14)	18 (± 10)	22 (±12)
Heart Rate			
HR (bpm, mean (SD))	95 (± 17)	102 (± 9)	98 (± 14)
RMSSD (ms, mean (SD))	28 (± 21)	15 (± 6)	23 (± 17)

Questionnaires

Register Questionnaire

LimeSurvey

Hello. Please answer this short survey to get you registered.

By participating in this study, you are actively participating in research regarding human visual behavior and stress levels during simulation training, which may lead to improved training for medical personel and patient safety.

Your data will be stored securely and anonymously. If you have any questions, contact Christoph at all times.

Section A: Registration

A1. **Please enter your ID number provided at the beginning**

A2. **Do you suffer from any of these conditions:**

	Yes	No
Glasses or contact lenses required for vision correction	☐	☐
Near- or farsightedness with a correction strength of more than +/- 4 diopters	☐	☐
Astigmatism (Hornhautverkrümmung)	☐	☐
Non-correctable strabism (Schielen)	☐	☐
Amblyopia (Schwachsichtigkeit) of any cause	☐	☐
Cataract	☐	☐
Skin condition prohibiting you from wearing gel covered skin electrodes	☐	☐

A3. **I feel comfortable enough in the English language (written and spoken) to complete the questionnaires associated with this study.**

Yes ☐

No ☐

A4. Please choose your specialty:

- Resident Internal Medicine - subspeciality intensive care ☐
- Resident Internal Medicine ☐
- Resident Anesthesiology ☐
- Senior physician - Anesthesiology ☐
- Senior physician - Internal medicine ☐
- General practitioner ☐
- General practitioner in training ☐

A5. Do you have more than 1 year of experience in intensive care medicine?

- Yes ☐
- No ☐

A6. Have you already passed your final exam?

- Yes ☐
- No ☐

A7. Have you already completed a rotation on an intensive care unit?

- Yes ☐
- No ☐

A8. Please answer the following. I...

	Yes	No
have a valid emergency medicine diploma (Notarztdiplom)	☐	☐
have participated in at least 3 simulation trainings including intubation exercises	☐	☐
have completed a rotation at an anesthesiological department	☐	☐

A9. Do you regularly work as an emergency physician (Notarzt) - at least once a month in the last 12 months?

- Yes ☐
- No ☐

A10. Please choose your gender

- Female ☐
- Male ☐
- Non-binary ☐
- No answer ☐

LimeSurvey

A11. **How old are you? (years)**

A12. **Please choose from 1 (not confident) to 5 (very confident) how confident you feel in handling internal medicine emergency situations**

	1 (not confident)	2	3	4	5 (very confident)
I feel:	☐	☐	☐	☐	☐

A13. **For how long have you been working as a senior physician (Oberarzt)?**

A14. **How many years are you into your residency/training for general practitioner?**

A15. **For how many years have you been working as a G.P.?**

Thank you! We will now check your data and let you know if you are eligible

LimeSurvey

Thanks for participating in our study. For debriefing, please answer the following questions regarding the equipment and your perceived emotions:

Section A: Eye tracking

These questions will help us evaluate how feasible the usage of eye-tracking glasses during simulation is.

A1. Please enter your ID number provided at the beginning

A2. The eye-tracking equipment...

Please let us know how you experienced using the eye-tracking equipment

	1 (fully agree)	2	3	4	5 (fully disagree)
distracted me from airway evaluation	☐	☐	☐	☐	☐
distracted me from bag-mask ventilation	☐	☐	☐	☐	☐
distracted me while intubating	☐	☐	☐	☐	☐
distracted me because of the wire	☐	☐	☐	☐	☐
distracted me during the whole scenario	☐	☐	☐	☐	☐
felt uncomfortable on my nose	☐	☐	☐	☐	☐
felt uncomfortable on my ears	☐	☐	☐	☐	☐
felt uncomfortable for my eyes	☐	☐	☐	☐	☐
sled down on my nose	☐	☐	☐	☐	☐
impaired my visual field when evaluating the patient in general	☐	☐	☐	☐	☐
impaired my ability to see the patient monitor	☐	☐	☐	☐	☐
made me feel like I was not part of the team	☐	☐	☐	☐	☐
During the scenario, I forgot that I was wearing the glasses	☐	☐	☐	☐	☐
I would wear eye tracking glasses during a real scenario	☐	☐	☐	☐	☐

LimeSurvey

Section B: Skin sensors

Now, let's hear about how you felt using the skin sensor

B1. Please rate your experience with the skin sensor

	1 (fully agree)	2	3	4	5 (fully disagree)
I forgot that I was wearing the sensor during simulation	☐	☐	☐	☐	☐
I tried to refrain from touching the sensors	☐	☐	☐	☐	☐
I was worried the sensor might fall off	☐	☐	☐	☐	☐
I behaved differently then I normally would during such a scenario because of wearing the sensor	☐	☐	☐	☐	☐
I was disturbed in my tasks	☐	☐	☐	☐	☐
I felt the sensor moving	☐	☐	☐	☐	☐
I was concerned about my safety while using the sensor	☐	☐	☐	☐	☐
I was afraid the sensor might cause skin irritation	☐	☐	☐	☐	☐
I acted cautiously to not destroy the sensor	☐	☐	☐	☐	☐
My skin felt warm while wearing the sensor	☐	☐	☐	☐	☐
Wearing the sensor was tiring	☐	☐	☐	☐	☐
The sensor hindered me moving	☐	☐	☐	☐	☐
The sensor felt weird to wear	☐	☐	☐	☐	☐
The sensor was pulling my hair	☐	☐	☐	☐	☐
The sensor was painful to wear	☐	☐	☐	☐	☐
The sensor felt heavy	☐	☐	☐	☐	☐
The sensor caused itching	☐	☐	☐	☐	☐

Section C: Stress / Cognitive Load

This part deals with how you felt during the simulation scenario

C1. Please rate how strongly you felt the following emotions on a scale from 1 (lowest) to 10 (highest)

	1	2	3	4	5	6	7	8	9	10
Stressed	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

LimeSurvey

	1	2	3	4	5	6	7	8	9	10
Overwhelmed	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Motivated	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Thank you for filling out the debriefing survey.

LimeSurvey

This section deals with your thoughts at certain timepoints/events. Please let us know in a few words what you were thinking at the specific point in time during the scenario, as stated in the question. After that, please rate the mental effort required for solving each of the tasks provided

Section A: Cardiac arrest

A1. Please enter your ID number provided at the beginning

A2. What were your thoughts at the time when the patient went into cardiac arrest?

A3. Please rate how much mental effort you had to invest to manage the situation when the patient went into cardiac arrest:

	Very very low mental effort	Very low mental effort	Low mental effort	Rather low mental effort	Neither high nor low mental effort	Rather high mental effort	High mental effort	Very high mental effort	Very very high mental effort
I had to invest...	☐	☐	☐	☐	☐	☐	☐	☐	☐

Section B: Distraction

B1. What were your thoughts when the distraction occured?

B2. **Please rate how much mental effort you had to invest to manage the situation when the distraction occured:**

	Very very low mental effort	Very low mental effort	Low mental effort	Rather low mental effort	Neither high nor low mental effort	Rather high mental effort	High mental effort	Very high mental effort	Very very high mental effort
I had to invest...	☐	☐	☐	☐	☐	☐	☐	☐	☐

Section C: Intubation

C1. **What were your thoughts while you were intubating the patient?**

C2. **Please rate how much mental effort you had to invest to manage the situation when you had to intubate the patient:**

	Very very low mental effort	Very low mental effort	Low mental effort	Rather low mental effort	Neither high nor low mental effort	Rather high mental effort	High mental effort	Very high mental effort	Very very high mental effort
I had to invest...	☐	☐	☐	☐	☐	☐	☐	☐	☐

NASA TLX Questionnaire

Figure 8.6

NASA Task Load Index

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

Exclusion criteria

- For the group "emergency medicine novices": Regular participation (= at least once monthly in the last 12 months) in emergency doctor services.
- Individuals who wear glasses and cannot wear contact lenses, and whose correction through the eye tracking system (corrective lenses available from -4 to +4 diopters) is not possible.
- Any subjects with visual impairment, corresponding to a visual acuity < 1 in both eyes, whose visual impairment cannot be corrected to a visual acuity of 1 even with the use of contact lenses or corrective lenses from the eye tracking system.
- Subjects with eye diseases that make eye tracking impossible:
- Subjects diagnosed with cataracts.
- Subjects with diagnosed non-correctable strabismus (squint).
- Subjects with amblyopia (lazy eye) of any cause.
- Subjects with skin diseases/irritations or defects that make it impossible to attach the adhesive electrodes for skin resistance measurement.

Inclusion criteria

- Inclusion Criteria for "Emergency Medicine Experts":
  - Specialists in Anesthesiology.
  - Specialists in Internal Medicine with at least 2 years of experience in internal intensive care medicine.
  - Senior residents in Anesthesiology in their final year of training with completed specialty board examination.
  - Senior residents in Internal Medicine and internal intensive care medicine in their final year of training with completed specialty board examination.
- Inclusion Criteria for "Emergency Medicine Newcomers":
  - Physicians in general medical training ("turnus").
  - General practitioners (GPs).
  - Junior residents in Anesthesiology before rotation to an intensive care unit.
  - Junior residents in Internal Medicine before rotation to an intensive care unit.
  - Additionally, for physicians not affiliated with Anesthesiology, at least one of the following criteria to demonstrate basic skills in airway management:
    - Valid emergency medical services (EMS) certification.
    - Completed rotation in an anesthesiology department (including operating room).

- Minimum of 3 completed simulation trainings with intubation exercises.

Defined patient characteristics for simulation

Patient Information:

<table>
<tr><td>Age:</td><td>55yo</td></tr>
<tr><td>Sex:</td><td>Male</td></tr>
<tr><td>Weight</td><td>95 kg</td></tr>
<tr><td colspan="2">Anamnesis:
Chest pain/discomfort for the last 7 months on exertion, resolving at rest. Increasing symptoms during the last two weeks with more intense pain and decreasing time to symptom onset. Sudden onset of chest pain approx. 3 hours before at rest on day of presentation, ongoing.

Past medical history:
Diabetes mellitus Type II – treated orally.
Heavy smoker – approx. 40 pack years
Hypertension
Hypercholesterolemia – not treated, medication declined by patient.
Brother with Myocardial infarction at age 50.
No regular medical check-ups, Diabetes Medication prescribed by general practitioner.
Medication not known in detail, but no anticoagulation, inconsistent medication compliance
No allergies.</td></tr>
</table>

<table>
<tr><th colspan="2">Vital signs at presentation:</th></tr>
<tr><td>Heart rate:</td><td>105/min</td></tr>
<tr><td>Blood pressure:</td><td>75/40 mmHg</td></tr>
<tr><td>Respiratory rate:</td><td>24/min</td></tr>
<tr><td>SpO2</td><td>89% at room conditions</td></tr>
<tr><td>Recap Time:</td><td>>3 sec</td></tr>
</table>

<table>
<tr><th colspan="2">Clinical status at presentation:</th></tr>
<tr><td>Heart sounds:</td><td>Tachycardia, rhythmic, no murmurs</td></tr>
<tr><td>Heart rhythm:</td><td>Sinus tachycardia, ongoing</td></tr>
<tr><td>Lung sounds:</td><td>Bilateral fine crackles – beginning lung edema</td></tr>
<tr><td>Abdomen:</td><td>Soft, no pain, normal peristaltic sounds</td></tr>
<tr><td>Skin:</td><td>Cold-sweated at forehead/torso</td></tr>
<tr><td>Extremities</td><td>Cold, no/very weak peripheral pulse palpable</td></tr>
<tr><td>Glasgow Coma Scale</td><td>15</td></tr>
<tr><td colspan="2">General appearance:
• Stressed patient with ongoing pain
• Weak/no pulses peripheral
• Cold-sweated forehead + torso</td></tr>
</table>

Medical history information sheet

Master Thesis MDH
Simulation Briefing V1.0

Please note that your performance is of no interest to the study investigators and will not be rated in any way other than it would help you to improve your emergency skills. No performance results or analyses will be published in any form, neither the master thesis connected to this project nor potential future publications. We are only interested in the data collected by the sensors. So, try to relax and have "fun" ;)

Simulation scenario:

You are on your shift in the emergency department and get called by the nurses because a patient presents with severe chest pain. The triage sheet states the following information.

Patient Information:

Age:	55yo
Sex:	Male
Weight	95 kg
Anamnesis: Patient admits himself to the Emergency Department with chest pain, sudden onset approx. 3 hours before, ongoing.	

Vital signs at presentation:	
Heart rate:	105/min
Blood pressure:	75/40 mmHg
Respiratory rate:	24/min
SpO2	89% at room conditions

There is no therapy restriction existing for the patient, so you should provide everything you see an indication for.

Additional information:

The scenario starts once you enter the room. During the scenario, the team consists of you and two ICU/CCU nurses who happen to be on site. Your role is the one of the team leader. Your position is at the head of the patient. The hospital's cardiac arrest team or other support will not be available for the duration of the simulation as they are currently handling another emergency. You will be provided some information sheets that shall support you in performing certain tasks if necessary.

Definition of Times of Interest

Master Thesis MDH
Definition of TOI

Author	Christoph Plank, dh221815@fhstp.ac.at

Item type:	Times of Interest

Items:

Name	Definition Start - End	Sensor Type
Total scenario	Start of scenario – end of scenario = Start marker of first medical contact to end marker of ROSC	
First Medical Contact	Moment of first patient contact (= first fixation of patient) – +4,5 minutes	Eye Tracking / GSR
Cardiac arrest	Time of cardiac arrest at minute 5 of scenario +1 minute	Eye Tracking/(GSR)
Intubation	Time of pick-up of laryngoscope – removement of laryngoscope from mouth)	Eye Tracking – Blink Rate?
First rhythm analysis	Time of first rhythm analysis on defibrillator (verbal command) – time of decision how to succeed (verbal command)	Eye Tracking / (GSR)
Return of spontaneous circulation - ROSC	Time of last rhythm analysis – time of acknowledgement of ROSC (verbal confirmation)	GSR
ECG analysis	Time of Focus on ECG – end of ECG analysis (focus away from ECG)	Eye Tracking
Distraction	Time of distracting noise – 10 seconds after distraction	Eye Tracking / (GSR)