Doctoral thesis submitted to
the Faculty of Behavioural and Cultural Studies
Heidelberg University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy (Dr. phil.)
in Psychology

Title of the publication-based thesis

The Potential of Virtual and Mixed Reality for Training Medical First Responders

presented by
Anke Sabine Baetzner

year of submission
2025

Dean: Prof. Dr. Guido Sprenger
Supervisors: Prof. Dr. Henning Plessner, Prof. Dr. Cornelia Wrzus

Abstract

Well-trained medical first responders (MFRs) are essential to the management of mass casualty incidents (MCIs), where they must respond to more patients than their resources allow. However, real-life exercises (RLEs), the gold standard training method, remain infrequent due to the limited resources of MFR organizations. To fill this gap, MCI training in immersive virtual reality (iVR) and mixed reality (MR) has been gaining increasing attention. However, a thorough evaluation of the effectiveness of such training methods is essential to ensure adequate preparation of MFRs. This dissertation examines the potential of iVR and MR training for preparing MFRs for MCIs.

First, a systematic review was conducted on MCI and disaster training methods for MFRs as well as on the measurement of their effectiveness (Manuscript 1). Training methods from 55 (quasi-)experimental studies were evaluated. Most studies used written knowledge tests and subjective measures, while less than a third employed performance indicators. Two performance indicators that can be objectively and potentially automatically assessed were decision accuracy and speed. Technology-based methods mostly included learning on a computer screen. Three studies evaluated training with iVR and concluded that iVR was associated with similar learning satisfaction and similar or even higher performance than traditional methods. The review highlights a lack of evaluation studies on iVR and MR training, particularly MR with haptic feedback, and emphasizes the need to identify and validate objective performance indicators for iVR and MR MCI training.

Next, a study was conducted comparing MR MCI training with haptic feedback (n = 34) to the gold standard, RLEs with patient actors (n_1 = 4, n_2 = 14; Manuscript 2). The MR group reported similar stress, exhaustion, and self-efficacy scores as the RLE groups, but stress and exhaustion tended to be low in all groups. The MR group reported only slightly lower physical presence but considerably lower social presence. Open-ended responses suggested that this was

mainly due to a lack of interaction opportunities with virtual patients. Perceived learning gains were moderate for the MR group and high for the RLE group. Participants reported that iVR/MR training had potential and could complement RLEs in a resource-efficient way. This study was the first to compare MR MCI training with RLEs and demonstrated the potential of such training while also identifying areas for improvement, such as incorporating more stressors (in both types of training) and enhancing virtual patient interaction.

To identify objective performance indicators that can be integrated in iVR/MR MCI training, a validation study on performance metrics for virtual training was conducted (Manuscript 3). Based on the known-groups approach, indicators identified in the systematic review as well as other medical education research were tested in their ability to differentiate between different levels of expertise. Seventy-six participants from the medical field completed five iVR MCI scenarios. While visual attention indicators did not distinguish between expertise levels, MFRs demonstrated significantly better triage accuracy, speed, and information transmission efficiency than non-MFRs. Triage accuracy also correlated positively with triage knowledge test scores. Self-rated performance did not correlate with any of the objective indicators. Immersive virtual MCI scenarios proved to be a valuable medium for assessing MFR performance, particularly in terms of accuracy, speed, and information transmission. The results suggest that immersive virtual training could be integrated into current MCI training curricula to provide frequent, objective performance assessments in a controlled environment.

In summary, this dissertation demonstrates that immersive virtual training has the potential to complement existing MCI training, thereby enhancing MFR preparation. Specifically, such training elicited similar psychological responses in several areas when compared to the gold-standard training and received positive trainee feedback. Additionally, immersive virtual MCI training can be used to assess MFR performance, and several performance indicators are recommended for the integration in iVR and MR training.

List of Scientific Publications of the Publication-Based Dissertation

Manuscript 1

Baetzner, A. S., Wespi, R., Hill, Y., Gyllencreutz, L., Sauter, T. C., Saveman, B. I., Mohr, S. Regal, G. Wrzus, C. & Frenkel, M. O. (2022). Preparing medical first responders for crises: a systematic literature review of disaster training programs and their effectiveness. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *30*(1), Article 76. https://doi.org/10.1186/s13049-022-01056-8

Author contributions: ASB, YH, LG, BS, CW and MOF were involved in the conception and design of the work. ASB, RW and YH contributed to the screening process, data extraction and quality assessment of papers; ASB interpreted the data and drafted the manuscript. The manuscript was critically reviewed and enriched with medical expertise by LG, TCS, BS and SM; with psychological expertise by RW, YH, CW, MOF; with technological expertise by RW and GR. All authors read and approved the final manuscript.

Manuscript 2

Baetzner, A. S., Gerwann, S., Uhl, J., Schuster, B., Sjöberg, D., Wenighofer, R., Frenkel, M. O., Wrzus, C. (under review). Mixed Reality and Real-Life Exercises for Mass Casualty Incidents: A Comparison of Psychological Responses and Learning. *Educational Technology Research and Development.*

Author contributions: Study conception: ASB, CW, MOF; design of the study: ASB, CW; conception, organization, and/or delivery of training: BS, RW, DS; data acquisition: ASB, SG, JCU; quantitative data analysis: ASB; qualitative data analysis: ASB, SG, and JCU; drafting of manuscript: ASB; critical review: SG, JCU, BS, DS, RW, MOF, CW.

Manuscript 3

Baetzner, A. S., Hill, Y., Roszipal, B., Gerwann, S., Beutel, M., Birrenbach, T., Karlseder, M., Mohr, S., Salg, G., Schrom-Feiertag, H., Frenkel, M. O., Wrzus, C. (2025). Mass Casualty Incident Training in Immersive Virtual Reality: Quasi-Experimental Evaluation of Multimethod Performance Indicators. *Journal of Medical Internet Research (JMIR), 27,* Article e63241. http://dx.doi.org/10.2196/63241

Author contributions: ASB, YH, MOF, and CW designed the study. ASB, YH, CW, BR, and MK planned the scenarios, which were constructed by BR, MK, and ASB. The scenarios and patient information were reviewed by GS, TB, and SM. SG supported the technological set-up and developed software for standardized behavioral data collection and audio recordings with triage information. ASB and MB were responsible for data collection. ASB analyzed the data and wrote the manuscript. All authors reviewed and approved the manuscript.

Table of Contents

List of Abbreviations

AOI: area of interest (eye-tracking)

AR: augmented reality

CAMIL: Cognitive Affective Model of Immersive Learning

HMD: head-mounted display

iVR: immersive virtual reality

MCI: mass casualty incident

MFR: medical first responder

MR: mixed reality

RLE: real-life exercise

SSSS scheme: scene, safety, situation, and support scheme

START: simple triage and rapid treatment (triage algorithm)

VI zone: vehicle impact zone

VR: virtual reality

Chapter 1. General Introduction

Mass casualty incidents (MCIs) are defined as “events which generate more patients at one time than locally available resources can manage using routine procedures” (World Health Organization, 2007, p. 9). They encompass a variety of uncommon and complex situations, including natural disasters, large-scale accidents and man-made disasters (Berndt, Wessel, Willer, et al., 2018). Medical first responders (MFRs) are usually responsible for the initial, prehospital care and patient transport (Chaput et al., 2007). However, during MCIs, MFRs must perform a variety of tasks beyond their routine responsibilities. These tasks include gaining an overview of the usually complex situation while ensuring their own safety, making decisions about who is most in need of help (i.e., triage), and providing medical care and transport based on established priorities (Chaput et al., 2007). In addition, these tasks must be performed in situations, in which insufficient information flow, technical communication failures and mismatches between contingency plans and reality are not unusual (Hugelius et al., 2020). Maintaining optimal attention and performance in these high-stress situations is critical to saving as many lives as possible.

Altogether, MCIs are particularly stressful events for MFRs (Hugelius et al., 2020), who are under high pressure to perform. However, the preparedness of emergency organizations for MCIs and disasters has been criticized as insufficient (Mills et al., 2020). For example, a German study with n = 1993 MFRs found that only 73% had participated in a real-life exercise (RLE) involving a triage process at any point in their careers (Ellebrecht, 2013). It is therefore not surprising that a large part of healthcare professionals perceives their preparedness for these situations as inadequate (Almukhlifi et al., 2021), despite their willingness to participate in MCI and disaster training (Goniewicz et al., 2021; Lim et al., 2013; Whetzel et al., 2013).

The gold standard for MCI training is real-life scenario training with patient actors, manikins and props such as wrecked vehicles (Berndt, Wessel, Willer, et al., 2018; Tin et al., 2021). Simulating potentially real scenarios during training is considered important as it gives trainees the chance to transfer their theoretical knowledge into practical skills as well as to experience difficulties (e.g., unclear responsibilities) and work on them without putting real patients or themselves at risk (Heldring et al., 2024; Lateef, 2010). However, real-life scenario training tends to be particularly resource-intensive in terms of costs and organizational effort, resulting in a low training frequency (Berndt, Wessel, Willer, et al., 2018). To fill this gap, immersive virtual reality (iVR) and mixed reality (MR) are considered promising new training technologies with the potential to substantially increase the availability of MCI training (e.g., Elsenbast et al., 2024; Pottle, 2019). However, there is still a lack of evaluation studies scientifically testing iVR and MR for MCI training. As a result, it remains difficult to make informed, theoretically solid and scientifically grounded decisions about the use and application of iVR/MR MCI training. This dissertation aims to fill this gap by examining the theoretical basis, scientifically examining the potential of iVR/MR training in preparing MFRs for MCIs, and highlighting implications and directions for future research.

1.1 iVR and MR

The term *virtual reality* encompasses technological systems that present computer-generated, synthetic environments, typically experienced through audio-visual input, which users can immerse themselves in and interact with (Mikropoulos & Natsis, 2011; Milgram & Kishino, 1994). Virtual reality can be classified based on the level of immersion, i.e., the ability of the technological system to create convincing and vivid virtual environments, while blocking out the physical world (Makransky & Petersen, 2021; Slater & Wilbur, 1997). Systems with low immersion usually refer to viewing the virtual world on a computer monitor (Fusco & Tieri, 2022; Slater & Wilbur, 1997). In contrast, high immersion can be achieved through a sensory-rich environment, a broad field of view, and highly realistic graphics (Slater & Wilbur, 1997).

iVR is usually experienced through head-mounted displays (HMDs) and the use of controllers to interact with the virtual environment (Gerwann et al., 2024; Makransky & Petersen, 2021). HMDs create the impression of being in a three-dimensional environment by using binocular projection, providing each eye with a slightly different perspective on two-dimensional screens within the HMD (Scarfe & Glennerster, 2019). Another, less frequently used iVR system is a room in which virtual environments are projected onto the walls, floor, and ceiling (CAVE; Andreatta et al., 2010; Slater & Wilbur, 1997). However, most of the recent research focuses on iVR with HMDs (Makransky & Petersen, 2021; Scarfe & Glennerster, 2019), mainly because of the increasing affordability and practical use such as its transportability (Jensen & Konradsen, 2018).

While users of iVR are completely surrounded by a virtual environment, MR can be defined as the combination of the virtual and physical world (Dwivedi et al., 2022; Milgram & Kishino, 1994). MR encompasses both the experience of the physical world with virtual objects placed inside (usually referred to as augmented reality) and the experience of being surrounded by a virtual world with physical elements or objects incorporated (Dwivedi et al., 2022; Wrzus, Schöne, et al., 2024). The latter is sometimes also referred to as augmented virtuality (Milgram & Kishino, 1994) and usually encompasses the integration of haptic feedback by mapping virtually visible objects on top of physical ones (Gerwann et al., 2024; Zechner, Uhl, et al., 2024). However, due to the inconsistent use of the terms iVR and MR (Dwivedi et al., 2022), it is recommended to carefully examine the description of the used technology in research articles (Gerwann et al., 2024). In addition to iVR, this dissertation primarily focuses on the type of MR that integrates objects of the physical world to an otherwise virtual environment (i.e., 'augmented virtuality'). Medical professionals rely heavily on haptic cues received through their hands, which are one of the most important diagnostic tools in patient interactions (Zechner, Uhl, et al., 2024). Therefore, in the context of medical education, embedded physical objects are usually patient manikins (Elsenbast et al., 2024; Girau et al., 2019; Uhl, Schrom-Feiertag, et al., 2023; Zechner, Uhl, et al., 2024). Table 1.1 provides an overview of training modalities as defined in this dissertation and ordered by the degree to which trainees experience the physical vs. virtual world.

iVR and MR have already proven to lead to the feeling of presence, i.e., the subjective feeling of ‘actually being there’ (Slater & Sanchez-Vives, 2016; Slater & Wilbur, 1997; Uhl, Schrom-Feiertag, et al., 2023). Furthermore, the technologies provide increasingly realistic graphics and sound, including spatial audio (Xu et al., 2020). While most setups can track movements by tracking the position of the HMD and controllers, full-body motion tracking has become possible as well (Gerwann et al., 2024; Kleygrewe et al., 2024b). Motion tracking is necessary for enabling users to move around and explore the virtual environment. For instance, while 360-degree videos offer a highly realistic environment and can be valuable for researching reactions from a static position, artificially created scenarios can allow for free movement and interaction with the scene, which are critical for effective training (Scarfe & Glennerster, 2019). Technological progress in recent years has also made it possible for multiple users to be in the same virtual environment working together (Nguyen & Bednarz, 2020). This can be beneficial for MFR training as they usually work in teams (Manser, 2009). Once set up with the necessary hardware and software, iVR and MR can be used almost everywhere, at any time and with little preparation (Düking et al., 2018).

The strengths of immersive virtual environments have led to a rapid expansion of the use of iVR across multiple sectors and research areas in recent years (Wrzus, Schöne, et al., 2024), particularly in the context of education and training (Makransky & Petersen, 2021). As the technology has become more affordable (Mills et al., 2020) and technological advances continue, its adoption in medical education has also increased substantially. Meanwhile, MR is

Table 1.1

Overview of training modalities as defined in this dissertation.

	Real-life exercises	Augmented reality	Non-immersive Virtual reality	Augmented virtuality	Immersive virtual reality
Typical set-up	Staged environment with patient actors and props (Berndt, Wessel, Willer, et al., 2018; Mills et al., 2020)	‘See-through’ HMDs, glasses or mobile devices displaying the physical environment and overlaying virtual elements (Nizam et al., 2018)	Desktop monitor (Milgram & Kishino, 1994; Slater & Wilbur, 1997)	HMDs and integrated physical elements (Milgram & Kishino, 1994)	HMDs, or less common: CAVE systems (Makransky & Petersen, 2021; Slater & Wilbur, 1997)
Perception of physical environment	Complete; physical environment only	High; virtual elements integrated into the physical environment	High to medium; virtual content confined to the screen; physical environment remains visible	Medium to low; users primarily perceive the virtual environment, selected elements from the physical world are integrated	Low, users primarily perceive the virtual environment
Typically addressed senses	Possibly all senses	Vision, potentially combined with hearing (Dong et al., 2016)	Vision (screen), hearing (Foronda et al., 2016; Fusco & Tieri, 2022)	Vision, hearing, often haptics, other senses possible as well (Uhl, Schrom-Feiertag, et al., 2023)	Vision and hearing (Zhang et al., 2018)
Interaction techniques and tools with training environment	Own hands and body, use of ‘normal’ equipment (Uhl et al., 2024)	Controllers, tracked gestures, etc. (Nizam et al., 2018)	Mouse, keyboard, joystick, etc. (Fusco & Tieri, 2022)	Body trackers, sensors within integrated objects, controllers etc. (Gerwann et al., 2024)	Controllers, body trackers, etc. (Uhl et al., 2024)

Note. This table displays an overview of the training modalities as they are defined in this dissertation and how they are typically used. As there are always modifications and technical adaptations, the aim is not to provide an all-encompassing representation.

emerging as a promising technology in the medical field, although it is still in its early stages with limited research conducted so far (Girau et al., 2019).

Despite their potential, iVR and MR still face technological limitations that affect the users' experience. For example, accurate representation and assessment of fine motor tasks remain challenging due to limitations in motion tracking and, in MR, imperfect mapping of virtual elements onto physical ones (Gerwann et al., 2024). Motor skills are important in medical treatment, but there are concerns that exact hand grips cannot be incorporated into MR and thus unnatural motor sequences are being learned (Giessing, 2021). In addition, users may experience cybersickness, which describes negative side effects such as nausea and headaches during or after using iVR/MR (Davis et al., 2014). However, the risk of cybersickness can be reduced by minimizing sensory conflicts between bodily perception and virtual input. Strategies include reducing the lag between user actions and system responses, implementing efficient motion tracking, and enabling natural locomotion (Davis et al., 2014). Overall, while iVR and MR are promising technologies with numerous potential applications, they currently also have limitations. Further research is needed to evaluate their suitability for MCI training.

1.2 Training in iVR and MR

The term *training* describes the systematic and goal-oriented execution of exercises to develop or improve specific competencies and skills (Altmann, 2019). To trigger learning processes, training must challenge trainees beyond their current abilities (e.g., higher difficulty, new training content) without being overwhelming (Altmann, 2019; Hill et al., 2024). In particular, for quick and effective decision-making in complex situations such as MCIs, training should support the development of intuitive expertise. Individuals with high intuitive expertise can draw on previous experience, recognizing when a situation's relevant cues resemble those of a prototypical situation they have encountered before (Kahneman & Klein, 2009). This allows them to quickly reach a solution without much apparent effort or even conscious awareness of the reasons for their decisions (Hogarth, 2001). The development of intuitive expertise is based on accumulated learning experience (Kahneman & Klein, 2009) and depends on sufficient training time in so-called *kind learning environments* (Hogarth, 2001). In contrast to *wicked learning environments*, *kind learning environments* provide valid cues that are representative of the actual performance situation, along with unambiguous and rapid feedback (Hogarth, 2001, 2010). Simulation training is particularly well suited for creating such favorable learning conditions by providing a high level of representativeness.

The primary goal of simulation training is to improve performance not only within the training context, but also in real-life situations (Hutter et al., 2023). To achieve this, a key requirement for a high-quality simulation is a well-designed training scenario (Hutter et al., 2023). The principles of *Representative Learning Design* (Pinder et al., 2011) provide guidance for developing realistic and effective training scenarios. Originally applied to sport, this concept has been adapted to first responder contexts such as police training (Hutter et al., 2023). The central idea is that environmental dynamics must be kept intact in simulation training, allowing trainees to engage with realistic demands and ensuring optimal transfer of learned skills into practice. *Representative Learning Design* argues that learning occurs through a dynamic interaction between the learner and the environment. To facilitate this process, two key characteristics are essential for effective training: functionality and action fidelity (Pinder et al., 2011). Functionality refers to the idea that the same sources of information for decision-making and action should be present in a training simulation as they would be in real-life performance situations. Action fidelity emphasizes that the actions performed during training should closely resemble those that can be performed in real situations. Combining these aspects, trainees should be exposed to challenges that closely simulate real-world conditions, allowing them to practice solutions that they can apply in actual performance situations (Hutter et al., 2023). This training approach promotes the effective integration of perception and action processes, thereby increasing the likelihood of successful transfer of skills to real-life performance situations (Pinder et al., 2011). Ideally, representative simulations even trigger states similar to those that may occur during real-life task execution, such as stress and anxiety (Hutter et al., 2023).

Although high-fidelity RLEs are considered the gold standard for representative learning experiences for medical personnel (Tin et al., 2021), difficulties in conducting them frequently have led to a need for other options that provide authentic learning. Previous research suggests that immersive learning can complement and enhance traditional training by providing an authentic learning environment that reinforces the transfer of learning to the real world (Dede, 2009).

The *Cognitive Affective Model of Immersive Learning* (CAMIL; Makransky & Petersen, 2021) describes how iVR facilitates knowledge acquisition and learning transfer, emphasizing the critical role of immersion and realism in virtual training. The effectiveness of learning in iVR is largely influenced by the concepts of *presence* – the subjective feeling of actually being part of the virtual scenario (Slater & Wilbur, 1997) – and *agency*, which refers to the feeling of having control over one's actions and their consequences (Moore, 2016). Technological factors such as immersion, the ability to exert (immediate) control over the virtual situation (i.e., control factors), and a realistic virtual experience that closely mirrors the physical world (i.e., representational fidelity) promote feelings of presence and agency (Makransky & Petersen, 2021). Affective and cognitive factors such as increased interest, motivation and self-efficacy mediate the effects of presence and agency. Through these mediating factors presence and agency have an impact on both the acquisition of factual, conceptual, and procedural knowledge as well as the transfer of learning (Makransky & Petersen, 2021). The *CAMIL*, however, is not based on the idea that immersive virtual technology is always superior to other training methods, but that the medium interacts with the instructional method. While highlighting the potential of immersive learning, successful application requires a careful training evaluation to ensure that the virtual system is used appropriately and is aligned with the desired learning processes and outcomes (Makransky & Petersen, 2021).

In terms of MCI training, iVR and MR may be particularly beneficial because they provide the opportunity to practice high-risk situations, such as fires and sudden explosions, in a safe and controlled setting (Murtinger et al., 2021). In addition, vulnerable groups such as the elderly, children, and people with disabilities can be included virtually, allowing to practice the interaction with them during MCIs (Kent & Hughes, 2022). iVR/MR scenarios can also be tailored to meet specific training objectives and needs of trainees (Zechner, Kleygrewe, et al., 2023), and the difficulty can be adjusted continuously to create a difficulty level that leads to optimal learning without being overwhelming (Gerwann et al., 2024).

Although prior research on immersive learning reported mixed findings for learning factual and conceptual knowledge, immersive virtual training seems to be promising for increasing procedural knowledge and transfer to real-life situations (Makransky & Petersen, 2021). Nevertheless, the quality of studies on iVR and learning conducted so far has been criticized as below average, lacking validated evaluation measures and theoretical grounding (Jensen & Konradsen, 2018; Makransky & Petersen, 2021). This indicates a need for further high-quality research systematically testing iVR in learning contexts.

1.3 Stress, Attention, and Performance in iVR and MR MCI training

The *integrative framework of stress, attention, and human performance* (Vine et al., 2016) describes the associations between stress, attention, and performance: Facing a stressful performance situation, individuals consciously or subconsciously assess whether they have sufficient resources to cope with the situation’s demands. If individuals consider their coping resources as sufficient, they perceive the situation as challenging. This perception leads to a balanced interplay between goal-directed and stimulus-driven attentional systems, which enables optimal perception of relevant information and ultimately improves task performance. In contrast, when coping resources are subjectively inadequate, stress arises, and individuals perceive the situation as threatening. This triggers heightened stimulus-driven attentional control, leading to increased distractibility and a suboptimal uptake of information, ultimately resulting in poor task performance.

Based on the framework, training can avoid potential performance losses in two important ways. First, thorough preparation can help MFRs perceive the situation as manageable, allowing them to perceive it as a challenge rather than a threat. A central mechanism for this is the enhancement of self-efficacy (Bandura, 1977). Self-efficacy refers to the belief in one's ability to successfully perform the behavior needed to achieve a desired outcome (Bandura, 1977). Self-efficacy is associated with an increased expectation that the situation is manageable and with greater coping effort and persistence during adversity (Bandura, 1977). A particularly influential source of self-efficacy are prior performance achievements (Bandura, 1977), including those in simulated performance situations (Cardós-Alonso et al., 2024; Vincent et al., 2008). Consequently, mastery experiences in MCI simulation training have been shown to increase trainees' self-efficacy to effectively manage real MCI situations (Cardós-Alonso et al., 2024; Vincent et al., 2008). Second, the *integrative framework of stress, attention, and human performance* suggests that compensatory strategies can be used to mitigate the negative effects on performance that come with perceiving the situation as a threat. These compensatory strategies, such as arousal management techniques, can be developed and reinforced through training.

Overall, training is essential to ensure optimal attention and performance in high-stress situations such as MCIs. Effective training for first responders should be able to induce stress, allowing trainees to practice under pressure, understand how stress affects their attention and performance, and learn strategies how they can counteract this (Hill et al., 2024; Hutter et al., 2023). Previous research indicates that iVR and MR can be used to create stressful learning environments and assess changes in attention and performance over time. Immersive virtual training can elicit psychophysiological stress responses that are barely distinguishable from those in the physical world (Schöne et al., 2023; Zimmer et al., 2019), and has been shown to induce stress in first responder groups like the police and firefighters (Giessing et al., 2019; Kleygrewe et al., 2024b; Paletta et al., 2022). Additionally, a study with iVR MCI scenarios demonstrated that iVR can elicit increased stress levels in MFRs (Prachyabrued et al., 2019), indicating the suitability of iVR and MR for MCI training.

Furthermore, iVR and MR allow for the implementation of objective visual attention and performance indicators (Gerwann et al., 2024; Shinnick, 2016). High-quality performance assessment in simulation-based training is critical to providing structured, systematic training and accurately evaluating the preparedness of MFRs (Salas et al., 2009). Furthermore, improvements in attention and performance are key metrics for evaluating training and comparing training methods (Markou-Pappas et al., 2024). However, there is still a need to identify appropriate metrics from relevant previous virtual and non-virtual training studies and test their applicability to iVR and MR MCI training.

1.4 Dissertation Overview and Research Questions

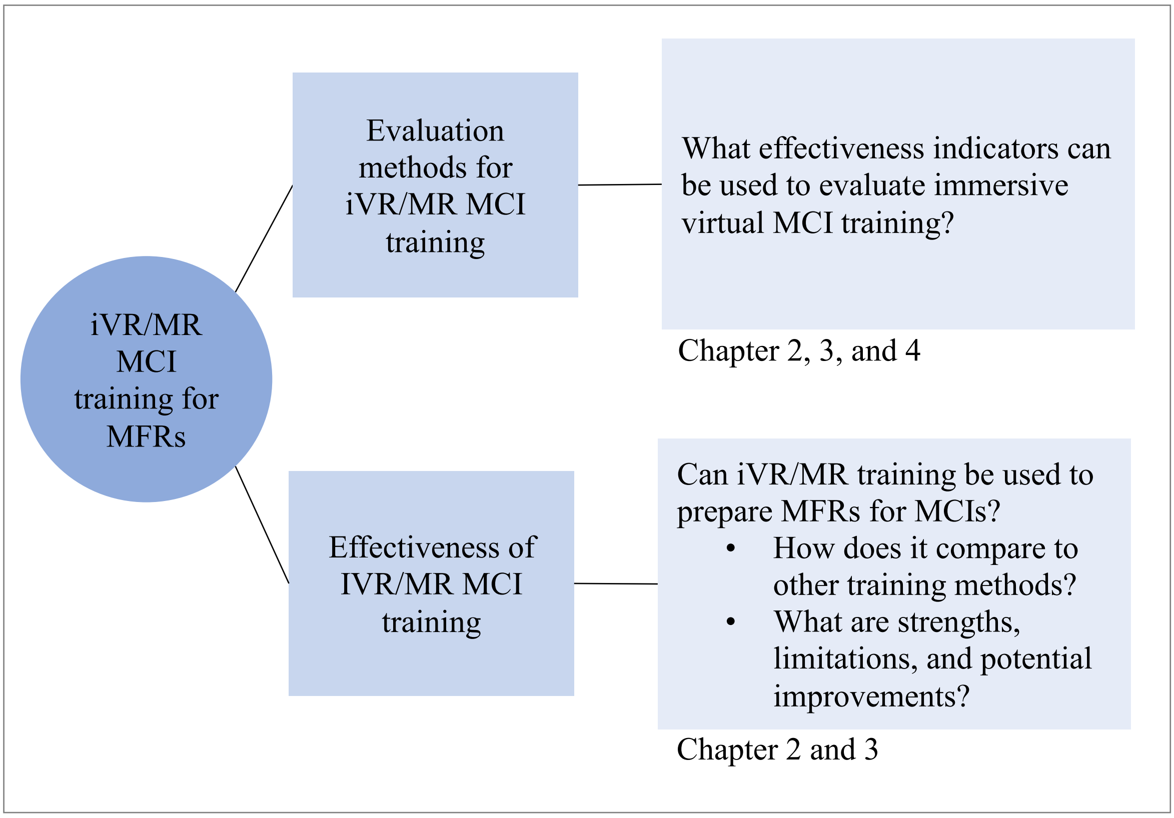
The overall research aim of this dissertation is to derive evidence-based insights into whether iVR and MR can be used to train MFRs for MCIs. This will enable MFR organizations to make more informed decisions about the use of these technologies. Figure 1.1 provides an overview of the main foci of this dissertation. This dissertation aims to answer the following research questions:

1) What effectiveness indicators can be used to evaluate immersive virtual MCI training?

2) Can iVR and MR training be used to prepare MFRs for MCIs?

Figure 1.1

Overview of the main foci of this dissertation.



Specifically, Chapter 2 systematically reviews prior research on MCI and disaster training for MFRs. It thereby provides an overview on the training methods that have already been evaluated and the effectiveness indicators that were used. A particular focus was on iVR and MR, with the aim of finding out to which extent they have already been tested. It further includes an overview of the effectiveness of the different methods.

Chapter 3 builds upon the results of Chapter 2, which found that there was no study testing an MR MCI training that fits the definition of MR in this dissertation, i.e., the integration of physical objects to an otherwise virtual environment. As a result, Chapter 3 compares an MR MCI training to the gold standard, RLEs with patient actors, regarding quantitative and qualitative outcome variables.

While Chapters 2 and 3 suggest the potential of immersive virtual training, results in Chapter 2 indicated that only few studies used objective performance indicators. The possibility to implement (at least partially) automated, objective performance assessments may be a particular strength of iVR and MR. However, there was still a need for the evaluation and validation of potential performance indicators that can be used for MCI training in immersive virtual environment. Using performance indicators identified in Chapter 2 as well as other indicators from medical education research, Chapter 4 evaluates the suitability of different performance indicators for virtual MCI training.

In Chapter 5, the study results are integrated and discussed, providing an overview of strengths and limitations as well as future directions and practical implications.

Chapter 2. Preparing Medical First Responders for Crises: a Systematic Literature Review of Disaster Training Programs and their Effectiveness

Anke S. Baetzner[1], Rafael Wespi[2,11], Yannick Hill[3,9,10], Lina Gyllencreutz[4,5], Thomas C. Sauter[2], Britt-Inger Saveman[4,5], Stefan Mohr[6], Georg Regal[7], Cornelia Wrzus[8], & Marie O. Frenkel[1]

[1]Institute for Sports and Sports Sciences, Heidelberg University, Germany
[2]Department of Emergency Medicine, Inselspital, University Hospital, University of Bern, Switzerland
[3] Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam, Netherlands
[4]Department of Nursing, Umeå University, Sweden
[5]Department of Surgical and Perioperative Sciences, Umeå University, Sweden
[6]Department of Anesthesiology, University Hospital Heidelberg, Germany
[7]Center for Technology Experience, AIT Austrian Institute of Technology GmbH, Austria
[8]Psychological Institute and Network Aging Research, Heidelberg University, Heidelberg, Germany
[9]Institute of Brain and Behaviour Amsterdam, Amsterdam, Netherlands.
[10]Lyda Hill Institute for Human Resilience, Colorado Springs, USA. 11 Graduate School for Health Sciences,
[11]University of Bern, Bern, Switzerland.

This chapter may not be an exact copy of record of the officially published paper. The original reference can be found here:

Baetzner, A. S., Wespi, R., Hill, Y., Gyllencreutz, L., Sauter, T. C., Saveman, B. I., Mohr, S., Regal, G., Wrzus, C., & Frenkel, M. O. (2022). Preparing medical first responders for crises: a systematic literature review of disaster training programs and their effectiveness. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 30,* Article 76. https://doi.org/10.1186/s13049-022-01056-8

2.1 Abstract

Background: Adequate training and preparation of medical first responders (MFRs) are essential for an optimal performance in highly demanding situations like disasters (e.g., mass accidents, natural catastrophes). The training needs to be as effective as possible, because precise and effective behavior of MFRs under stress is central for ensuring patients' survival and recovery. This systematic review offers an overview of scientifically evaluated training methods used to prepare MFRs for disasters. It identifies different effectiveness indicators and provides an additional analysis of how and to what extent the innovative training technologies virtual (VR) and mixed reality (MR) are included in disaster training research.

Methods: The systematic review was conducted according to the PRISMA guidelines and focused specifically on (quasi-)experimental studies published between January 2010 and September 2021. The literature search was conducted via Web of Science and PubMed and led to the inclusion of 55 articles.

Results: The search identified several types of training, including traditional (e.g., lectures, real-life scenario training) and technology-based training (e.g., computer-based learning, educational videos). Most trainings consisted of more than one method. The effectiveness of the trainings was mainly assessed through pre-post comparisons of knowledge tests or self-reported measures although some studies also used behavioral performance measures (e.g., triage accuracy). While all methods demonstrated effectiveness, the literature indicates that technology-based methods often lead to similar or greater training outcomes than traditional trainings. Currently, few studies systematically evaluated immersive VR and MR training.

Conclusion: To determine the success of a training, proper and scientifically sound evaluation is necessary. Of the effectiveness indicators found, performance assessments in simulated scenarios are closest to the target behavior during real disasters. For valid yet inexpensive evaluations, objectively assessible performance measures, such as accuracy, time, and order of actions could be used. However, performance assessments have not been applied often. Furthermore, we found that technology-based training methods represent a promising approach to train many MFRs repeatedly and efficiently. These technologies offer great potential to supplement or partially replace traditional training. Further research is needed on those methods that have been underrepresented, especially serious gaming, immersive VR, and MR.

Keywords: Emergency medical technicians, emergency medicine, mass casualty incident, medical education, mixed reality, paramedics, performance, prehospital care, simulation, virtual reality

2.2 Background

Natural and man-made disasters such as floods, mass-accidents, and terrorist attacks are ubiquitous and cause loss of life, human suffering, and infrastructural damage (Centre for Research on the Epidemiology of Disasters, 2020; Srivastava, 2010). They create particularly demanding situations for emergency services, as they are unforeseen and usually sudden events that exceed local capacity and resources to rescue and care (Centre for Research on the Epidemiology of Disasters, 2020). During such disasters, medical first responders (MFRs), who are responsible for the initial prehospital care in medical emergencies, play a key role (Alharbi, 2018; Chaput et al., 2007). However, numerous healthcare professionals, including MFRs, perceive their preparedness for the response to disasters as inadequate (Almukhlifi et al., 2021). As previous research indicates, a higher training frequency and better training quality are associated with increased disaster preparedness (Almukhlifi et al., 2021). To enhance the overall quality of MFR training, the aim of this review is to provide an overview of scientifically evaluated training methods and to examine whether certain methods seem to be particularly effective. Furthermore, indicators used to evaluate the training effectiveness will be identified so that future research can be guided by existing training evaluation methods. Finally, the emergence of new, immersive technologies, including virtual (VR) and mixed reality (MR; Milgram & Kishino, 1994), has led to the development of new training programs which are becoming increasingly accessible to educators in the medical sector (Barrie et al., 2019). Therefore, we will draw particular attention to the role of immersive technologies by providing an additional analysis of how and to what extent VR and MR specifically are included in current disaster training research.

MFRs typically include paramedics and emergency medical technicians (Chaput et al., 2007), but the term may also refer to physicians, ambulance specialist nurses, and trained volunteers depending on a country’s emergency medical service systems (Beyramijam et al., 2021; Yafe et al., 2019). During disasters, MFRs take on a variety of tasks such as the initial scene evaluation, triage, medical care, and the transport of patients (Chaput et al., 2007). They have to perform those tasks under stressful and challenging conditions, such as difficult access to the disaster site, multiple injured people and disruption in communication systems (Sorani et al., 2018). In order for MFRs to adapt to these unusual conditions, they require specifically tailored training.

Effective training involves a systematic and goal-oriented execution of exercises for the acquisition or increase of specific competences and skills (Altmann, 2019). The general idea of training is to challenge the current level of performance (e.g., higher intensity, higher difficulty, new content) without being too overwhelming, so that the trainee can adapt and reach a higher performance level (Altmann, 2019; Giessing & Frenkel, 2022; Hill et al., 2020; Hill et al., 2021; Kiefer et al., 2018). However, training resources, including time, budget, and facilities, are usually limited. Therefore, training methods must be not only effective, but also match the resources of the rescue organization.

Despite the necessity of adequately preparing MFRs for disasters, no systematic and up-to-date overview of scientifically evaluated training methods and their effectiveness exists. Ingrassia and colleagues conducted an internet-based search via Google and Bing and identified several disaster management curricula at a postgraduate level with a large variety of methods, e.g., lectures and discussion-based exercises (Ingrassia, Foletti, et al., 2014). The trainings' effectiveness, however, was not evaluated. Assessing studies published between 2000 and 2005, Williams and colleagues (2008) concluded that the available evidence had not been sufficient to determine whether disaster training can effectively increase the knowledge and skills of MFRs and in-hospital staff. Because these findings are derived from studies conducted more than 15 years ago, new insights have most likely emerged and new training methods may have been added following recent technological advances.

Two of those new methods are VR and MR. In VR training, users are placed inside a simulated, artificial, three-dimensional environment in which they can interact with their digital surroundings (Milgram & Kishino, 1994). VR can either be screen-based using computer monitors or experienced in more immersive forms: Through head-mounted displays or certain rooms equipped with several large screens or projections on several walls (i.e., CAVE system; Giessing, 2021; Milgram & Kishino, 1994). In contrast, mixed reality combines the real and virtual worlds and refers to the whole spectrum between reality and VR. MR, for example, includes augmented reality (AR) in which users see their real surroundings supplemented with virtual objects (Milgram & Kishino, 1994). A specific application from the medical field may be the visual insertion of patient information during practice. Given the rapid development of immersive technology, this review provides an additional analysis of the role of VR and MR training.

Altogether, the following research questions are addressed:

1. Which current disaster training methods for MFRs have already been scientifically evaluated?
2. Which effectiveness indicators are used to evaluate MFR disaster training methods?
3. Based on the findings of the reviewed studies, which methods for MFR disaster training seem to be effective?
4. How and to what extent are VR and MR used to prepare MFRs for disasters?

2.3 Methods

The preregistered (osf.io/yn5v3) systematic literature search was conducted in accordance with the PRISMA guidelines (Moher et al., 2015).

2.3.1 Search strategy

The search strategy was prepared with support of a medical information specialist to ensure the appropriateness of the search terms. Using the search engines Web of Science and PubMed, we applied search terms such as *health personnel*, *training*, and *disaster* (see Additional File 2.1[1] for the search string). To ensure that the results reflect current training methods, the electronic search was limited to studies published between January 2010 and September 2021. A filter limited our results to studies with a full text in English.

2.3.2 Inclusion and exclusion criteria

We included articles that described a training or training session (e.g., drill, lectures, mixed methods training, etc.) conducted to improve the participants' *prehospital* disaster response. The training had to address prehospital content, but was allowed to also contain in-hospital topics. Participants had to be MFRs, regardless of whether they were still in training or already had work experience. In addition, to ensure adequate assessments of the effectiveness, we only considered (quasi-)experimental designs in which outcomes were compared to a control or comparison group [i.e., randomized controlled trials (RCTs), non-RCTs and at minimum pre-post testing of the same group; (Evans, 2003; Stratton, 2019)].

We excluded studies that a) did not test the effectiveness of a disaster training for MFRs, b) contained other occupational groups or not sufficiently specified groups (e.g., "others") in addition to MFRs without reporting separate analyses for the MFRs, c) were not primary studies published in peer-reviewed journals, and d) had no full-text available.

[1] Please note that in this dissertation, chapter numbers have been added to numbers of tables, figures, and supplementary files for clearer identification.

2.3.3 Selection process

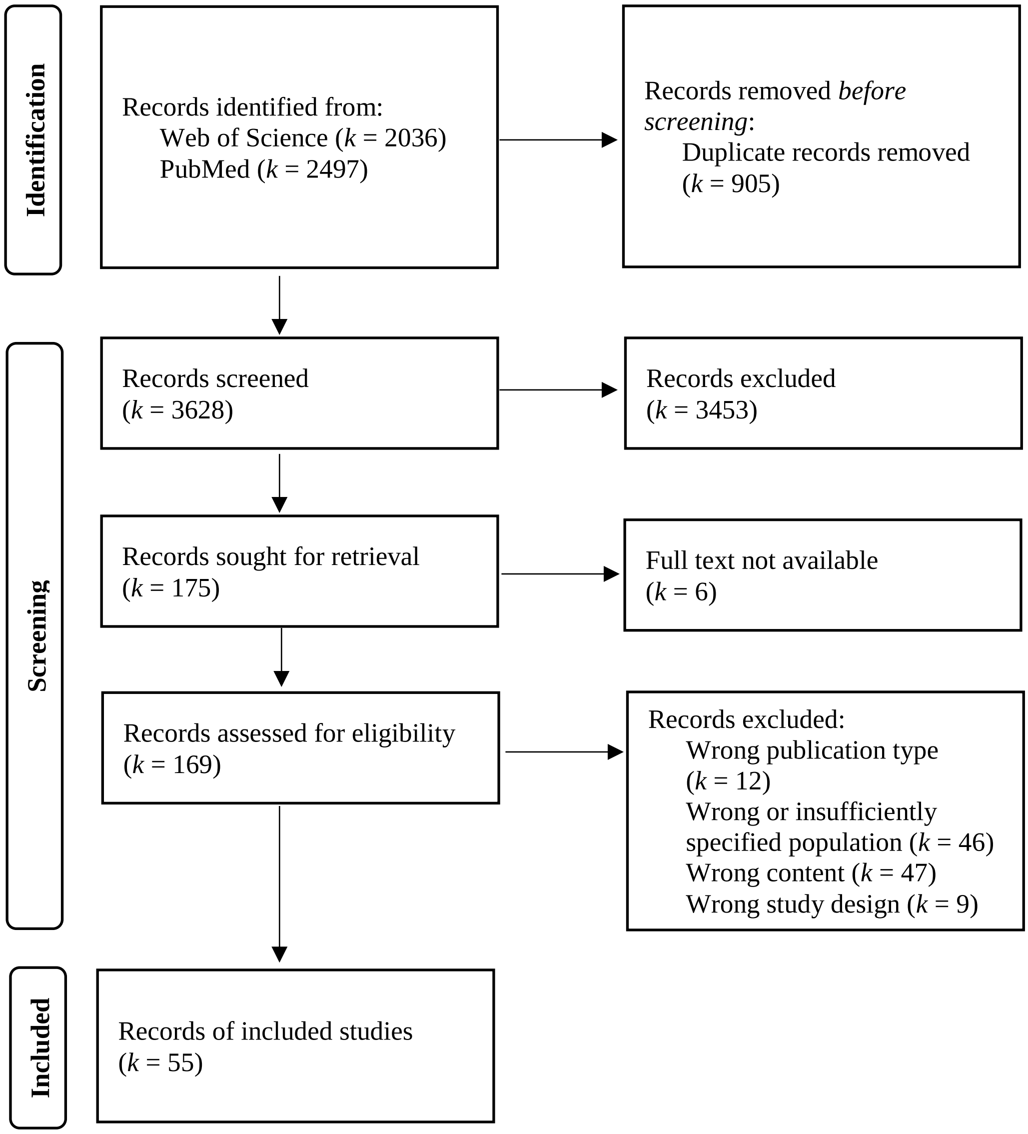
The search was conducted on 28th October 2021 and led to 4533 hits (Figure 2.1). Duplicates were identified with the software Endnote™ (Version 20.1) and additional visual screening. Two raters (ASB and RW) independently screened the remaining hits and performed the study selection using the web application Rayyan (Ouzzani et al., 2016). Discrepancies in the study selection process were resolved by consensus or, if necessary, together with a third rater (YH). Fifty-five studies were included in the review.

2.3.4 Data collection and analysis

Two raters (ASB and RW) extracted the relevant information for each article. Again, discrepancies were resolved by consensus, and when necessary together with the third rater (YH). Whenever studies used multiple methods at different time points we only considered those applied between the pre- and post-measurement. For trainings evaluated in (non-)RCTs without a pre-test, methods must have been applied before the post-test comparison with control groups. Similarly, only effectiveness indicators with sufficient informative value about training success or failure were considered (i.e., indicators used for pre-post comparisons or for comparisons with control groups). To assess the studies' quality and risk of bias, we used the Joanna Briggs Institute (JBI) critical appraisal checklists for RCTs and quasi-experimental studies (Tufanaru et al., 2020). The JBI tool for quasi-experimental studies contains nine questions and the JBI tool for experimental studies consists of 13 questions (e.g., "Were outcomes measured in a reliable way?"). There are four possible answer options: *yes*, *no*, *unclear*, *not applicable*. The answer *yes* indicates quality while *no* indicates a risk of bias.

Figure 2.1

Flow diagram of study selection.



2.4 Results

The majority of studies used a single group pre-post design (*k*=35). Other study designs included non-RCTs (*k*=6) and RCTs (*k*=14) with 15 out of 20 containing pre-post testing (see Table 2.1 for a full overview and Additional File 2.1: Table 2.1 for further information). The sample sizes varied largely between studies (range: 6 – 524). Trainings took place on several continents with the majority of trainings conducted in North America (*k*=24), followed by Asia (*k*=18), Europe (*k*=8), Australia and Africa (both *k*=2), unclear (*k*=1). The majority of tested trainings addressed general disaster management or several disaster-related topics (*k*=31), followed by triage (*k*=14), trauma management/sonography (*k*=3) etc. Furthermore, the time spans varied between one day or less (*k*=22) to up to eight months (*k*=33).

2.4.1 Research question 1: Overview of training methods

The majority of studies reported trainings that contain a combination of several methods, either in the intervention group, control group, or in both (*k*=42). Training methods could be categorized into traditional and technology-based methods (Figure 2.2). Traditional categories reflect lectures, real-life scenario training (e.g., mass casualty incident simulations with actors or manikins), discussion-based training (including seminars, workshops, in-class games, tabletop exercises), practical skills training (e.g., regional anesthesia), field visits (e.g., the visit of disaster affected sites or riding with the prehospital physician vehicle), and debriefings. In contrast, the technology-based category is composed of computer-based learning (i.e., online learning, educational computer programs), screen-based serious gaming, educational videos, and VR/MR. The term serious gaming refers to computer-based learning that additionally contains game elements, such as cooperation, competition, and stories (Ma et al., 2021).

Table 2.1

Overview of included studies.

First author, year	Study design	N[1]	Professions	Training content	Training methods	Duration, timespan	Effectiveness indicators	Effectiveness confirmation
Aghababaeian et al. (2013)	Non-RCT (pre-post)	144	Paramedics, EMTs	Triage	IG: Educational video CG: Real-life scenario	0.5h, 1 day	Basic knowledge test	No sign. difference between groups
							Applied knowledge test (triage accuracy)	IG partially better than CG
Alenyo et al. (2018)	Single group, pre-post	129	Emergency medical service providers	General disaster management	Not specified	NA, NA	Applied knowledge test (triage accuracy)	Yes
Alim et al. (2015)	Single group, pre-post	309	Nursing students	General disaster management	Not sufficiently clear ("in-class training")	8h, 1 day	Basic knowledge test	Yes
Aluisio et al. (2016)	Single group, pre-post	12	Physicians, nurses	Lower extremity regional anesthesia for earthquake victims	Discussion-based training, practical skills	5.25h, 1 day	Basic knowledge test	Yes
Andreatta et al. (2010)	Stratified RCT (pre-post)	15	Physicians	Triage	IG: Lecture, VR with CAVE CG: Lecture, real-life scenario	NA, 1 day	Basic knowledge test	CG better than IG (effect sizes only)
							Observed performance with a self-developed instrument to compose an overall performance score during real-life / VR simulation (e.g., ensure safety on scene, call for additional help, accuracy)	IG better than CG (effect sizes only)

Andreatta et al. (2015)	Non-RCT (pre-post)	204	Nurses, paramedics, medical students	Cholinergic crisis management	IG (use of virtual animal model): Discussion-based training, real-life scenario, computer-based learning CG (use of virtual human model): Discussion-based training, real-life scenario, computer-based learning	4h, 1 day	Basic knowledge test Observed performance with a previously validated instrument (Andreatta et al., 2014) to compose an overall performance score during real-life simulation (multiple performance dimensions associated with managing a nerve agent casualty) Self-reports of self-efficacy and state affect	Yes, for both groups (no sign. difference between groups)
Bajow et al. (2016)	Single group, pre-post	29	Medical students	General disaster management	Lecture, discussion-based training, real-life scenario, VR on screen, educational video, debriefing, field visit	53h, 2 weeks	Basic knowledge test	Yes
Betka et al. (2021)	Single group, pre-post	17	Medical students, nursing students[2]	General disaster management	Discussion-based training, computer-based learning	NA, NA	Self-rep. interprofessional collaborative competencies	Partially confirmed
							Self-rep. disaster management competence	Yes
							Self-rep. self-confidence in managing disasters	Partially confirmed
Chan et al. (2010)	Single group, pre-post	138	Nursing students	General disaster management	Lecture, discussion-based training, practical skills, field visit	NA, 2 weeks	Self-rep. competence in disaster nursing	Yes
Chandra et al. (2014)	Single group, pre-post	76	Professional volunteers	Psychological first aid	Lecture, discussion-based training, educational video	2h 1 day	Basic knowledge test	No
							Self-rep. capability in using psychological first aid	Yes
Chou et al. (2021)	Single group, pre-post	48	Medical students	General disaster management	Lecture, discussion-based training, real-life scenario	NA, 2 days	Basic knowledge test	Partially confirmed
							Self-rep. willingness to pursue further training	No
							Interest in disaster training	No

Cicero et al. (2012)	Single group, pre-post	50	Physicians	Triage	Lecture, real-life scenario, debriefing	3.5h, 5 months	Observed triage performance during real-life simulation (accuracy)	Yes
Cicero et al. (2017)	Single group, pre-post	261	Paramedics, EMTs, paramedic students	Triage	Real-life scenario, debriefing, computer-based learning	5h, 6.5 months	Observed triage performance during real-life simulation (accuracy)	Yes
Cowling et al. (2021)	Single group, pre-post	26	EMTs in training	General disaster management	Real-life scenario, debriefing	NA, 1 day	Basic knowledge test	No
							Self-rep. knowledge	Yes
							Self-rep. confidence in managing a structural collapse scenario	Yes
Cuttance et al. (2017)	RCT (only post)	292	Paramedics	Triage	IG1: lecture, other (aid memoire) IG2: lecture IG3: other (aid memoire) CG: no intervention	NA, NA	Applied knowledge test (triage accuracy)	Yes, all IGs improved with the greatest improvements in IG1 and IG3
Dittmar et al. (2018)	Single group, pre-post	19	Paramedics[2]	Triage	Lecture	0.75h, 1 day	Observed triage performance during real-life simulation (time + accuracy + overall performance score assessed with a self-developed instrument that incorporated e.g., accuracy, time, airway handling and bleeding control measures)	Yes
Edinger et al. (2019)	Single group, pre-post	19	Emergency medical service providers	Interacting with individuals with developmental disabilities during disaster response	Computer-based learning	1h, 1 day	Basic knowledge test	Partially confirmed
							Self-rep. self-efficacy for caring for developmentally disabled individuals	Yes

Farra et al. (2013)	RCT (pre-post)	47	Nursing students	General disaster management	IG: computer-based learning, VR on screen CG: computer-based learning	NA, NA	Basic knowledge test	IG partially better but CG improved as well (within-group effect of CG not tested for significance)
Fernandez-Pacheco et al. (2017)	Single group, pre-post	35	Nursing students	Triage	Debriefing with video	NA, NA	Change in self-perception (number of behaviors, moments, thoughts, feelings, strengths + weaknesses being described)	Partially confirmed
Foronda et al. (2016)	Single group, pre-post	6	Nursing students	Triage	Lecture, VR on screen, debriefing	1.25, 1 day	Applied knowledge test (triage accuracy)	No
Furseth et al. (2016)	Non-RCT (pre-post)	189	Nursing students, paramedic students	General disaster management	IG with a training focus on handoff communication: lecture, real-life scenario, debriefing CG: lecture, real-life scenario, debriefing	NA, 1 day	Self-rep. attitude towards Interprofessional education and interprofessional healthcare teams Self-rep. confidence Satisfaction with the training	IG partially better
Greco et al. (2019)	Single group, pre-post	90	Nursing students	Triage	Real-life scenario, debriefing	NA, NA	Self-rep. ethical reasoning confidence perceived importance of ethical reasoning	Yes
Huh and Kang (2019)	RCT (pre-post)	60	Nursing students	General disaster management	IG: Lecture, discussion-based training, practical skills, educational video CG: No intervention	8h, 4 weeks	Basic knowledge test Self-rep. disaster readiness	Yes
Hutchinson et al. (2011)	Single group, pre-post	81	Nursing students	General disaster management	Lecture, discussion-based training, real-life scenario, educational video, computer-based learning	NA, NA	Basic knowledge test	Partially confirmed

Ingrassia et al. (2015)	RCT (pre-post) cross-over design	56	Medical students	Triage	IG: Lecture, VR on screen CG: Lecture, real-life scenario training	NA, 3 days	Triage performance observed during real-life/virtual simulation (accuracy of triage decisions and of decisions about the need for lifesaving treatments + time)	No sign. difference between groups but both groups improved significantly
Ingrassia, Ragazzoni, et al. (2014)	Single group, pre-post	524	Medical students	General disaster management	Discussion-based training, computer-based learning, debriefing	47h, 1 month	Basic knowledge test Applied knowledge test (triage accuracy)	Yes
James et al. (2021)	Single group, pre-post	34	Nursing students	General disaster management	Real-life scenario, debriefing	4h, 1 day	Self-rep. attitudes towards teamwork in training	Yes
Jones et al. (2014)	Single group, pre-post	224	Paramedics, EMTs	Active shooter incident response	Lecture, real-life scenario	4h, 1 day	Self-rep. preparedness and attitudes towards active shooter incident response	Yes (but only descriptive statistics)
Kim and Lee (2020)	Single group, pre-post	34	Nursing students	General disaster management	Lecture, discussion-based training, real-life scenario, debriefing	3h, 1 day	Self-rep. attitudes towards responding to MCIs	Yes
Knight et al. (2010)	Non-RCT (only post)	91	Physicians, nurses, paramedics	Triage	IG: lecture, serious gaming with VR on screen CG: lecture, discussion-based training	NA, NA	Observed triage performance during real-life simulation (triage accuracy + compliance with the correct procedure + time)	Accuracy: IG better than CG, step accuracy: IG partially better, time: no sign. difference between groups
Koca and Arkan (2020)	RCT (pre-post)	235	Nursing students	General disaster management	IG: computer-based learning CG: no intervention	11h, 2 weeks	Self-reports of preparedness and self-efficacy	Yes
Koutitas et al. (2021)	RCT (only post)	30	EMTs	Familiarization with the ambulance bus (AMBUS)	IG1: lecture, VR with HMD IG2: lecture, AR with a head-mounted display CG: lecture, field visit	NA, 1 week	Observed performance in a real ambulance bus (orientation in ambulance bus; accuracy + time + overall performance score calculated from accuracy and time)	Yes, for both IGs with better results in IG1 (only descriptive statistics reported)

Kuhls et al. (2017)	Single group, pre-post	78	Physicians, nurses[2]	General disaster management	Not specified	8h, 1 day	Self-rep. confidence to manage disaster scenarios	Yes
Lampi et al. (2013)	Single group, pre-post	153	Physicians	Trauma life support	Lecture, discussion-based training, practical skills	NA, NA	Applied knowledge test (triage accuracy)	No
Lennquist Montán et al. (2015)	Single group, pre-post	83	Emergency medical services providers[2]	General disaster management	Discussion-based training	NA, 3 days	Self-rep. knowledge	Yes
							Self-rep. skills	
Ma et al. (2021)	RCT (pre-post)	104	Nursing students	General disaster management	IG: screen-based serious gaming CG: discussion-based training, real-life scenario, practical skills, debriefing	2h, NA	Self-rep. competence in disaster nursing	IG better than CG
Merlin et al. (2010)	Single group, pre-post	46	Medical students	General disaster management	Lecture, practical skills, field visit/clerkship	34.5h, 4 weeks	Self-rep. knowledge	Yes
							Self-rep. opinions about prehospital issues	
Mills et al. (2020)	RCT (only post) cross-over design	29	Paramedic students	Triage	IG: VR with HMD CG: real-life scenario	NA, 1 day	Triage performance (accuracy) during real-life/VR simulation	No sign. difference between groups
							Immersion (via heart-rate and self-rep.)	Greater immersion in real-life training
							Learning satisfaction	No sign. difference between groups
Motola et al. (2015)	RCT (pre-post)	91	Paramedics	Managing CBRNE incidents	IG: Educational video Waiting CG: no intervention	NA, NA	Basic knowledge test	Yes
							Observed performance in treating a CBRNE patient (overall performance score assessed with an instrument based on a previous study; Scott et al., 2006)	Partially confirmed

Paddock et al. (2015)	Stratified RCT (pre-post)	36	Physicians, nurses, paramedics, EMTs	Prehospital focused assessment with sonography in trauma	IG1: Educational video, computer-based learning CG: Educational video, practical skills IG2: Both of the trainings above	4h, 1 day	Basic knowledge test	No sign. difference between groups but all trainings led to significant improvement
							Observed sonography performance (overall performance scores for image acquisition and interpretation assessed with a self-developed instruments)	no sign. difference between groups
Phattharapornjaroen et al. (2020)	Single group, pre-post	52	Physicians	General disaster management	Discussion-based training	NA, 2 days	Self-rep. knowledge	Yes
Pollard et al. (2015)	Single group, pre-post	41	Medical students	General disaster management	Lecture, real-life scenario, computer-based learning	NA, 8 months	Basic knowledge test	Yes
Pouraghaei et al. (2017)	Single group, pre-post	205	EMTs	Triage	Lecture, discussion-based training	2h, 1 day	Basic knowledge test	Yes
							Applied knowledge test (triage accuracy)	
							Observed performance of managing the jaw trust airway maneuver (overall performance score via expert evaluations of success vs. failure)	
Ripoll-Gallardo et al. (2020)	Single group, pre-post	8	Physicians	General disaster management	Discussion-based training, real-life scenario, educational videos, computer-based learning, field visits	NA, 6 months	Basic knowledge test	Yes
							Observed performance in real-life simulation in a low-resource emergency room (overall performance score assessed with the TIGR (Franc et al., 2017))	Yes
							Self-rep. attitude towards disaster management domains	No
Rivkind et al. (2015)	Single group, pre-post	309[3]	Medical students	Trauma management	Lecture, discussion-based training, real-life scenario, practical skills, educational video, computer-based learning, debriefing	77h, 2 weeks	Basic knowledge test	Yes

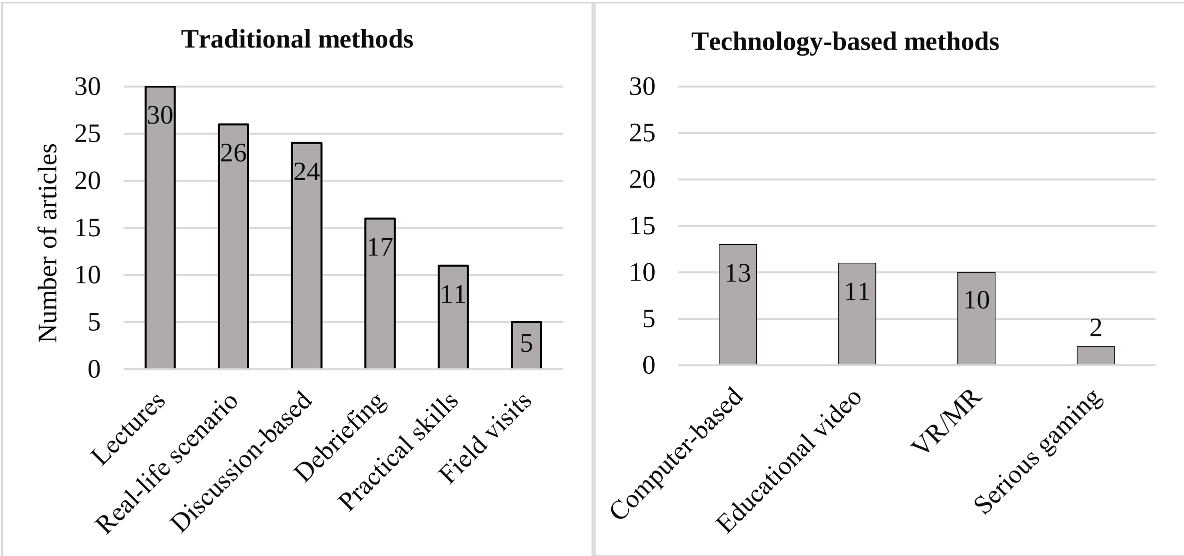
Saiboon et al. (2021)	Single group, pre-post	168	Medical students	General disaster management	Educational video	0.5h, 7 days	Basic knowledge test	Yes
Scott et al. (2010)	Single group, pre-post	61	Medical students	General disaster management	Lecture, real-life scenario	3h, 1 day	Basic knowledge test Self-rep. knowledge	Yes (but only descriptive statistics)
Sena et al. (2021)	Single group, pre-post	22	Physicians	General disaster management	Lecture, discussion-based training, debriefing	2h, 1 day	Basic knowledge test	No
							Self-rep. confidence	Yes
							Perceived importance of disaster medicine training	No
Smith et al. (2015)	Single group, pre-post	65	Nursing students	Nursing leadership skills in disaster response	Lecture, discussion-based training, real-life scenario, VR on screen with movement tracking via webcam	8h, 1 day	Self-rep. self-efficacy	Yes
Unver et al. (2018)	Single group, pre-post	87	Nursing students	General disaster management	Lecture, real-life scenario, debriefing	NA, 8 weeks	Self-rep. disaster preparedness	Yes
Wiese et al. (2021)	Non-RCT (pre-post), cross-over design	80	Nursing students	General disaster management	IG: computer-based learning CG: practical skills, debriefing	NA, NA	Basic knowledge test	IG better than CG
							Perceptions about learning	No sign. difference between groups
Xia et al. (2020)	RCT (pre-post)	63	Nursing students	General disaster management	IG: lecture, discussion-based training, real-life scenario, educational video, debriefing CG: No intervention	7h, NA	Basic knowledge test	Partially confirmed
							Self-rep. attitude (one score combining attitudes towards the training, towards disaster preparedness and family disaster preparation)	No sign. difference between groups
Yanagawa et al. (2018)	Non-RCT (only post)	63	EMTs	General disaster management	IG: Lecture, real-life scenario, practical skills CG: No intervention	NA, 1 day	Observed performance of whole team during real-life simulation (accuracy + overall performance score assessed with a self-developed instrument)	No

Zhang et al. (2021)	RCT (pre-post)	120	Nurses	General disaster management	IG: VR on screen, discussion-based training, real-life scenario CG: Lecture, discussion-based training, real-life scenario, practical skills	48h, 3 months	Basic knowledge test	IG better than CG; CG improved as well but within-group effects not tested for significance
							Observed performance during real-life simulation (overall performance score assessed with the self-developed emergency care capability rating scale)	
							Observed performance in technical skills (overall performance score assessed with a self-developed instrument)	
							Self-rep. disaster preparedness	
Zheng et al. (2020)	RCT (pre-post)	103	Medical students	Triage	IG: Lecture, discussion-based training CG: Lecture	NA, 3 weeks	Basic knowledge test	IG better than CG
							Observed triage performance (accuracy + time) during real-life simulation	No sign. difference between groups
							Satisfaction with training	IG partially better than CG

Note. [1]Number of participants in analyses. [2]Only referring to relevant subsample as there were separate analyses reported. [3]number refers to trainees from the years 2010-2012. Sign. = significant, CG = control group, IG = intervention group, HMD = head-mounted display; Self-rep. = self-reported.

Figure 2.2

Overview of the distribution of traditional and technology-based training methods. Note that several methods can be combined in a single article.

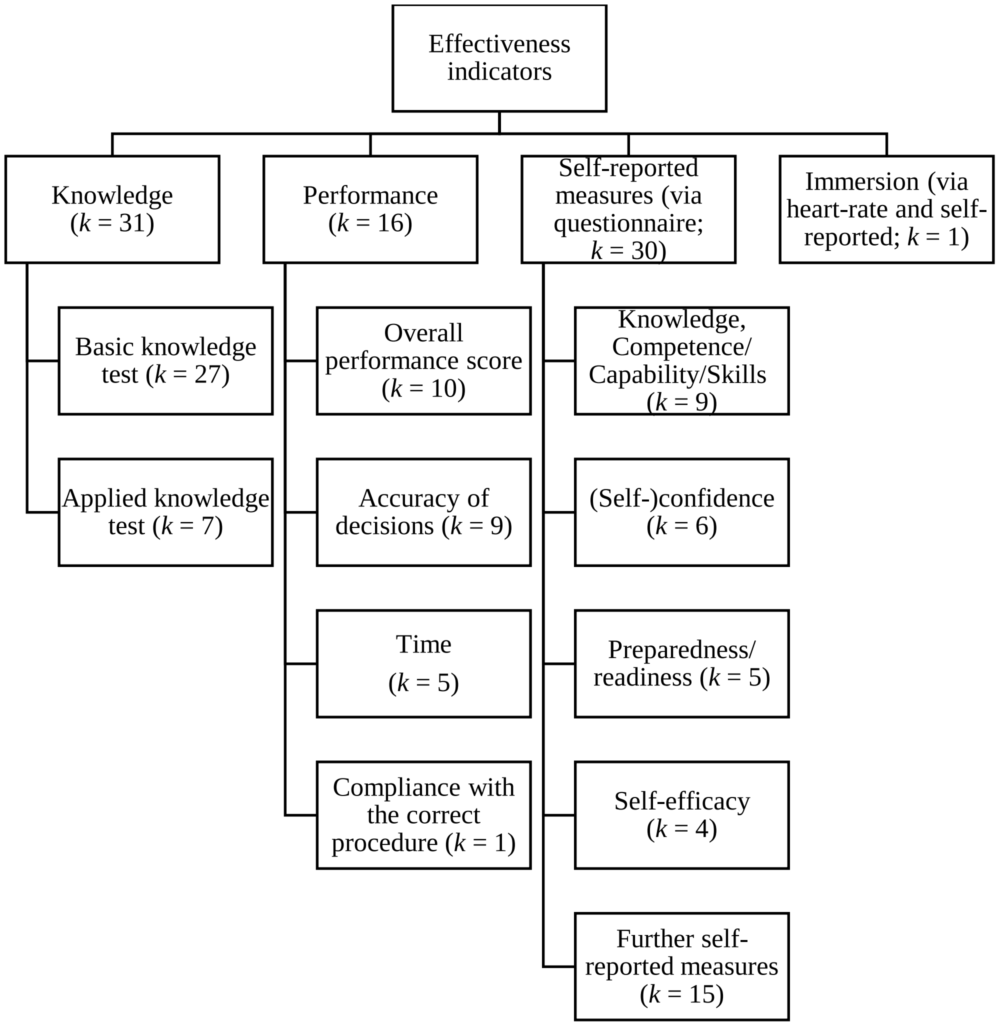


2.4.2 Research question 2: Effectiveness indicators

The trainings were evaluated with several effectiveness indicators, including knowledge and performance, but also self-reported measures (Figure 2.3).

Figure 2.3

Effectiveness indicators with the number of articles that used them.



Note. Note that several indicators can be combined in a single article.

Most frequently, knowledge gain was used as an indicator. Knowledge was mainly assessed with a basic knowledge test on the training content, often in a multiple-choice format. Some studies used an applied knowledge test that consisted of a written test with several patient descriptions which had to be classified into triage categories. Less than one third of the studies used performance as an indicator. Performance assessments were frequently conducted in triage simulations (Andreatta et al., 2010; Cicero et al., 2012; Cicero et al., 2017; Dittmar et al., 2018; Ingrassia et al., 2015; Knight et al., 2010; Mills et al., 2020; Yanagawa et al., 2018; Zheng et al., 2020) but also in in other contexts, e.g., the management of patients affected by chemical, biological, radiological, nuclear and/or explosive events (CBNRE; Andreatta et al., 2015; Motola et al., 2015) and the execution of specific medical procedures (Paddock et al., 2015; Pouraghaei et al., 2017). Several of those studies focused on measures that could be determined easily and relatively well objectively, including accuracy of triage or treatment decisions (Cicero et al., 2012; Cicero et al., 2017; Dittmar et al., 2018; Ingrassia et al., 2015; Knight et al., 2010; Koutitas et al., 2021; Mills et al., 2020; Yanagawa et al., 2018; Zheng et al., 2020), time needed (Dittmar et al., 2018; Ingrassia et al., 2015; Knight et al., 2010; Koutitas et al., 2021; Zheng et al., 2020) or compliance with the correct procedure (Knight et al., 2010). In ten studies, raters composed an overall performance score based on several criteria (Andreatta et al., 2015; Andreatta et al., 2010; Dittmar et al., 2018; Koutitas et al., 2021; Motola et al., 2015; Paddock et al., 2015; Pouraghaei et al., 2017; Ripoll-Gallardo et al., 2020; Yanagawa et al., 2018; Zhang et al., 2021), e.g., the evaluation of safety on site (Andreatta et al., 2010; Yanagawa et al., 2018) and airway/breathing interventions (Andreatta et al., 2010; Dittmar et al., 2018). Three of those studies used already existing assessment instruments, either for treating CBNRE patients (Andreatta et al., 2015; Motola et al., 2015) or for single patient care in low-resource countries (Ripoll-Gallardo et al., 2020). Only one study used team performance as a measure of effectiveness by letting raters compose an overall team performance score for the management of simulated disaster scenes (Yanagawa et al., 2018). All other studies measured individual performance only. Furthermore, several studies used self-reported measures, including preparedness/readiness (Huh & Kang, 2019; Jones et al., 2014; Koca & Arkan, 2020; Unver et al., 2018; Zhang et al., 2021) and (self-) confidence (Betka et al., 2021; Cowling et al., 2021; Furseth et al., 2016; Greco et al., 2019; Kuhls et al., 2017; Sena et al., 2021). In addition to knowledge, performance and self-reported measures, one study compared the level of immersion in VR to real-life scenario training (Mills et al., 2020).

2.4.3 Research question 3: Effectiveness of training methods

All training methods demonstrated certain effectiveness, as most studies reported positive or at least partially positive effects of the different methods (see Table 2.2 for an overview of the methods' effectiveness).

Table 2.2

Effectiveness of methods.

Method	**Indicator**	**Confirmed**	**Partially confirmed**	**No effect found**
Lectures	Knowledge	**Cuttance et al. (2017),** Pouraghaei et al. (2017), Huh and Kang (2019), Scott et al. (2010), Bajow et al. (2016), Pollard et al. (2015), Rivkind et al. (2015)	Chou et al. (2021), Xia et al. (2020), Hutchinson et al. (2011)	Chandra et al. (2014), Sena et al. (2021), Foronda et al. (2016), Lampi et al. (2013)
	Performance	**Dittmar et al. (2018),** Cicero et al. (2012), Ingrassia et al. (2015), Pouraghaei et al. (2017)		Yanagawa et al. (2018)
	Self-rep. preparedness	Huh and Kang (2019), Jones et al. (2014), Unver et al. (2018)		
	Self-rep. knowledge and competence	Chan et al. (2010), Scott et al. (2010), Chandra et al. (2014), Merlin et al. (2010)		
	Self-rep. confidence	Sena et al. (2021)		
	Self-rep. self-efficacy	Smith et al. (2015)		
	Further self-rep. measures	Jones et al. (2014), Merlin et al. (2010), Kim and Lee (2020)		Chou et al. (2021), Sena et al. (2021), Xia et al. (2020)

Method	Indicator	Confirmed	Partially confirmed	No effect found
Real-life scenario training	Knowledge	Andreatta et al. (2015), Bajow et al. (2016), Ripoll-Gallardo et al. (2020), Scott et al. (2010), Rivkind et al. (2015), Pollard et al. (2015)	Chou et al. (2021), Xia et al. (2020), Hutchinson et al. (2011)	Cowling et al. (2021)
	Performance	Andreatta et al. (2015), Cicero et al. (2012), Cicero et al. (2017), Ripoll-Gallardo et al. (2020), Ingrassia et al. (2015)		Yanagawa et al. (2018)
	Self-rep. preparedness	Jones et al. (2014), Unver et al. (2018)		
	Self-rep. knowledge and competence	Cowling et al. (2021), Scott et al. (2010)		
	Self-rep. confidence	Cowling et al. (2021), Greco et al. (2019)		
	Self-rep. self-efficacy	Andreatta et al. (2015), Smith et al. (2015)		
	Further self-rep. measures	Andreatta et al. (2015), Jones et al. (2014), Greco et al. (2019), James et al. (2021), Kim and Lee (2020)		Ripoll-Gallardo et al. (2020), Xia et al. (2020), Chou et al. (2021)
Discus-sion-based learning	Knowledge	Aluisio et al. (2016), Andreatta et al. (2015), Ingrassia, Ragazzoni, et al. (2014), Huh and Kang (2019), Pouraghaei et al. (2017), Bajow et al. (2016), Ripoll-Gallardo et al. (2020), Rivkind et al. (2015), Zheng et al. (2020)	Chou et al. (2021), Xia et al. (2020), Hutchinson et al. (2011)	Chandra et al. (2014), Sena et al. (2021), Lampi et al. (2013)
	Performance	Andreatta et al. (2015), Pouraghaei et al. (2017), Ripoll-Gallardo et al. (2020)		
	Self-rep. preparedness	Huh and Kang (2019)		
	Self-rep. knowledge and competence	**Lennquist Montán et al. (2015), Phattharapornjaroen et al. (2020)**, Chan et al. (2010), Chandra et al. (2014)	Betka et al. (2021)	
	Self-rep. confidence	Sena et al. (2021)	Betka et al. (2021)	
	Self-rep. self-efficacy	Andreatta et al. (2015), Smith et al. (2015)		
	Further self-rep. measures	Andreatta et al. (2015), Kim and Lee (2020)		Ripoll-Gallardo et al. (2020), Sena et al. (2021), Chou et al. (2021), Xia et al. (2020)

Method	Indicator	Confirmed	Partially confirmed	No effect found
Practical skills training	Knowledge	Aluisio et al. (2016), Paddock et al. (2015), Rivkind et al. (2015), Huh and Kang (2019)		Lampi et al. (2013)
	Performance			Yanagawa et al. (2018)
	Self-rep. preparedness	Huh and Kang (2019)		
	Self-rep. knowledge and competence	Merlin et al. (2010), Chan et al. (2010)		
	Self-rep. confidence			
	Self-rep. self-efficacy			
	Further self-rep. measures	Merlin et al. (2010)		
Field visit	Knowledge	Bajow et al. (2016), Ripoll-Gallardo et al. (2020)		
	Performance	Ripoll-Gallardo et al. (2020)		
	Self-rep. preparedness			
	Self-rep. knowledge and competence	Merlin et al. (2010), Chan et al. (2010)		
	Self-rep. confidence			
	Self-rep. self-efficacy			
	Further self-rep. measures	Merlin et al. (2010)		Ripoll-Gallardo et al. (2020)
Debrief-ing	Knowledge	Cicero et al. (2012), Cicero et al. (2017), Bajow et al. (2016), Ingrassia, Ragazzoni, et al. (2014), Rivkind et al. (2015)	Xia et al. (2020)	Cowling et al. (2021), Sena et al. (2021), Foronda et al. (2016)
	Performance			
	Self-rep. preparedness	Unver et al. (2018)		
	Self-rep. knowledge and competence	Cowling et al. (2021)		
	Self-rep. confidence	Cowling et al. (2021), Sena et al. (2021), Greco et al. (2019)		
	Self-rep. self-efficacy			
	Further self-rep. measures	Greco et al. (2019), Kim and Lee (2020), James et al. (2021)	**Fernandez-Pacheco et al. (2017)**	Xia et al. (2020)
Com-puter-based learning	Knowledge	Andreatta et al. (2015), Paddock et al. (2015), Ripoll-Gallardo et al. (2020), Ingrassia, Ragazzoni, et al. (2014), Pollard et al. (2015), Rivkind et al. (2015)	Edinger et al. (2019), Hutchinson et al. (2011)	

Method	Indicator	Confirmed	Partially confirmed	No effect found
	Performance	Cicero et al. (2017), Andreatta et al. (2015), Ripoll-Gallardo et al. (2020)		
	Self-rep. preparedness	**Koca and Arkan (2020)**		
	Self-rep. knowledge and competence		Betka et al. (2021)	
	Self-rep. confidence		Betka et al. (2021)	
	Self-rep. self-efficacy	**Koca and Arkan (2020), Edinger et al. (2019)**, Andreatta et al. (2015)		
	Further self-rep. measures	Andreatta et al. (2015)		Ripoll-Gallardo et al. (2020)
Educa-tional videos	Knowledge	**Motola et al. (2015), Saiboon et al. (2021),** Huh and Kang (2019), Bajow et al. (2016), Ripoll-Gallardo et al. (2020), Paddock et al. (2015), Rivkind et al. (2015)	Xia et al. (2020), Hutchinson et al. (2011)	Chandra et al. (2014)
	Performance	Ripoll-Gallardo et al. (2020)	**Motola et al. (2015)**	
	Self-rep. preparedness	Huh and Kang (2019)		
	Self-rep. knowledge and competence	Chandra et al. (2014)		
	Self-rep. confidence			
	Self-rep. self-efficacy			
	Further self-rep. measures			Ripoll-Gallardo et al. (2020), Xia et al. (2020)
VR (on screen)	Knowledge	Bajow et al. (2016)		Foronda et al. (2016)
	Performance	Ingrassia et al. (2015)		
	Self-rep. preparedness			
	Self-rep. knowledge and competence			
	Self-rep. confidence			
	Self-rep. self-efficacy	Smith et al. (2015)		
	Further self-rep. measures			

Note. Bold numbers[2] indicate that the study tested the method as a sole training method; Self-rep. = self-reported; *Self-reported preparedness* also covers readiness; *Self-reported knowledge and competence* also covers self-reported capability and skills; Note that the table only lists studies that test pre-post comparisons of a training, while between-method comparisons are described in the text; The categories *serious gaming* and *immersive VR* are not part of this table because the studies in which they were used focused on between-method comparisons.

[2] The officially published article used a citation style with numbers. For consistency, the chapters of this dissertation use the APA 7 citation style.

Lectures were mostly used in combination with other methods and often served the initial theoretical knowledge transfer (Chandra et al., 2014; Chou et al., 2021; Kim & Lee, 2020; Merlin et al., 2010; Scott et al., 2010; Sena et al., 2021; Xia et al., 2020). There were three studies in which only lectures occurred between the pre-test and post-test. Two of these evaluated educational refresher sessions and reported a positive impact on knowledge (Cuttance et al., 2017) and performance (Dittmar et al., 2018). The third one concluded that lectures led to similar performance but lower knowledge gain and partially lower training satisfaction than the combination of lectures and discussion-based training (Zheng et al., 2020). Multimethod trainings with lectures showed mixed results regarding knowledge and performance but positive effects on self-reports of preparedness, knowledge, competence, confidence, and self-efficacy.

Real-life scenario training was often similarly or less effective compared to technology-based training. Studies that compared real-life scenario training to either educational videos (Aghababaeian et al., 2013) or VR (Mills et al., 2020) reported a partially lower impact of real-life practice on knowledge (Aghababaeian et al., 2013) and similar impacts on performance (Mills et al., 2020) and training satisfaction (Mills et al., 2020). In combination with other methods, the training also resulted in similar (Ingrassia et al., 2015) or slightly lower (Andreatta et al., 2010) performance but greater knowledge gain (Andreatta et al., 2010) than VR training and lower self-reported competence than serious gaming (Ma et al., 2021).

Discussion-based learning was often combined with other methods and resulted in mixed knowledge outcomes but at least partially positive effects on performance and self-reports of preparedness, competence, confidence, and self-efficacy. However, two studies reported smaller performance improvements (Knight et al., 2010) and self-reported competence gain (Ma et al., 2021) than trainings that contained serious gaming.

Practical skills training was never tested as a sole method. Compared to technology-based training, multimethod training with practical skills exercises always resulted in similar or smaller effects. Trainings containing practical skills exercises led to similar (Paddock et al., 2015) or lower (Wiese et al., 2021) knowledge gain as well as similar performance levels (Paddock et al., 2015) and self-reported learning gains (Wiese et al., 2021) than trainings that contained computer-based learning instead. Furthermore, multimethod training with practical skills exercises resulted in lower performance, self-reported preparedness, and self-reported competence than screen-based VR (Zhang et al., 2021) and lower self-reported competence than serious gaming (Ma et al., 2021).

Field visits were part of five trainings and varied considerably in their content and length. Evidence suggests positive effects on knowledge, performance, and self-reports of knowledge and competence. One paper compared a visit of a large ambulance bus to VR and MR training and concluded that the visit was less effective in increasing performance (Koutitas et al., 2021). However, trainees only had one hour in the ambulance bus to practice finding essential objects while the VR and MR group could practice as many times as they wanted within one week (at least three times).

Debriefings were only explicitly tested once. The study used drone videos from a real-life scenario training that the trainees had previously undergone (Fernandez-Pacheco et al., 2017) and partially confirmed a positive effect on (self-)perception. In combination with other methods, debriefings led to positive outcomes on performance as well as on self-reports of knowledge, confidence, and preparedness. There were mixed findings regarding objectively measured knowledge. Furthermore, multimethod training with debriefings led to lower knowledge scores and similar self-reported learning gains than computer-based learning (Wiese et al., 2021) as well as lower self-reported competence than serious gaming (Ma et al., 2021).

Computer-based learning as a stand-alone method or in combination with other methods led to improvement or partial improvement in knowledge, performance, and self-reports of preparedness, competence, and self-efficacy. Computer-based training resulted in greater knowledge gain and similar self-reported learning gains compared to traditional training (Wiese et al., 2021). Computer-based learning also led to similar knowledge and performance improvements as practical skills training, both combined with videos (Paddock et al., 2015).

Educational videos usually led to at least partial knowledge gain and performance improvements as well as a partially greater knowledge gain than real-life scenario training (Aghababaeian et al., 2013). Only one multimethod study did not find an effect on knowledge. Studies also reported positive outcomes on self-reported preparedness and competence.

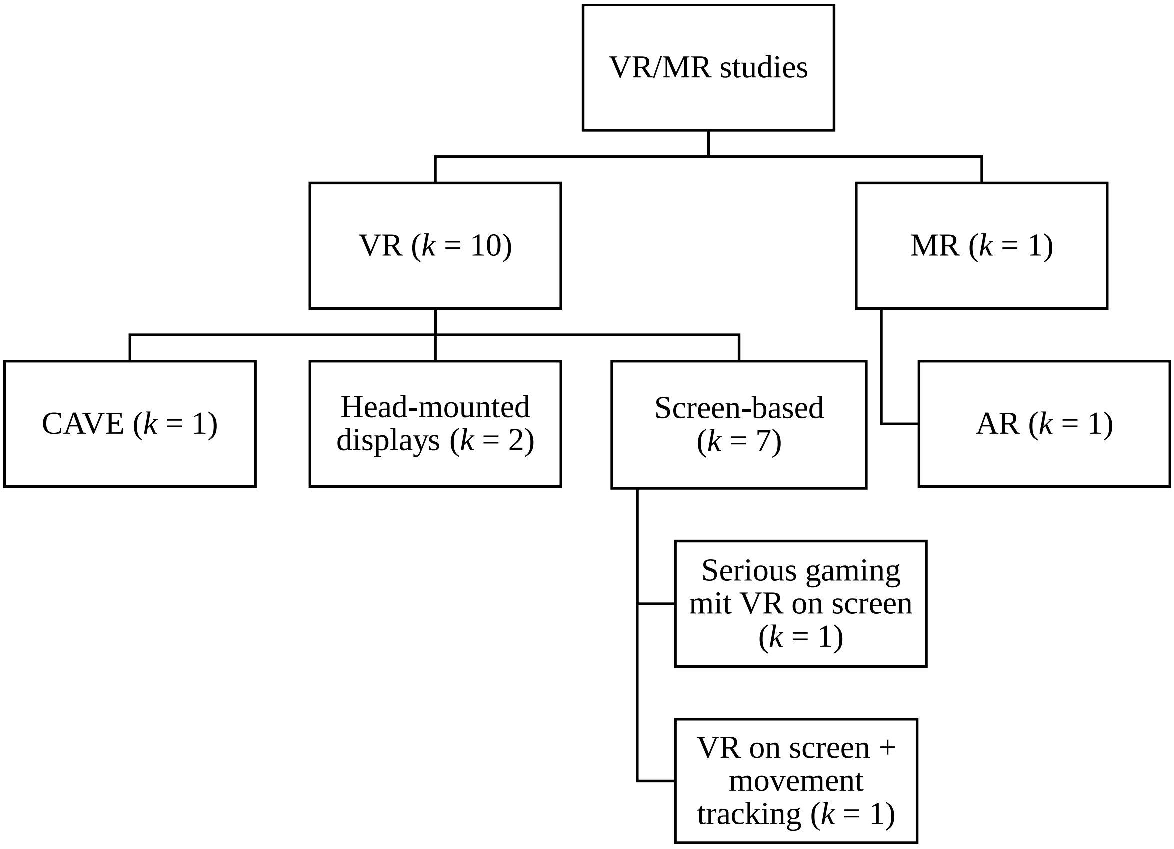
Serious gaming was only evaluated in two studies (Knight et al., 2010; Ma et al., 2021). Ma and colleagues reported that game-based teaching resulted in significantly higher self-reported disaster nursing competence than traditional training (Ma et al., 2021). Knight and colleagues tested a multimethod training including a lecture and serious gaming within VR (Knight et al., 2010). Compared to traditional training, it fostered better triage accuracy and partially better step accuracy. The time needed to triage did not differ between groups.

2.4.4 Research questions 3 and 4: Current role and effectiveness of VR and MR

The VR/MR training systems were mostly used for MFR groups with little or no work experience, including students (Bajow et al., 2016; Farra et al., 2013; Foronda et al., 2016; Ingrassia et al., 2015; Mills et al., 2020; Smith et al., 2015), cadets (Koutitas et al., 2021) or job starters (Andreatta et al., 2010). Seven studies tested trainings that contained PC-screen-based VR (Figure 2.4), although always in combination with other methods (Bajow et al., 2016; Farra et al., 2013; Foronda et al., 2016; Ingrassia et al., 2015; Knight et al., 2010; Smith et al., 2015; Zhang et al., 2021). Five of them covered the topic of triage (Bajow et al., 2016; Farra et al., 2013; Foronda et al., 2016; Ingrassia et al., 2015; Knight et al., 2010), two decontamination (Farra et al., 2013; Smith et al., 2015), one the management of COVID-19 patients (Zhang et al., 2021), and one general disaster scene management (Bajow et al., 2016). The virtual scenarios mainly included manmade disasters such as traffic accidents (Ingrassia et al., 2015), explosions in busy areas (Farra et al., 2013; Knight et al., 2010), building collapse and fire on boats at a seaport (Bajow et al., 2016) while one simulated a major earthquake (Foronda et al., 2016). Two studies that tested pre- and in-hospital trainings used either scenarios in both settings (Zhang et al., 2021) or only an in-hospital scenario (Smith et al., 2015).

Figure 2.4

Overview of VR/MR studies.



During the VR exercises, trainees were able to move their avatar around and perform a variety of intervention, e.g., breathing/airway checks (Foronda et al., 2016; Ingrassia et al., 2015; Knight et al., 2010). While the participants usually used a mouse, keyboard and/or a joystick, one screen-based VR system tracked the trainees' movements with a webcam as they performed decontamination exercises (Smith et al., 2015). Training that contained screen-based VR led to mixed findings regarding knowledge but to positive performance and self-efficacy outcomes. Compared to exclusively traditional trainings, training with screen-based VR led to greater knowledge gain (Zhang et al., 2021) and self-reported preparedness (Zhang et al., 2021) as well as partially greater (Knight et al., 2010; Zhang et al., 2021) or similar performance levels (Ingrassia et al., 2015). Furthermore, the combination with computer-based learning led to greater knowledge gain than computer-based learning alone (Farra et al., 2013).

Three studies evaluated immersive VR technology (Andreatta et al., 2010; Koutitas et al., 2021; Mills et al., 2020). The first one evaluated triage training in a VR CAVE (Andreatta et al., 2010). The scenario was an explosion in an office building. To perform the triage, trainees observed virtual patients to assess their respiratory rate and verbally requested pulse rates. In terms of training effectiveness, the VR exercise resulted in slightly better performance but poorer knowledge scores than real-life scenario training, both in combination with lectures. Two studies evaluated VR training with head-mounted displays in which trainees used controllers to interact with their virtual surroundings (Koutitas et al., 2021; Mills et al., 2020). The first one tested a triage training with a car chase and shooting scenario (Mills et al., 2020). Participants could click on icons attached to each casualty to gather basic clinical information and allocate triage cards. The other VR training was designed to help MFRs get a better orientation in a large ambulance bus by practicing to find essential medical equipment (Koutitas et al., 2021). Both VR systems provided feedback regarding the correctness and time of task execution. Overall, these two VR trainings with head-mounted displays led to similar (Mills et al., 2020) or greater (Koutitas et al., 2021) performance than traditional training and to a similar learning satisfaction (Mills et al., 2020). One of those studies, however, indicated a higher immersion level during real-life simulations which seemed to be caused by the subscale physical demand (Mills et al., 2020). The only study using MR compared AR training to VR with head-mounted displays and traditional education (Koutitas et al., 2021). The AR training was completely similar to the VR ambulance bus training except for the use of an AR headset with transparent lenses. The device projected holograms in the trainees' field of view. With click gestures, they were able to interact with their environment, like opening/closing drawers. The AR training resulted in a better performance than traditional training, but not as much as the VR training.

2.4.5 Quality assessment

Overall, the study quality was satisfactory (for a detailed overview see Supplementary Tables 2.2 and 2.3, Additional File). For the experimental studies, either none (k=9) or one question (k=5) out of 13 were answered with *no*. For the quasi-experimental studies, usually none (k=4), one question (k=23) or two questions (k=13) were answered with *no*. There was only one paper for which four out of nine questions were answered in the negative (Sena et al., 2021). The higher risk of bias in the quasi-experimental studies was mainly based on question 4, which assesses the control group because a large part of the studies had a single group pre-post design (k=35). Furthermore, some studies did not have a complete follow-up or a detailed explanation or analysis for the dropout (k=9).

2.5 Discussion

Well-trained MFRs are essential for managing disaster situations with multiple casualties (Alharbi, 2018; Chaput et al., 2007). To ensure that future disaster training is as effective as possible, we conducted this review on scientifically-evaluated trainings which comprised both traditional and technology-based methods. The trainings were evaluated with several different effectiveness indicators, including knowledge, performance, self-reported measures, and immersion. Despite the heterogeneity of methods and outcome measures, some conclusions could be synergized. While all methods demonstrated effectiveness, the results of this review suggest that technology-based methods often lead to similar or greater training outcomes than exclusively traditional training. Furthermore, we found ten studies that used VR, although usually combined with other methods and often PC-screen-based. Only one study evaluated MR training (Koutitas et al., 2021).

Although trends in effectiveness could be identified, the data basis was not sufficient to declare some methods as unequivocally more effective than others. Training methods were often tested in combination, which impaired drawing unbiased conclusions about individual methods. Furthermore, the various effectiveness indicators that were used had only limited comparability. Fewer than one-third of the included studies used performance observation as an evaluation tool. Instead, several studies used knowledge tests or self-assessments (e.g., confidence) although these have limited predictive value for actual performance (Barnsley et al., 2004; Liaw et al., 2012; Rodgers et al., 2010). Despite the great variety in studies, the data basis strongly suggests the strength of technology-based methods. Several studies compared technology-based training to training with real-life scenario exercises which are usually considered the gold standard of disaster training (Tin et al., 2021). While these studies suggest the great potential of technology-based methods, there may be a certain degree of bias. Real-life scenario training often served as (part of) the exclusively traditional training for control groups. Therefore, studies may not have been published that did not find at least an equivalent effect of their newly developed technological methods. Instead, the training technology might have been improved and retested until it was similarly or more effective, leading to a publication bias. The same might apply to practical skills training which was always used in combination with other methods and resulted in similar or lower training effectiveness than trainings that contained technology-based methods.

Generally, the current literature indicates that technology-based methods are well suited to train MFRs for disasters. Given the usually limited resources of MFR organizations, these methods promise to be particularly beneficial. Although initial investment in the technology is required, it can then be used flexibly and repeatedly. Thus, a higher, more individually adapted training frequency can be created than with many traditional methods, especially real-life scenario training.

2.5.1 Current use of VR/MR and its future potential

Seven out of ten studies that tested VR training focused on non-immersive, screen-based VR. The advantage of screen-based VR is that usually no hardware other than normal computer accessories is required. However, more immersive trainings offer greater similarity to experiencing real disaster situations and could therefore be even more useful for preparing MFRs for stressful and unfamiliar situations. Given that high stress can affect the performance of MFRs, training should explicitly address stress responses (Ignacio et al., 2016; LeBlanc, 2009). Although some of the reviewed trainings contained in-class teaching about dealing with emotions or stress (e.g., Chan et al., 2010; Kim & Lee, 2020; Zhang et al., 2021), we found no studies that explicitly conducted scenario-based training while assessing and controlling for stress responses. To provide more insight into behavioral changes under stress, future studies should conduct and evaluate explicit disaster training with (continuous) stress measurements to investigate its potential for MFRs. The ongoing improvement of immersive VR and MR technology (Anthes et al., 2016) seems quite promising as it can provide increasingly realistic immersive training scenarios with fewer organizational demands than real-life simulations regarding time and space. Users can experience and practice an almost unlimited number of scenarios in which demands and difficulty levels can be designed as needed (Düking et al., 2018). Our results indicate that practical exercises with immersive technology can be conducted nearly everywhere, at any time, and with relatively little preparation, i.e., without setting up a real disaster scene. Furthermore, technical progress in recent years now allows several people to interact within the same virtual environment (Nguyen & Bednarz, 2020) and treat patients together as in realistic rescue operations.

2.5.2 Future research

Given the heterogeneity of the current literature, future research should further investigate the effectiveness of individual training methods but also systematically assess whether certain combinations work particularly well. Furthermore, training methods and validated training evaluation tools should be developed not only in terms of effectiveness, but also in terms of efficiency as (financial) resources are often limited. The results of this review suggest, for example, that technological methods such as serious gaming and VR are similarly good or better than traditional methods so that complex real-life scenario trainings with actors could be at least partially replaced. There is also initial evidence that lectures, as an easily implemented method, are well suited for refresher sessions. Future research still needs to clarify the usefulness of immersive VR and especially MR as we only found one MR experimental study that matched our inclusion criteria.

The effectiveness-efficiency trade-off also applies to training evaluation. While knowledge tests offer the advantage of being very easy to conduct and evaluate, the transferability of training success to actual operations is unclear. Performance evaluations during (virtual or real-life) scenario training may be more suitable as they are closer to the target behavior of MFRs during disasters. This review has already identified some indicators, including accuracy of decisions, time needed and compliance with the correct procedure. Future research should focus on finding the appropriate performance measures for diverse disaster training contents in terms of resource efficiency, usability, and relevance. New training technologies could also provide further opportunities for performance assessment, e.g., eye-tracking to gain insights into attentional processes. Furthermore, the assessment of team performance has hardly been considered in disaster training research, although MFRs mainly work in teams. Disaster management is a team effort and is often done in ad-hoc teams similar to other domains of acute care medicine (Manser, 2009). Improved and trained teamwork improves medical performance (McEwan et al., 2017). Future studies should also assess long-term benefits of the different training methods and their combination as most of the studies we found only conducted pre-post testing within a few days or weeks.

2.5.3 Limitations

Our review has three main limitations. First, we only included studies published in English so we might have missed relevant studies published in other languages. Second, we only kept studies in which it was either evident that the sample only consisted of MFRs or in which separate analyses for MFRs were provided. This led to the exclusion of some studies with insufficiently specified sample categories such as *others*. However, it might be possible that the participants were also MFRs. Third, we decided to include only quasi-experimental and experimental studies. We consider this a strength of this systematic review, as it allowed us to create a better overview of the trainings' effectiveness. Nevertheless, we cannot draw conclusions about what training methods are generally used in disaster training research and whether new methods have been added without being tested in (quasi-)experiments.

2.5.4 Conclusion

We found several traditional and technology-based trainings methods. The trainings were mainly evaluated with knowledge tests and self-reported measures, while less than one third also used actual performance measures. For valid and yet inexpensive evaluations, objectively assessible performance measures, such as accuracy, time, and order of certain actions can be used. In this review, we found that technology-based methods were often similarly or more effective than traditional training. They therefore offer great potential to supplement or at least partially replace traditional training as especially the organization of the gold-standard, real-life scenario training, can be costly and time-consuming. Two training technologies that have become increasingly popular and affordable are VR and MR. This review suggests that they have great potential which is why further assessments of these technologies are required.

Chapter 3. Mixed Reality and Real-Life Exercises for Mass Casualty Incidents: A Comparison of Psychological Responses and Learning

Baetzner, A. S. [1], Gerwann, S. [1], Uhl, J. C. [2,3], Schuster, B.[4], Sjöberg, D.[5], Wenighofer, R.[6], *Frenkel, M. O.[7], *Wrzus, C.[8] (*shared last authorship)

[1]Institute of Sports and Sports Sciences, Heidelberg University, Germany
[2]AIT Austrian Institute of Technology, Austria
[3]University of Salzburg, Austria
[4]Johanniter Research and Innovation Centre, Austria
[5]Unit of Police Work, Umeå University, Umeå, Sweden
[6]Chair of Subsurface Engineering, Montanuniversität Leoben, Austria
[7]Hochschule Furtwangen University, Psychology in Health Care, Faculty Health, Safety, Society, Furtwangen, Germany
[8]Psychological Institute and Network Aging Research, Heidelberg University, Germany

This chapter presents an earlier version of the article, which was revised during the peer review process and differs from the final published version. The Version of Record of this article is published in the journal Educational Technology Research and Development, and is available online at https://doi.org/10.1007/s11423-025-10517-7.

3.1 Abstract

Well-prepared medical first responders (MFRs) are indispensable for effectively managing mass casualty incidents (MCIs). Still, the gold standard for training, high-fidelity real-life exercises (RLEs), is infrequently implemented due to high organizational effort and costs. Mixed reality (MR), where MFRs train in a virtual environment with haptic feedback from manikins, may be a viable training alternative. This study aimed to explore strengths, limitations, and potentials for improvement of MR-MCI training compared to two RLEs.

Thirty-four MFRs (M_{age}=29.7, SD_{age}=7.7, 82% male) participated in MR training, 14 MFRs in RLEs (RLE1, akin to MR: n=4, M_{age}=32.0, SD_{age}=9.5; RLE2, near-ideal: n=14, M_{age}=26.9, SD_{age}=6.7; 100% male). Stress, exhaustion, self-efficacy, presence, and perceived learning gain were assessed using questionnaires and analyzed descriptively. Participants further answered open-ended questions about perceived opportunities and limitations of virtual training. The MR and RLE groups reported similar stress, exhaustion, and self-efficacy levels. The MR group reported slightly lower physical presence but considerably lower social presence compared to the RLE groups. Perceived learning gains were moderate for MR participants and high for RLE participants. Qualitative data indicated a need to improve interaction opportunities with virtual patients. Also, participants viewed virtual training as a resource-efficient supplement, not a replacement for RLEs. Future studies should explore which content and groups benefit most from MR and further evaluate it through larger, experimental studies. MR-MCI training shows promise in preparing MFRs for MCIs and seems to be a valuable addition to RLEs, with the potential to increase training frequency and practice scenarios otherwise difficult to simulate.

Keywords: mixed reality; triage training; emergency medicine; paramedics; stress; learning

3.2 Introduction

Medical first responders (MFRs) play an important role in the management of mass casualty incidents (MCIs), and their effective preparation for such situations is critical to patient survival and public safety in general (Uhl, Schrom-Feiertag, et al., 2023; Zechner, Uhl, et al., 2024). Real-life exercises (RLEs) with patient actors and manikins are considered the gold standard for preparing MFRs for these rare, yet usually chaotic and demanding situations (Tin et al., 2021). Due to their high organizational and financial costs, however, RLEs cannot be conducted frequently (Mills et al., 2020). While traditional MCI training methods, such as lectures and table-top exercises (Baetzner et al., 2022), are less resource-intensive, they do not adequately represent the complexity, which a realistic simulation of an MCI can provide (Berndt, Wessel, Willer, et al., 2018). To fill this gap, immersive virtual reality (iVR) and mixed reality (MR) have emerged as new training solutions that could help increase the frequency of simulation training (Baetzner et al., 2022).

iVR and MR give trainees the impression of being surrounded by a virtual world in which they can move and interact with their virtual environment (Milgram & Kishino, 1994). To create the impression of ‘actually being there’ (i.e., presence; Slater, 2018), iVR/MR systems must provide a certain level of immersion, i.e., objectively measurable system properties leading to a multimodal stimulation of the perceptual system (Gerwann et al., 2024; Slater, 2018). While iVR often allows interaction with the environment via controllers, MR involves additional senses beyond sight and hearing, such as haptic feedback (Gerwann et al., 2024; Milgram & Kishino, 1994).

Immersive virtual training may be particularly beneficial for MCI training, as it allows for procedural training under stress (Voigt & Frenkel, 2023; Wrzus, Schöne, et al., 2024) without endangering real patients (Zechner, Uhl, et al., 2024). Well-designed simulation training should be realistic enough to create a sense of presence (Berndt, Wessel, Willer, et al., 2018), and elicit responses similar to the real situation to enable learning transfer (Hutter et al., 2023). As MCIs are not only physically but also psychologically demanding for MFRs (Hugelius et al., 2020), effective simulation training should aim to induce exhaustion and stress. While research on MR MCI training is still scarce (Baetzner et al., 2022), iVR MCI training has been shown to successfully induce a sense of presence and stress reactions (Prachyabrued et al., 2019; Servotte et al., 2020). In addition, iVR MCI training leads to improved self-efficacy to successfully manage MCI situations (Thompson, 2023; Vincent et al., 2008) and perceived learning gain (Bilek et al., 2021). iVR training, however, lacks certain aspects of realism due to the absence of physical contact with patients and in some cases the use of unnatural movement like teleportation (Mills et al., 2020; Prachyabrued et al., 2019), which can result in reduced physical demands (Mills et al., 2020). The development of MR systems with tangible patients using integrated manikins is still relatively new but demonstrated success in first aid and CPR training (Ricci et al., 2022; Uhl et al., 2024). MR MCI training with integrated manikins, a large training area for natural movement and the possibility of team training could further bridge the gap between iVR and real situations, thereby improving MFR training.

As a first step towards a holistic understanding of the potential of MR, this study specifically focused on the trainees' individual experiences. To identify strengths, limitations, and potential improvements, this study aimed to compare an MR MCI system with the gold standard, an RLE with built-up scenery and patient actors. For this purpose, we compared stress, exhaustion, and presence as key metrics of representative learning experiences, and MCI-related self-efficacy and perceived learning gain as indicators of perceived training effectiveness. In addition to the comparison with a similarly built RLE, the MR training was also compared to an 'optimal' RLE, including a larger number of trainees and cooperation with other emergency services. We thus contribute to the understanding of which training modalities should be used for different training goals and illustrate key future developments needed in MR training to fulfill the vision of immersive, safe, but realistic virtual training.

3.3 Methods

The ethics committee of the faculty of behavioral and empirical cultural sciences at Heidelberg University approved the study (AZ Beu 2023 1/1). This study was part of the EU project MED1stMR (www.med1stmr.eu) which focused on the development of MR training for MFRs. Although the MR training was performed in six European countries, we only used data from the same country, where the RLEs were executed, to ensure comparability.

3.3.1 Participants

Thirty-six MFRs took part in the MR training. Two of them did not fill out the pre-questionnaire, resulting in a sample size of *n*=34. Fourteen participants from the same MFR organization took part in the RLEs, *n*=4 in RLE1 and *n*=14 in RLE2. For further information on the groups, see Table 3.1[3]. One person participated in the MR training and both RLEs.

[3] Please note that in this dissertation, chapter numbers have been added to numbers of tables, figures, and supplementary files for clearer identification.

Table 3.1

Description of groups.

	Mixed reality (*n*=34)	Real-life exercise 1 (*n*=4)	Real-life exercise 2 (*n*=14)
Age in years			
Mean (SD)	29.7 (7.73)	32.0 (9.45)	26.9 (6.74)
Gender			
Male	28 (82.4%)	4 (100%)	14 (100%)
Female	6 (17.6%)	0 (0%)	0 (0%)
Profession			
Emergency doctor	1 (2.9%)	0 (0%)	0 (0%)
Paramedic	32 (94.1%)	3 (75%)	13 (92.9%)
Paramedic + medical student	1 (2.9%)	1 (25%)	1 (7.1%)
Job experience			
Mean (SD)	10.0 (6.24)	12.5 (9.68)	7.37 (6.40)
MCI training experience			
0-10 hrs	11 (32.3%)	2 (50%)	4 (28.6%)
10-20 hrs	5 (14.7%)	0 (0%)	1 (7.1%)
20-30 hrs	4 (11.8%)	0 (0%)	2 (14.3%)
30-40 hrs	4 (11.8%)	1 (25.0%)	2 (14.3%)
40-50 hrs	0 (0%)	0 (0%)	1 (7.1%)
more than 50 hrs	10 (29.4%)	1 (25.0%)	3 (21.4%)
Missing	0 (0%)	0 (0%)	1 (7.1%)
Experience with MCIs (number of incidents)			
Mean (SD)	5.62 (9.53)	2.75 (2.63)	2.43 (3.65)
Experience with MCIs in hours			
Mean (SD)	3.50 (2.51)	3.00 (1.63)	2.93 (2.50)

3.3.2 Procedure

All training was conducted in Austria, the MR training in July 2023 and the RLEs in September 2023. Prior to the training, participants received a link to an online questionnaire, polling demographics and person characteristics. Trainees also completed baseline stress, exhaustion, and self-efficacy questionnaires on site. In both the MR and RLE modalities, participants were tasked with managing the MCI situation and performing the first triage as the first MFRs on site. Based on a user-centered approach, scenarios were built and tested together with MFR organizations.

Mixed reality training

Teams of three to four MFRs participated in the MR training (Figure 3.1a-d). Wireless head-mounted displays (HTC Vive Focus 3) allowed for free movement within a 10x10m training. Body trackers in the head-mounted displays, on the back, hands, and feet were used to effectively capture pose and movement of the participants with up to 40 cameras, attached to a truss built around the training area. First, participants were placed in a virtual waiting room to familiarize themselves with the MR system (e.g., the use of triage cards and tourniquets). The training session consisted of two scenarios, both involving a bus crash with 21 casualties: an MCI in a motorway tunnel (average duration: 09:24 min) and a similar MCI on a country road (average duration: 09:12 min). Both scenarios were taken into account to approximate the duration of (complex) RLEs.

The virtual environments depicted a damaged bus and other vehicles nearby. Patients were spread out outside the bus, either lying on the ground, sitting or walking around (Figure 3.1c-d). Environmental noises included patient sounds, such as cries for help, crying, and moaning, as well as background noises, e.g., engine sounds. Patients displayed various visible injuries, ranging from minor blood wounds to head injuries and severed body parts. Patient movements were matched to their injuries. Patients with minor injuries (triage color green) walked to a designated safe spot after receiving a green triage card or reacted to verbally being sent off by an MFR and a digital trigger being activated by the trainer. As an MR feature, two virtual patients were mapped onto manikins with a tangible pulse and chest movement corresponding to their respiration rate. Post-questionnaires were administered immediately after each scenario (stress, exhaustion), and a final questionnaire (self-efficacy, perceived learning gain, presence, and open questions) after the entire training session. For more information on the development of the system and characteristics of the MR training, see Zechner, Uhl, et al. (2024).

Figure 3.1

MR setup and RLEs.

Notes. MR setup (a) and manikin (b), tunnel scenario from above (c) and inside scenario (d), RLE1 (e) and RLE2 (f).

Real-life training

The real-life training consisted of two exercises in a motorway tunnel dedicated to research and training (Zentrum am Berg, www.zab.at). Both RLEs were set up as bus accidents with damaged vehicles, patient actors and manikins. A close resemblance between the MR and RLE environments was achieved by building the MR tunnel to closely mirror the real one, including a door to a security area within the tunnel. Patient actors underwent detailed theatrical makeup application and vocalized distress signals to achieve realistic appearances (Figure 3.1e-f).

The first exercise (RLE1, Figure 3.1e) was similar to the MR tunnel scenario described above, including damaged vehicles and a bus. Because the MR training was mainly conducted with teams of four MFRs, only four MFRs took part in the RLE1 (duration: 16:35min). To maintain a high similarity with the MR training, the casualty area in RLE1 was of comparable size and no non-medical emergency personnel were involved in the scenario. In contrast, RLE2 was set up closer to a typical RLE, with all 14 trainees taking part simultaneously and coordinating with members of the fire brigade and police, who also participated (40:43 min, Figure 3.1f). The RLE2 included forty actors performing as patients (see Figure 3.1e-f). Additionally, RLE2 extended beyond RLE1 by incorporating arrival and initial build-up, as well as treatment and transportation of patients, thereby presenting a more comprehensive exercise. After each RLE, participants were asked about their stress and exhaustion levels, their self-efficacy, perceived learning gain, and feeling of presence. After both RLEs, participants answered open-ended questions.

3.3.3 Data and Measures

Demographics included age, gender, profession, years of job experience in the medical sector, experience with MCIs (number of incidents and total hours), and MCI training experience in hours.

Experience with iVR was assessed with the question "How much prior experience do you have with VR?" (9 options from "no experience" to "daily use of VR").

Subjective stress and exhaustion were measured with one item each: "Right now I feel stressed" (Giessing et al., 2020) and "Right now I feel exhausted" with a scale from 1 (not at all) to 7 (very). For the MR group, post values of the two scenarios were averaged (Spearman-Brown coefficient for stress = .79, for exhaustion = .61).

MCI-self-efficacy was assessed with five items based on Vincent et al. (2008) on a scale from 1 = "not at all" to 5 "very" (e.g., "I feel confident that I am an effective first responder"; Cronbach's Alpha[4] in the MR group = .91, in the RLE group = .91).

Presence was assessed separately for physical presence (i.e., the illusion of actually being in the simulated place) and for social presence (i.e., the illusion of being with others, even if they are only virtual/actors and manikins). Items were based on the multimodal presence scale (MPS), adapted for iVR research (Makransky et al., 2017). Four items were used for physical (e.g., "I was completely captivated by the virtual world"; Cronbach's alpha for both groups: .79) and two items for social presence (e.g., "I had a sense that I was interacting with other people in the virtual environment, rather than a computer simulation"; Spearman-Brown coefficient for MR group = .69, for RLE group =.36). For the RLEs, the items were rephrased to fit the context, e.g. "I was completely captivated by the staged scenario" (scale: 1=strongly disagree, 5=strongly agree).

[4] Please note that variables from RLE1 were not included in the calculation of Cronbach's Alpha due to the small sample size and items without variance in that group.

Perceived learning gain was measured with 7 items. They cover perceived learning gains in risk awareness, systematic care/triage, systematic examination, situational awareness, mental stress, communication, and decision-making (e.g. "How much have you learned today about recognizing risks at the scene of an incident?"; scale: 1="nothing at all", 4="quite a lot"). Item values were summed, leading to possible values between 7 and 28 (Cronbach's Alpha in MR group = .77, in RLE group = .75).

Open-ended questions

To gain further insight into the differences between the evaluated training modalities and what aspects could be improved in iVR/MR training, participants were asked about their thoughts on iVR/MR training compared to RLEs, and what opportunities and concerns they perceived regarding iVR/MR training. The responses from the RLEs provide an insight into the attitudes of MFRs, who have participated in usual training so far. These themes could be addressed during a potential MR training implementation process. Responses were either audio-recorded (RLEs group) or written in a free text field (MR group). Supplementary File 1[5] provides the exact questionnaire items and further information on the qualitative data assessment.

3.3.4 Analyses

Quantitative data was analyzed using the software R (Version 4.3.2). Due to the small sample size (n_{MR} = 34, n_{RLE1} = 4, n_{RLE2} = 14), results are reported descriptively. Extreme values were defined as values 3 standard deviations above or below the mean. To prevent these values from excessively influencing the outcome variables without excluding them, they were

[5]As stated at the end of the original manuscript, supplementary material and data are available online: https://osf.io/2f764/?view_only=a2154c9b844748cd943631eb817e16a2

winsorized, i.e., set to mean +/- 3 standard deviations (see Supplementary File 2[5] for R script and Supplementary File 3 for openly accessible, quantitative data).

Qualitative data was analyzed with the software MAXQDA. Data was analyzed inductively and based on Mayring's qualitative content analysis (Mayring, 2022) to ensure a systematic coding and categorization process (see Supplementary File 1[5]). Responses were coded inductively and categories formed on the basis of content proximity. Qualitative data from *n*=10 participants were analyzed for the RLE group due to incomplete responses from some participants.

3.4 Results

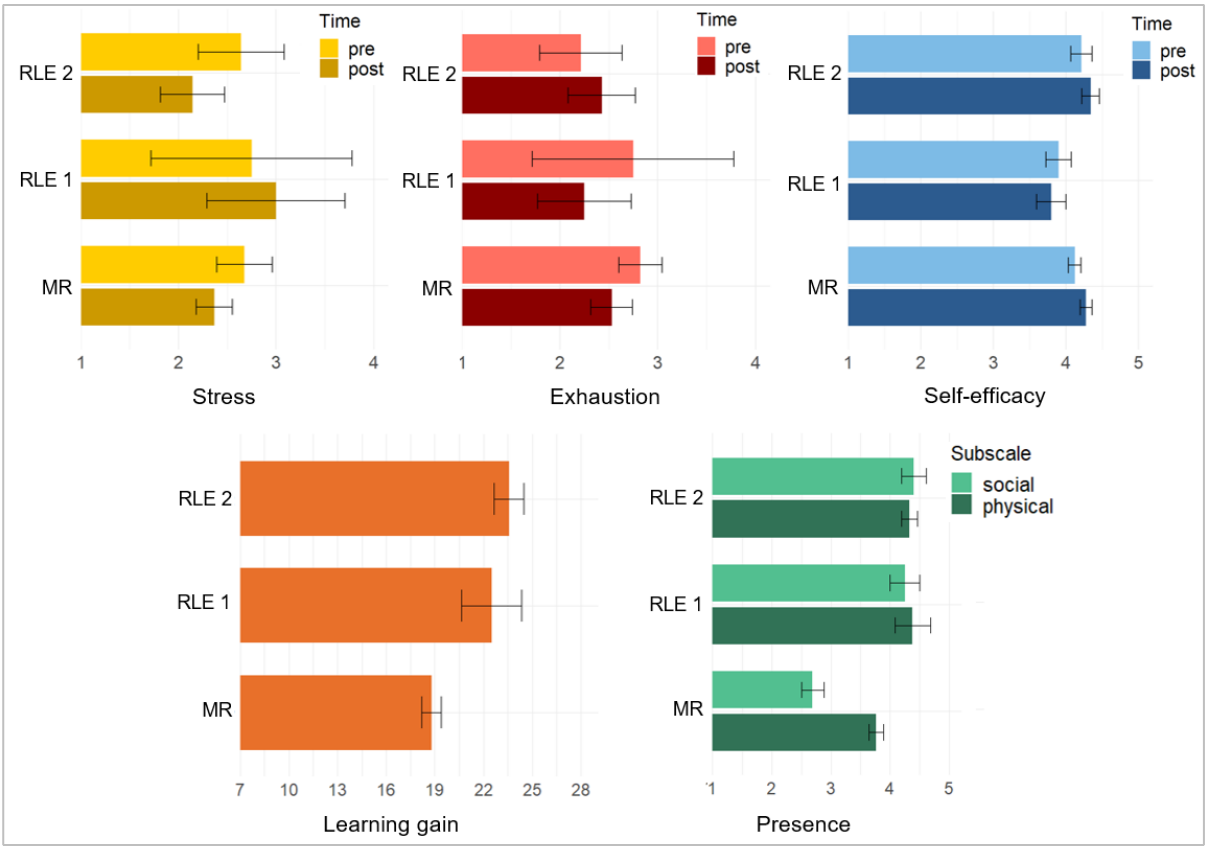
Demographic information about the groups is presented in Table 3.1. Regarding the MR group, 13 (38.2%) had no prior experience with iVR, ten had tried it 1-2 times, ten had tried it several times and one person used iVR once per month. In RLE1, participants had either tried iVR 1-2 times (n=2) or several times (n=1; n=1 missing answer). Half of the RLE2 participants had no experience with iVR (n=7), four had tried it 1-2 times (28.6%), and two several times (14.3%; n=1 missing answer).

3.4.1 Quantitative results

Figure 3.2 shows the descriptive results of the outcome variables per group with standard errors. Table 3.2 contains the means and standard deviations of all outcome variables.

Figure 3.2

Comparison of the groups' training experiences.



Stress and exhaustion

In all three groups, participants reported similar and generally low average stress levels after the training (between 2 and 3; scale: 1-7). Only the RLE1 group reported higher stress levels after (*M*=3.00) the training than before (*M*=2.75). In the MR and RLE2 groups, participants reported slightly greater stress levels before than after the training.

Average exhaustion values were also all between 2 and 3, i.e., in the lower half of the scale during both the MR and the RLEs. For the MR and RLE1 groups, mean exhaustion was slightly lower before than after the training (MR: from 2.82 to 2.53, RLE1: from 2.75 to 2.25), while for the RLE2 group, mean exhaustion was higher after (*M*=2.21) the training than before (*M*=2.43).

Table 3.2

Means and standard deviations of outcome variables per group.

	Mixed reality (*n*=34)	Real-life exercise 1 (*n*=4)	Real-life exercise 2 (*n*=14)
Stress pre	2.68 (1.66)	2.75 (2.06)	2.64 (1.65)
Stress post	2.37 (1.08)	3.00 (1.41)	2.14 (1.23)
Exhaustion pre	2.82 (1.29)	2.75 (2.06)	2.21 (1.58)
Exhaustion post	2.53 (1.25)	2.25 (0.96)	2.43 (1.28)
MCI self-efficacy pre	4.12 (0.50)	3.90 (0.35)	4.21 (0.55)
MCI self-efficacy post	4.28 (0.47)	3.80 (0.40)	4.34 (0.45)
Perceived learning gain	18.8 (3.52)	22.5 (3.70)	23.6 (3.41)
Physical presence	3.76 (0.71)	4.38 (0.60)	4.32 (0.51)
Social presence	2.69 (1.09)	4.25 (0.50)	4.39 (0.78)

Self-efficacy

All three groups had already reported high self-efficacy values prior to the training (between 3.90 and 4.21 on a scale of 1-5). While self-efficacy was higher after the training in the MR and RLE2 groups, the RLE1 group reported slightly lower self-efficacy after the training.

Presence

All three groups tended to perceive a high physical presence. The RLE1 (*M*=4.38) and RLE2 group (*M*=4.32) reported a greater physical presence than the MR group (*M*=3.76). In terms of social presence, the RLE1 (*M*=4.25) and RLE2 (*M*=4.39) groups again reported high mean scores, while the MR group reported lower social presence (*M*=2.69).

Perceived learning gain

After the training, the RLE2 group reported the greatest perceived learning progress (M=23.60 with possible values between 7 and 28), closely followed by the RLE1 group (M=22.50). The MR group reported medium perceived learning gains (M=18.80).

3.4.2 Qualitative results

The results are organized according to the inductively determined categories, and the number of participants per group, who made the statements, are reported in brackets. Supplementary File 1 provides example statements and the corresponding categories.

Resource requirements (n_{RLE}=3, n_{MR}=24)

Both groups expected lower resource requirements of iVR/MR training, particularly in organizational (n_{RLE}= 2, n_{MR}=16) and personnel effort (n_{RLE}=2, n_{MR}=8). Especially the MR group valued the flexibility in creating various and complex scenarios (n_{MR}=7), including those difficult to implement in RLEs (n_{RLE}=1; n_{MR}=3). While current costs were considered high (n_{MR}=4), iVR/MR was expected to become economically advantageous in the long term (n_{RLE}=2, n_{MR}=1) or when shared by multiple regions (n_{MR}=1).

Stress and mental burden (n_{RLE}=2; n_{MR}=6)

Both groups considered stress induction essential in simulation training and believed iVR/MR to be less stressful or mentally demanding than RLEs or real operations (n_{RLE}=2; n_{MR}=6), mainly due to lower realism of virtual patients compared to real humans.

Presence and immersion (n_{RLE}=8, n_{MR}=10)

The RLE group suspected some constraints in realism in virtual training compared to RLEs (n=6), particularly regarding interactions with virtual patients (n=3). While one person appreciated the potential for good visualization in immersive virtual training (*n*=1), others assumed that the absence of certain sensory aspects, such as the feeling of wind, was disadvantageous compared to RLEs (*n*=3). Overall, the RLE group perceived the experience of presence to be more difficult to achieve in virtual environments than in RLEs (*n*=5), although imagination skills might be required for both training modalities (*n*=1).

In contrast, the MR group did not mention concerns about presence and found iVR/MR immersive (*n*=2). They considered iVR/MR both more realistic (*n*=3) and less realistic than RLEs (*n*=6), the latter due to lower perceived realism in interaction and communication with virtual patients (*n*=3).

Learning (n_{RLE}=4, n_{MR}=19)

The RLE group believed virtual training could increase training frequency and thus help to develop a routine (*n*=2). There were assumptions that virtual training could be useful for leadership training (*n*=1) and for medical, but not organizational training (*n*=1).

Similarly, the MR group valued iVR/MR for its potential to increase training frequency (*n*=2), repeat scenarios (*n*=3), and offer built-in debriefing capabilities through recordings (*n*=5). Concerns were raised about its usability for older trainees, those with poor eyesight or those prone to motion sickness (each *n*=1). The MR group found iVR/MR valuable for inexperienced MFRs (*n*=3) but were more critical of the technologies' usefulness for leadership training (*n*=3 critical vs. *n*=1 in favor).

Attitude towards iVR/MR (n_{RLE}=9, n_{MR}=20)

Although the RLE group saw potential in virtual training (*n*=3), they believed that RLEs are better (*n*=5; *n*=2 neutral) and cannot be replaced by virtual training (*n*=2). The MR group also saw potential in iVR/MR training (*n*=9) and viewed it as a good complement (*n*=6), not a replacement for RLEs (*n*=7).

3.5 Discussion

This multi-method study explored the strengths, limitations, and potential improvements of MR MCI training compared to RLEs, with both modalities designed for team training in a high-fidelity setting. Descriptive coefficients indicated that the MR group reported similar mean stress, exhaustion and self-efficacy scores as the RLE groups. Regarding presence, the MR group experienced only slightly lower physical presence indicating a high realism of the virtual environments. For social presence, however, the differences were considerably larger, suggesting that currently social interactions in the virtual environment were perceived as less realistic than interactions with patient actors. Perhaps as a consequence, the MR group reported medium average perceived learning gains, while the RLE groups reported high gains.

Qualitative data highlighted practical advantages of iVR/MR, including reduced organizational and personnel requirements, greater flexibility in scenario creation, and potentially reduced future costs compared to RLEs. The RLE group who had previously received ‘conventional’ training, expressed concern that virtual training might offer less realism and presence, which aligns with the quantitative findings on (social) presence. The MR group had mixed views on the realism of iVR/MR scenarios, as they perceived the interaction with virtual patients as less realistic compared to with patient actors. Both groups mentioned positive aspects of iVR/MR for learning, such as increased training frequency and scenario repetition. However, the ideal contents and target groups of iVR/MR training were less clear. The MR group suggested it could be useful for inexperienced MFRs. Both groups saw potential in iVR/MR and suggested the technologies as a complement to RLEs rather than a replacement.

This study was the first to compare MCI training in MR to RLEs. Previous research comparing iVR MCI training to RLEs found no significant differences in various variables like subjective performance, effort, and satisfaction (Mills et al., 2020). Lower physical demand was assumed to occur due to participants standing stationary in iVR (Mills et al., 2020). To address this, the current study allowed for natural movement. Participants in previous iVR MCI research criticized the lack of haptic feedback (Servotte et al., 2020). A recent comparison of a single patient scenario in RLE, iVR, and MR demonstrated that MR was closer to RLEs than iVR in terms of presence and acceptance like facilitating conditions (i.e., intuitive and easy use) (Uhl et al., 2024).

Consistent with prior iVR research, social presence was lower in MR than in RLEs and participants wished for more and better interaction opportunities with virtual patients (Servotte et al., 2020; Uhl et al., 2024). In the current MR system, interactions included obtaining vital parameters through haptic feedback or hand movements, applying triage cards and tourniquets, and directing lightly injured patients to green-triaged zones. However, trainees did not receive verbal responses during interactions. Although recent work has shown that virtual training environments can elicit similar levels of social presence, even if the interaction is highly scripted (Uhl, Neundlinger, et al., 2023), the pre-recorded voice lines of the MR simulation did not reach real-life levels of social presence. Furthermore, a recent study has shown that real-time transcription and artificial intelligence can enable communication in a single-patient MR system (Uhl et al., 2024), suggesting that this feature could be extended to multiple patients in an MCI training scenario in the future.

For all three training groups, differences between pre- and post-measures of stress were small. While the pre-questionnaire stress levels could be due to anticipatory stress, the survey wording may also have influenced results, as participants were asked about their current stress level rather than during the scenario. By the time of answering post-questionnaires, stress levels may have already dropped.

3.5.1 Limitations and future research

In addition to the mentioned technical limitations, the small sample sizes due to demanding RLEs, led to only descriptive analyses of quantitative data. While this study provides valuable insights into current strengths and areas for improvement in MR training, future research could incorporate the suggested improvements and test the updated MR training system with larger sample sizes. Furthermore, future studies should especially concentrate on improving MFR-patient interaction and perceived stress, and experimentally test the systems. Testing objective learning outcomes of different modalities could help to determine which training content is best suited for each modality.

In addition, a combination of psychological, physiological and behavioral data could provide a more holistic understanding of the benefits and challenges of MR MCI training. Physiological stress parameters such as heart rate, heart rate variability, and electrodermal activity could be tracked in real-time during the training (e.g., Prachyabrued et al., 2019). In terms of behavioral data, performance assessments would be valuable for the evaluation of training effectiveness, comparisons between training methods, and the identification of individual training needs (Baetzner et al., 2025; Wakasugi et al., 2009). For instance, recently validated performance indicators for immersive virtual MCI training include triage accuracy, triage speed, and efficient information transmission to the control center (Baetzner et al., 2025). Furthermore, the evaluation of teamwork parameters could be beneficial. The tested MR system allowed simultaneous training of up to four MFRs. Trainees could interact by viewing each other's avatars and gestures, and by communicating verbally. While this study focused on individual experiences, future research investigating team dynamics and performance in MR and RLEs could provide further insights for the training of teamwork.

3.5.2 Practical implications

First responder training has a slowly evolving and meticulously developed curriculum. As with many aspects of disaster response and critical care, any change to the current standard must demonstrate a clear and compelling added value. MR training solutions meet this criterion by enabling significantly higher training frequencies than RLEs and great flexibility in scenario design and complexity levels. The technology can be integrated seamlessly into existing curricula, potentially in a phased plan, starting with an initial supplement to partial replacement of those RLE scenarios that are exceedingly hard or impossible to implement in routine institutional training. It could also be used to reduce the gap between theoretical classroom teaching and large-scale RLEs (Baetzner et al., 2025), with the possibility of creating individual training plans of increasing difficulty.

3.5.3 Conclusion

MR MCI training showed the potential to achieve similar levels of perceived stress, exhaustion, and self-efficacy, as well as nearly similar levels of physical presence as RLEs. Social presence and perceived learning gain could, however, benefit from further enhancement. Potential improvements include the ability to adjust scenario complexity for stress elicitation and the implementation of communication capabilities with virtual patients. Overall, MR MCI training shows promise in preparing MFRs for MCIs. As a complement to traditional MCI training, MR has the potential to increase training frequency and practice scenarios otherwise difficult to simulate.

Supplementary Material

Supplementary material and the data can be retrieved from the following link:

https://osf.io/2f764/?view_only=a2154c9b844748cd943631eb817e16a2

Chapter 4. Mass Casualty Incident Training in Immersive Virtual Reality: Quasi-Experimental Evaluation of Multimethod Performance Indicators

Baetzner[1*], Hill[2,3,4], Roszipal[5], Gerwann[1], Beutel[1], Birrenbach[6], Karlseder[5], Mohr[7,8], Salg[7,9], Schrom-Feiertag[10], Frenkel[1,11], Wrzus[12]

[1]Institute of Sports and Sports Sciences, Heidelberg University, Germany
[2]Department of Human Movement Sciences, Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, the Netherlands
[3]Institute of Brain and Behaviour Amsterdam, Amsterdam, the Netherlands
[4]Lyda Hill Institute for Human Resilience, University of Colorado Colorado Springs, Colorado Springs, USA
[5]Mindconsole GmbH, Austria
[6]Department of Emergency Medicine, Inselspital, University Hospital, University of Bern, Switzerland
[7]Medical Faculty, Heidelberg University, Germany
[8]Department of Anesthesiology, University Hospital Heidelberg, Germany
[9]General-, Visceral- and Transplantation Surgery, University Hospital Heidelberg, Germany
[10]Center for Technology Experience, Austrian Institute of Technology GmbH, Austria
[11]Furtwangen University, Psychology in Health Care, Faculty Health, Safety, Society
[12]Psychological Institute and Network Aging Research, Heidelberg University, Germany

This chapter may not be an exact copy of record of the officially published paper. The original reference can be found here:

Baetzner, A. S., Hill, Y., Roszipal, B., Gerwann, S., Beutel, M., Birrenbach, T., Karlseder, M., Mohr, S., Salg, G., Schrom-Feiertag, H., Frenkel, M. O., Wrzus, C. (2025). Mass Casualty Incident Training in Immersive Virtual Reality: Quasi-Experimental Evaluation of Multimethod Performance Indicators. *Journal of Medical Internet Research (JMIR).* http://dx.doi.org/10.2196/63241

4.1 Abstract

Background: Immersive virtual reality (iVR) has emerged as a training method to prepare medical first responders (MFRs) for mass casualty incidents (MCIs) and disasters in a resource-efficient, flexible, and safe manner. However, systematic evaluations and validations of potential performance indicators for virtual MCI training are still lacking. This study aimed to investigate whether different performance indicators based on visual attention, triage performance, and information transmission can be effectively extended to MCI training in iVR by testing if they can discriminate between different levels of expertise. Furthermore, the study examined the extent to which such objective indicators correlate with subjective performance assessments.

Methods: Seventy-six participants (age M = 25.54 years, SD = 6.01, 59% male) with different medical expertise (MFRs: paramedics and emergency physicians; non-MFRs: medical students, in-hospital nurses, and other physicians) participated in five virtual MCI scenarios of varying complexity in a randomized order. Tasks involved assessing the situation, triaging virtual patients, and transmitting relevant information to a control center. Performance indicators included eye-tracking based attention, triage accuracy, triage speed, information transmission efficiency, and self-assessment of performance. Expertise was determined based on the occupational group (39 MFRs vs. 37 non-MFRs) and a knowledge test with patient vignettes.

Results: Triage accuracy (d = 0.48), triage speed (d = 0.42), and information transmission efficiency (d = 1.13) differentiated significantly between MFRs and non-MFRs. Additionally, higher triage accuracy was significantly associated with higher triage knowledge test scores (Spearman’s rho = .40). Eye-tracking based attention was not significantly associated with expertise. Furthermore, subjective performance was not correlated with any other performance indicator.

Conclusion: iVR-based MCI scenarios proved to be a valuable tool for assessing the performance of MFRs. The results suggest that iVR could be integrated into current MCI training curricula to provide frequent, objective, and potentially (partly) automated performance assessments in a controlled environment. In particular, performance indicators such as triage accuracy, triage speed, and information transmission efficiency capture multiple aspects of performance and are recommended for integration. While the examined visual attention indicators did not function as valid performance indicators in this study, future research could further explore visual attention in MCI training and examine other indicators such as holistic gaze patterns. Overall, the results underscore the importance of integrating objective indicators to enhance trainers' feedback and provide trainees with guidance on evaluating and reflecting on their own performance.

Keywords: pre-hospital decision-making; disaster medicine; emergency medicine; mass casualty incident; medical education; eye tracking; emergency simulation; virtual reality

4.2 Introduction

4.2.1 Overview

Medical first responders (MFRs) are confronted with extreme demands at mass casualty incidents (MCIs), during which they must attend to more patients than their resources allow (Lomaglio et al., 2020). Whether those situations are due to natural disasters, accidents, or terrorist attacks, peak performance of MFRs under such demanding circumstances is crucial. However, current MCI training is insufficient to prepare MFRs adequately for MCIs due to low immersion and realism of classroom learning and scarce real-life exercises that are typically used (Almukhlifi et al., 2021; Berndt, Wessel, Willer, et al., 2018; Mills et al., 2020).

An increasingly popular tool for MCI training is immersive virtual reality (iVR; (Baetzner et al., 2022; Berndt, Wessel, Willer, et al., 2018). iVR is typically experienced through a head-mounted display (HMD) that elicits the user's impression of being completely surrounded by a 3D, virtual world (Gerwann et al., 2024; Milgram & Kishino, 1994; Wrzus, Schöne, et al., 2024). In addition, users are often able to move around, explore, and interact with their environment (e.g., by walking and using controllers). Especially in the context of MCI training, iVR offers the possibility to train numerous scenarios in a safe, flexible, and resource-efficient way (Baetzner et al., 2022). iVR training applications also allow for high-quality assessments of performance, which are crucial for enabling systematic and structured training (Salas et al., 2009) and for assessing the degree to which MFRs are prepared for MCIs. Moreover, high-quality performance assessments allow for the evaluation of training effectiveness and the comparison of training methods, contributing to an ongoing improvement of MCI preparedness (Markou-Pappas et al., 2024). However, a systematic evaluation of performance indicators for iVR MCI training is still missing. This study aims to fill this gap by evaluating the usefulness of different performance indicators in virtual MCI scenarios. Specifically, we first provide an overview of the potential of iVR for MCI simulation training.

Next, we discuss various performance indicators that cover different aspects of the performance of MFRs who arrive first at an MCI scene (i.e., visual attention, triage accuracy, triage speed, and information transmission). Finally, we introduce self-rated performance as a potential tool for a holistic assessment.

4.2.2 MCI Training and iVR

The gold standard for MCI training is to take part in real-life scenarios with patient actors and manikins (Tin et al., 2021). However, such exercises are particularly resource-intensive in terms of preparation time and costs. Therefore, technological training solutions have been developed to increase the level of preparedness, such as training in iVR (Baetzner et al., 2022).

iVR provides near-realistic training sessions that require fewer personnel and financial resources compared to real-life MCI exercises with patient actors (Mills et al., 2020). iVR also allows the creation of an infinite number of scenarios that can be quickly and flexibly adapted to training objectives (Gerwann et al., 2024). By providing a safe training environment where mistakes can be made and scenarios can be repeated until they are handled correctly (Bowyer et al., 2008; Heldring et al., 2024), iVR can be used in training for otherwise dangerous scenarios, such as fire or terrorist attacks. The training and its challenges can also be adapted to individual needs and expertise levels, thus providing an opportunity to maintain a balance by being complex enough to trigger learning without leading to overload (Gerwann et al., 2024; Hill et al., 2024). On the basis of their current skills, MFRs may be placed in virtual scenarios with few patients and distractions or may train in large-scale MCIs with more patients under complex conditions (e.g., nighttime and difficult terrain). Another advantage of iVR-based training is the ability to implement automated monitoring of objective performance indicators, which can help improve both the trainees’ learning process and their preparedness (Kman et al., 2023). Furthermore, iVR facilitates the inclusion of novel performance indicators, such as visual attention assessed with eye-tracking, which integrates seamlessly with the technology (Bischof et al., 2024).

4.2.3 Performance Indicators

Overview

Performance indicators can be validated with the known-groups validation approach (Shinnick, 2016; Ziegler & Bühner, 2012) which tests the expectation that distinct groups (e.g., age groups, experts and novices) differ on a certain measure (Cronbach & Meehl, 1955). Effective indicators should distinguish between experts and novices, as experts, by definition, possess the knowledge and skills to consistently perform at high levels (Ullén et al., 2016). Previous studies have classified experts and novices based on various measures, including having or not having a specific profession or certificate, often intertwined with (years of) job experience (e.g., medical students vs physicians) (Blondon et al., 2015). In addition, experts generally possess more knowledge than novices (Schmidt & Boshuizen, 1993).

Because performance assessments should capture multiple dimensions of performance (Salas et al., 2009), we evaluate three areas for a broad performance assessment of the first MFRs who arrive at an MCI: (1) the process of gaining an overview of the situation when arriving on scene, which could be assessed based on visual attention, (2) the triage process, and (3) the efficiency of information transmission to the control center. Furthermore, to obtain a holistic assessment of performance, self-reports are also included.

Visual Attention During Orientation at the MCI Scene

Visual attention has been suggested as a potential factor contributing to medical performance (Shinnick, 2016). Consistent with the eye-mind hypothesis, gaze fixations are usually the focus of thought (Just & Carpenter, 1980), which is why the interest in gaze behavior is increasing in research on medical decision-making and medical training (Al-Moteri et al., 2017; Brunyé et al., 2019). Previous research demonstrated that perception and recognition skills improve with increasing medical experience (Al-Moteri et al., 2017). Furthermore, visual attention differs with expertise in several medical tasks, such as surgical simulations and diagnostic decision-making (Al-Moteri et al., 2017; Currie et al., 2019). Among other aspects, experts fixate on task-relevant cues more often and fixate less on task-irrelevant areas than novices (Gegenfurtner et al., 2011).

Only a few studies have tested visual attention in prehospital settings thus far. Eye tracking while viewing photos of accident scenes revealed that MFRs fixated significantly longer on task-relevant cues, such as injured people or safety hazards, than non-MFRs (Prytz et al., 2018). However, the association between visual attention and expertise was less clear when tracking the eye movements of physicians and nurses watching a 2D video of an MCI (Price et al., 2022). Regardless of the number of years working in emergency medical services, all participants spent more time observing the patients who were more severely injured. In addition, participants navigated through the scenes with less predetermined or structured patterns than initially expected. Another study assessed the visual attention of college students viewing a computer screen with photos of injuries and information boxes about airway, breathing, and circulation (Loth et al., 2019). One group was trained in the *simple triage and rapid treatment* (START) triage algorithm, while the active control group was instructed in patient transport. After training, the START triage group fixated on relevant information boxes significantly faster. However, all of these studies relied on 2D stimuli presented on computer screens, and none tested visual attention of MFRs in 3D iVR MCI scenarios, although differences between 2D input and iVR have been documented for many cognitive and behavioral processes (Wrzus, Schöne, et al., 2024).

Accuracy and Speed of the Triage Process

Typical objective performance assessments in MCI training address accuracy and speed of the triage process (Baetzner et al., 2022; Markou-Pappas et al., 2024). Effective allocation of available resources and a well-decided sequence of patient care and transport are essential for managing MCI situations (Mills et al., 2020) and require the ability to triage patients correctly even under high stress. Triage accuracy refers to the correct choice of triage levels based on a specific triage algorithm. The START triage algorithm is an internationally known algorithm developed in the United States and used in several emergency medical services organizations (Super et al., 1994). According to START, patients are classified based on their injuries as green (minor priority), yellow (delayed priority), red (immediate priority), and black (dead or fatally injured). In addition to accuracy, triage speed is also considered crucial in the management of MCI situations so that patients with immediate priority receive medical treatment as quickly as possible (Bazyar et al., 2020). In particular triage accuracy, but also triage speed, have been used as effectiveness indicators for evaluating MCI training in previous studies (Baetzner et al., 2022). Regarding iVR studies, only a few have included objective performance assessments within the virtual environment. These studies suggest that triage accuracy in iVR improves with more iVR training (Vincent et al., 2008). Furthermore, when comparing iVR to real-life exercises, MFRs had similar triage accuracy scores (Ferrandini Price et al., 2018; Mills et al., 2020), but completed the triage process faster in iVR (Mills et al., 2020). However, emergency physicians with more experience did not have higher triage accuracy scores in iVR than those with less experience (Lowe et al., 2020). This may be because the study did not aim to compare experts and novices, and those with less experience were already practicing residents. Therefore, the difference in practical experience was probably not large enough to be considered a typical expert-novice comparison.

Transmission of Information to the Control Center

One of the main problems commonly found during major incidents is poor communication with the control center (Lowes & Cosgrove, 2016; Markou-Pappas et al., 2024). For instance, communication problems such as excessive radio traffic can lead to several negative consequences including increased mortality rates and reduced safety for MFRs at the scene (Perry et al., 2021). Although information transmission to the control center is crucial to ensure proper coordination of all rescue services, studies measuring the quality (Markou-Pappas et al., 2024) and efficiency (i.e., accurate information in few words) of such radio messages are still missing. In Germany, the widely used *scene, safety, situation, and support* (SSSS) *scheme* serves as a standard for the assessment of onsite emergency situations (Schmid et al., 2022). This scheme provides a structured approach to assessing critical aspects of an incident. Sometimes referred to as the 3S scheme, with s*upport* not specifically named, the scheme assists MFRs in identifying and communicating all relevant information, especially potential hazards and environmental risks (Zechner, Schrom-Feiertag, et al., 2024). *Scene* refers to the assessment of the emergency site. In terms of safety, MFRs assess the risk to themselves and others. *Situation* refers to an estimation of the number of patients and evaluation of injury mechanisms. *Support* refers to the possible need for reinforcements and other emergency services, such as police and fire brigades (Schmid et al., 2022). Based on interviews and workshops with European MFRs, communication and the correct use of the scheme were identified to be key performance indicators in training for high-stress situations like MCIs (Zechner, Schrom-Feiertag, et al., 2024). While no studies seem to have evaluated the efficiency of information transmission, it was found that trained physicians reported more accurately than untrained physicians in a real-life MCI exercise (Ingrassia et al., 2013).

Subjective Performance Ratings

A large part of MCI-training studies evaluates the training effectiveness with self-rated indicators, such as self-ratings in knowledge and skills. For example, emergency medical technicians with more work experience rate their triage skills more favorably than those with less work experience (Soola et al., 2022). However, general skill assessment does not necessarily relate to performance in a specific situation and studies assessing self-rated MCI performance directly related to a previously completed simulation are still missing (Baetzner et al., 2022). Not specific to MCI training, but to medical education in general, previous research suggests that self-assessed performance has at least a low to moderate validity (Blanch-Hartigan, 2011; Gordon, 1991). Self-rated indicators have the advantage of being easily implemented in all possible training modalities without any technical effort. Although objective performance assessment can be implemented in iVR, such assessment may be more complex in other training modalities, impairing potential comparison studies. If a global subjective performance indicator is accurate, it may be a low-threshold indicator for such studies. Including self-assessment alongside objective performance evaluation could also enhance the debriefing process by encouraging trainees to reflect on their performance and improve self-awareness (Gow, 2013). However, comparisons between (global) subjective performance and objectively measured behavior in iVR MCI training are still lacking.

4.2.4 Research Aims and Hypotheses

This study aimed to investigate whether different performance indicators that were used in previous MCI training studies can be effectively extended to MCI training in iVR. As a method of validation (Shinnick, 2016), indicators were tested with a focus on their ability to differentiate between different levels of expertise. For comprehensive testing, 2 indicators of expertise were used, including occupational group as a dichotomous indicator (MFR or non-

MFR) and triage knowledge test scores as a continuous indicator. For a broad assessment, we investigated visual attention after arriving on scene, triage accuracy, triage speed, information transmission efficiency, and subjective performance. Overall, this study intends to provide insights into which indicators are suitable for incorporation into the design of effective iVR MCI training programs and MFR performance assessments. The following hypotheses were formulated:

hypothesis 1a: With greater expertise, medical staff pays more attention to task-relevant information during MCIs;
hypothesis 1b: With greater expertise, medical staff show greater triage accuracy during MCIs;
hypothesis 1c: With greater expertise, medical staff complete the triage process faster during MCIs;
hypothesis 1d: With greater expertise, medical staff transmit information more efficiently;
hypothesis 2: Attentional indicators demonstrate incremental value in the discrimination of different levels of expertise beyond performance indicators such as speed, accuracy, and information transmission during triage;
hypothesis 3: Subjective evaluation of one’s own performance is better with greater attention to task-relevant information, greater triage accuracy, faster triage time, and more efficient transmission of information.

The wording of 2 hypotheses has changed from the preregistration. Previously, they were “Hypothesis 1d: with greater expertise, medical staff transmits information more quickly and completely” and “Hypothesis 3: subjective evaluation of own performance is better with greater attention to task-relevant information, greater triage accuracy, faster triage time, and faster transmission of information."

4.3 Methods

4.3.1 Study Design

In this quasi-experimental, multimethod study, participants (MFRs and non-MFRs) completed virtual MCI scenarios while their performance was assessed.

4.3.2 Procedure

Overview

Interested people visited a website following the link or QR code provided on the recruitment materials. On the website, they received information on the study aims and procedure, provided informed consent, and answered a web-based preliminary questionnaire covering demographics, personal characteristics, and a triage knowledge test. The web-based questionnaire was answered before participants came to their appointment to shorten the time in the laboratory. At the end of the questionnaire, participants scheduled a 2-hour appointment in the VR laboratory. In the laboratory, participants filled out additional questionnaires, were shown the START triage algorithm, and engaged in an iVR familiarization scenario with people who were uninjured and at no accident site. During the familiarization, participants practiced information gathering and navigation within the virtual environment. The familiarization ended when participants announced that they feel sufficiently prepared. The main phase of the study involved 5 virtual MCI scenarios to cover the different performance levels of our sample, thus avoiding floor and ceiling effects and increasing the reliability of the measurements. The scenario order was randomized for each participant using a random number generator.

MCI Scenarios in iVR

The scenarios were built with the XVR software (XVR Simulation B.V.) and consisted of traffic accidents that varied in scenario complexity (Table 4.1[6]; Multimedia Appendix 4.1 gives a detailed description). Tasks included gaining an overview of the situation, performing triage, and transferring all relevant information to the control center via radio messages.

Table 4.1

Scenario descriptions. After the familiarization scenario, the scenario order was randomized for each participant.

Scenario	**Scene description**	**Number and triage levels of patients**	**Duration in minutes: mean (SD)**
Familiarization	Quiet residential area, no accident, clear vision, middle of the day; 6 healthy people present, 3 of them sitting in cars	0 patients but healthy people to practice the iVR[a] handling	5.78 (1.71)
1 (very low difficulty)	Accicent with 2 cars on a countryside road, clear vision, middle of the day	3 green patients 1 yellow patients	3.32 (1.03)
2 (low difficulty)	Accident with a car, a van, and a motorcyclist on a motorway, clear vision, cloudy day	1 green patients 3 yellow patients 1 red patients	5.33 (1.37)
3 (medium difficulty)	Accident with a car and an SUV[b] in a busy inner-city area with a crowd of bystanders around, clear vision, cloudy day	1 green patients 3 yellow patients 2 red patients	5.06 (1.59)
4 (greater difficulty)	Express way accident with 3 cars and a van; several bystanders, twilight, end of the day	1 green patients 3 yellow patients 3 red patients 1 black patients	7.28 (1.97)
5 (highest difficulty)	Express way accident involving a bus and a truck, nighttime and fog	4 green patients 6 yellow patients 4 red patients 4 black patients	10.66 (3.08)

Note. [a]iVR: immersive virtual reality. [b]SUV: sports utility vehicle.

[6] Please note that in this dissertation, chapter numbers have been added to numbers of tables, figures, and supplementary files for clearer identification.

Participants were asked to remain in the starting position for the first 30 seconds during each scenario and were allowed to turn and look around. This time frame was used to assess initial attention processes measured via eye tracking. During the scenario, participants could use visual cues like wounds to select triage levels. Further information about the patients could be acquired verbally and was provided with the help of standardized audio tracks. For example, if participants inquired about the respiration rate, they would hear a pre-recorded answer. Obtainable information included whether the patient could walk, whether the airways were clear, respiration rate, recapillarization time, presence of a radial pulse, existence of heavy bleeding, the patient's responsiveness to simple instructions, and localization of injuries (e.g., head injury and leg injury). The participants could announce their intention to perform actions, including clearing the airway, stopping heavy bleeding, and conducting triage. After each scenario, they filled out questionnaires and had the option to reread the START algorithm.

Hardware Description

The Varjo Aero HMD had a display resolution of 2880×2720 pixels per eye at 90 Hz. Its field of view was 115° horizontally and 134° diagonally (at 12mm eye relief), and the gaze data output frequency was 200 Hz. The HMD was connected to a laptop (ROG Strix G with Windows 10, Intel Core i7-9750H central processing unit @ 2.60 Hz 2.59 GHz, NVIDIA GeForce TTX 2070 graphical processing unit, 32 GB of RAM) through the standard HMD cable and Varjo laptop adapter. Participants used the Varjo Aero HMD from Varjo Technologies Oy and could freely move within a 3×4m size area. For greater distances, participants used controllers to teleport (Figure 4.1).

Figure 4.1

Immersive virtual reality (iVR) lab (left) and use of teleportation in iVR (right).

4.3.3 Participants

Emergency services personnel, hospital personnel (physicians and nurses), and medical students (all semesters) were recruited to participate in this study. Inclusion criteria were a minimum age of 18 years, proficiency in German, and the absence of a hearing aid. Participants were recruited through social media, email distribution lists, flyers, and short presentations at local hospitals, emergency medical services, and university courses. Data collection lasted from February to October 2023. Of the 150 participants who filled out the web-based, preliminary questionnaire, 76 (50.7%) scheduled an appointment with our iVR laboratory and participated in the study. Participants were aged between 18 and 49 (mean 25.54, SD 6.01) years, and 59% (n=45) of participants were male. Among the 76 participants, 51% (n=39) were categorized as MFRs (n=36, 92% paramedics and n=3, 8% emergency physicians). The remaining 49% (37/76) non-MFR participants were mainly medical students (n=30, 81%), other medical staff not working in prehospital emergency settings (n=4, 11%), or both at the same time (n=3, 8%).

4.3.4 Measures

Study Data

The data presented in this paper were derived from a larger project, which focused on investigating the stress dynamics and performance of MFRs in virtual MCI training. For detailed information of all measures and instruments used during the study, refer to the construct overview on the web (Wrzus, Baetzner, et al., 2024).

Web-Based, Preliminary Questionnaire

Demographic information included age, gender, profession, years of job experience in the medical sector, and prior MCI training in hours. In the case of profession, participants could choose multiple options from the following: emergency physicians, other physicians, paramedics, emergency medical technicians, medical students, or specify another option through an open text field. Prior MCI training was assessed with the following answering options: none, 1 to 5 hours, 5 to 10 hours, 10 to 20 hours, 20 to 30 hours, 30 to 40 hours, 40 to 50 hours, or >50 hours.

In terms of expertise, participants were classified as MFRs if they were emergency physicians or paramedics; otherwise, they were classified as non-MFRs. The second expertise measure was a triage knowledge test based on Cuttance et al. (2017) and adapted to the START algorithm. Participants had 10 minutes to assign triage levels to 20 case descriptions, after which the test ended, and the next page opened. One point was awarded for each correctly assigned color, resulting in a score range of 0 to 20.

Triage algorithm or algorithms. Participants were asked which triage algorithm they typically used in the field or in training sessions, including the option to select “none.”

Prior experience with iVR was assessed with the item “How much prior experience do you have with VR?” and a dropdown menu with 9 answering options from “no experience” to “daily use of VR.”

Visual Attention During Orientation in Each iVR MCI Scenario (Onsite)

Attention to task-relevant information was assessed using eye tracking in the first 30 seconds of each scenario. Eye-tracking data were collected using Varjo Base and analyzed with the software iMotions (2022). The gaze behavior was analyzed in terms of average fixation durations and number of fixations of the areas of interest (AOIs; Clay et al., 2019; Nguyen et al., 2017). AOIs included *patients, vehicle impact zones, safety aspects,* and a *distractor* (Multimedia Appendix File 4.1 gives scenario descriptions). Within the iMotions software, AOIs were defined graphically, and the software automatically computed the fixation parameter (Farnsworth, 2024).

Objective Triage Accuracy and Speed in Each iVR MCI Scenario (Onsite)

Triage accuracy was measured by the number of correctly assigned triage colors per scenario. The accuracy score for each scenario was divided by the number of patients in that scenario. The accuracy scores of all 5 scenarios were then averaged, leading to possible values between 0 and 1 (standardized Cronbach’s $\alpha = 0.76$).

Speed of triage was measured as the time from the start of the scenario to the completion of triaging the last patient (standardized Cronbach’s $\alpha = 0.82$). Note, however, that due to the different lengths of the scenarios, averaging with non-standardized times would lead to a stronger weighting of the more complex scenarios with longer duration and a larger temporal variance. To avoid this, a standardization was performed: values were divided by the average triage speed of the specific scenario before being averaged. Consequently, values <1 signified a triage process faster than the average, while values >1 denoted a slower-than-average triage pace.

Information Transmission to Control Center in Each iVR MCI Scenario (Onsite)

To assess how quickly and completely medical staff transmit information, an efficiency measure was formed as a combination of completeness and speed. Participants' radio messages during the scenarios were recorded and transcribed according to the content-semantic transcription by Dresing and Pehl (2018). A coding system based on the qualitative content analysis by Mayring (2015) was then applied, with categories derived from the SSSS scheme (Schmid et al., 2022) and added to a coding template. Next, 2 independent raters (university students specifically trained for this) analyzed the transcripts with the software MAXQDA (VERBI GmbH), without access to any information on the participants (interrater agreement: 97% for categorization and 93% for assessment of correctness). Discrepancies were resolved through discussion and, if necessary, together with a third rater (ASB). Statements were categorized, with a maximum of 1 point awarded for each of the 4 categories. Half a point was awarded for s*cene* if participants only mentioned that it was a traffic accident, and a full point if more information was given (e.g., the type of street or number of vehicles involved). Furthermore, a penalty for mistakes was implemented: 0.3 points were deducted for each error in the respective category. Regarding the category s*ituation*, an error was given if the number of patients was not stated correctly. In the 2 most complex scenarios, no penalty was applied if an estimate close to the correct number was reported (+/- 2; e.g., for scenario 4 with 8 patients "almost 10" or "between 7 and 9"; n = 4 cases). No penalty was given for incorrectly stated triage colors, as these are already covered by the triage accuracy variable. The scores of the 4 categories were added up (range: 0–4 points) and standardized based on the number of words in the radio message (i.e., divided by the word count). This standardization accounted for unclear, redundant and superfluous communication (possibly even in the same time by simply speaking faster), which poses a risk to the management of MCIs (Hutchins & Timmons, 2006). Finally, the average score across all 5 scenarios was calculated (standardized Cronbach's α = 0.83; range: 0.03–0.15).

Subjective Performance of Each iVR MCI Scenario (Onsite)

Subjective performance was assessed with 2 items after each scenario: "How do you rate your performance in the last scenario?" (1 = very bad and 10 = very good) and "What school grade does your performance correspond to?" (scale: 1-6, with 1 being the best grade). We used 2 items to increase the reliability in capturing the construct. Before the items were averaged, the second item was inverted and transformed to match the 1 to 10 scale of the first item (Cronbach's α = 0.86):

$$11 - (\left(\frac{Original\ Value - 1}{6 - 1}\right) x\ (10 - 1) + 1)$$

4.3.5 Analyses

Analyses were conducted with R (version 4.3.2; R Core Team). Hypotheses H1a to 1d were tested with both expertise measures separately, distinguishing between MFRs and non-MFRs as well as using the triage knowledge test. For hypothesis 1a specifically, a multivariate ANOVA was used because of the multiple AOI categories. According to the Mardia test, the assumption of multivariate normal distribution was violated, but multivariate ANOVAs are robust to this specific violation, particularly in cases of homogeneous covariance matrices, which was given (Finch, 2006). For hypotheses 1b, 1c, and 1d, independent *t* tests were performed to test for differences between MFRs and non-MFRs. For hypothesis 1b, the Welch *t* test was used because the Levene test indicated a violation of the assumption of homogeneity of variance. Hypotheses 1c and 1d were tested with the student *t* test. According to the Shapiro-

Wilk test, the assumption of normality was violated testing hypothesis 1b and hypothesis 1d. However, *t* tests are largely robust against this violation, especially because both group sample sizes were relatively similar and >30 (Pagano, 2013; Rasch & Guiard, 2004). The Bonferroni-Holm method was used to control for multiple testing in the *t* tests (hypotheses 1b-1d). Spearman rank-order correlations were used to examine the relationship between the knowledge test score and the outcome variables in hypotheses 1a to 1d, as they are robust to distributional violations and outliers. In addition, when correlations were significant, regression analyses were conducted to test the relationship while controlling for age and gender. We selected these 2 control variables because they are potential influencing factors in iVR performance situations (Felnhofer et al., 2012; Parra & Kaplan, 2019). For the regression analyses, independent variables and covariates were centered (for gender: -1 = female and 1 = male). Hypothesis 3 was tested with Spearman correlation analyses.

Outliers were defined as values that were 3 SDs above or below the mean. These values were winsorized (i.e., set to mean +/- 3 SDs, respectively) to reduce the influence of outliers on the results (Ratcliff, 1993). Analyses were conducted with adjusted outliers and with original variables to examine the robustness of results. When technical problems occurred during the assessment (e.g., technical failure of the eye tracking or the microphone for radio message recording), the respective measure was excluded instead of excluding the entire dataset of the participant. Therefore, hypothesis 1a, which addressed eye-tracking data, was tested with a sample of $n = 71$ for the AOIs *patients, safety*, and *vehicle impact zone*, and $n = 57$ for the *distractor*. Hypotheses 1b and 1c, which examined triage behavior, were tested with $n = 74$ and hypothesis1d, which focused on information transmission, was tested with $n = 70$ participants. There were no missing values for subjective performance (hypothesis 3).

4.3.6 Ethical Considerations

This study was preregistered (Baetzner et al., 2023) and approved by the ethics committee of the Faculty of Behavioral and Empirical Cultural Sciences at Heidelberg University (AZ Bae 2023 1/1), and carried out in accordance with the Helsinki Declaration on Ethical Principles of Research Involving Humans. To ensure transparency in the analysis process, the R Markdown file, containing the R code and all results, is available on the web (Wrzus, Baetzner, et al., 2024). Data was collected in a pseudonymized manner and subsequently anonymized for analysis. As an incentive, the participants received €25 (US $27.48 on Feb 1, 2023).

4.4 Results

4.4.1 Analyses

To capture 2 different measures of expertise, analyses were conducted with occupational group as a dichotomous measure (MFRs vs. non-MFRs) and the triage knowledge test scores as a continuous measure. Means and SDs of performance indicators per scenario can be found in Table 4.1 in Multimedia Appendix 4.1.

4.4.2 Demographics

The classification of expertise based on the profession was supported by the results of the knowledge test. MFRs had a significantly higher knowledge score than non-MFRs, $t_{Welch}(66.17) = -3.43$, $p = .001$, $d = 0.79$. More information on the groups can be found in Table 4.2. Most participants were not familiar with any triage system before this study (43/76, 57%; 34/43, 79% of those being non-MFRs). Some (27/76, 36%) participants knew the mSTART (modified START) algorithm, 12% (9/76) knew PRIOR (Primäres Ranking zur Initialen Orientierung im Rettungsdienst), and 1% (1/76) the START algorithm. Of the non-MFR group, 65% (24/37) had no prior MCI training, and 30% (11/37) non-MFRs had <5 hours. 3% (1/37) had up to 20 hours, and 3% (1/37) had up to 30 hours of MCI training. Regarding the MFRs' prior MCI training experience, 8% (3/39) reported none, 31% (12/39) had up to 5 hours, 18% (7/39) had up to 10 hours, 18% (7/39) had up to 20 hours, 5% (2/39) had up to 30 hours, 8% (3/39) had up to 40 hours, 5% (2/39) had up to 50 hours, and 8% (3/39) had >50 hours.

Table 4.2

Description of expertise groups: medical first responders (MFRs) and medical students and other medical staff (non-MFRs).

Characteristics	**MFRs (*n* = 39)**	**Non-MFRs (*n* = 37)**
Age in years (*M*, *SD*)	27.00 (7.15)	24.00 (4.07)
Gender: female, *n* (%)	11 (28)	20 (54)
Job experience in years, mean (SD)	6.53 (7.01)	1.62 (1.94)
Experience with iVR[a], none or only 1-2 times, *n* (%)	33 (85)	34 (92)
Knowledge test score, mean (SD)[b]	13.46 (2.73)	10.89 (3.70)

Note. [a]iVR: immersive virtual reality.

[b]The knowledge test score could range from 0 to 20.

4.4.3 Visual Attention During Orientation and Expertise (Hypothesis 1a)

MFRs and non-MFRs did not differ significantly in their mean durations of average fixation in the 4 AOI categories of patients, vehicle impact zones, safety aspects and distractor, Pillai's trace = 0.03, $F(4, 52) = 0.40$, $p = .81$. Furthermore, there were no significant correlations between the knowledge test score and the average fixation duration in any of the 4 AOI categories (Table 4.3).

Table 4.3

Duration of average fixation (DOAF) of specific area of interest (AOI) categories across the 5 scenarios.

DOAF AOIs	Patients	Safety	VI[a] zone	Distractor
Examples of AOI cues	All patients and no bystanders	Scenario 1: spilled oil; scenario 2: ongoing traffic; scenario 5: ongoing traffic, spilled oil, broken glass	VI zones in all scenarios; in scenario 3 and scenario 4 mean value of 2 VI zones	Only in scenario 3: filming bystander
Value per group, mean (SD)				
MFRs[b]	397.41 (105.55)	319.09 (105.69)	314.50 (80.39)	215.13 (92.28)
Non-MFRs	403.18 (106.97)	319.37 (91.20)	315.21 (92.61)	241.35 (118.76)
Correlation with triage knowledge test				
ρ	.06	-.10	.17	.23
p	.61	.40	.15	.08

Note. Time in milliseconds; fixation duration was first averaged per scenario and then averaged across the five scenarios.

[a]VI: vehicle impact.

[b]MFR: medical first responder.

Regarding the fixation count, MFRs and non-MFRs did not differ across the 4 AOI categories, Pillai's trace = 0.15, $F(4, 52) = 2.37$, $p = .06$. Again, the triage knowledge test was not significantly associated with the fixation count in any of the AOI categories (Table 4.4).

Table 4.4

Number of fixations of specific area of interest (AOI) categories across 5 scenarios.

FC[a] AOIs	Patients	Safety	VI[b] zone	Distractor
Examples of AOI cues	All patients and no bystanders	Scenario 1: spilled oil; scenario 2: ongoing traffic; scenario 5: ongoing traffic, spilled oil, broken glass	VI zones in all scenarios; in scenario 3 and scenario 4 mean value of 2 VI zones	Only in scenario 3: filming bystander
Value per group, mean (SD)				
MFRs[c]	12.51 (4.43)	8.90 (5.46)	8.84 (3.29)	2.73 (2.13)
Non-MFRs	12.76 (4.30)	11.51 (6.47)	7.71 (2.80)	3.85 (2.36)
Correlation with triage knowledge				
ρ	-.09	-.16	.02	.04
p	.46	.18	.89	.76

Note. Fixation count was summed up per scenario and then averaged across the five scenarios.

[a]FC: fixation count.

[b]VI: vehicle impact.

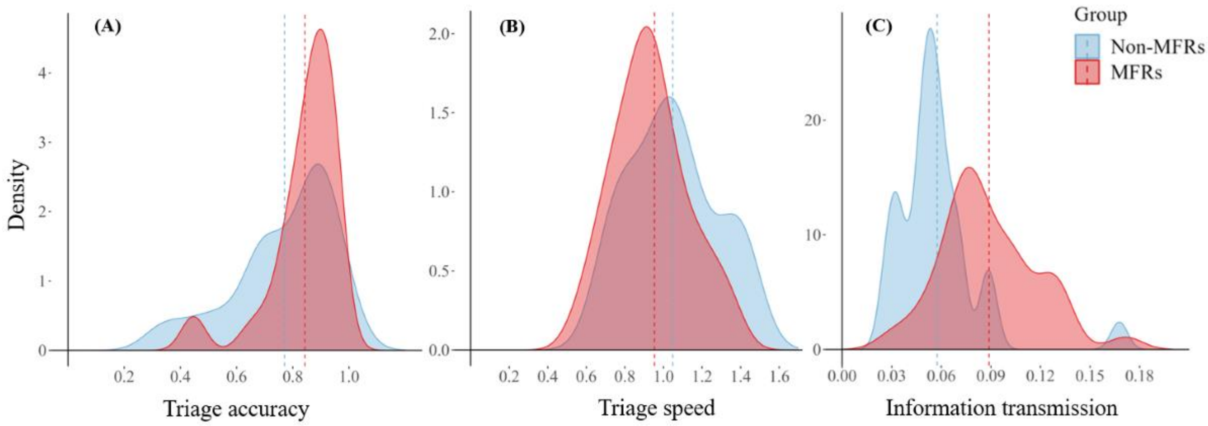
[c]MFR: medical first responder.

4.4.4 Triage Accuracy and Expertise (Hypothesis 1b)

MFRs had a significantly higher triage accuracy than non-MFRs, $t_{Welch}(61.62) = -2.04$, $p = .02$ (Bonferroni-Holm corrected $p = .045$), $d = 0.48$. On average, MFRs triaged 84% of the patients correctly ($SD = 12\%$) while non-MFRs triaged 77% correctly ($SD = 18\%$; see Figure 4.2A depicts the distributions).

Figure 4.2

(A) density plots per group for triage accuracy, (B) triage speed, and (C) information transmission efficiency. The dashed lines mark the average per group.



Furthermore, a higher knowledge test score was associated with significantly higher triage accuracy, Spearman's ρ = .40, *p* < .001. A multiple regression analysis was used to test this relationship while controlling for age and gender. More prior triage knowledge significantly predicted higher triage accuracy during the scenarios, *b* = 0.02, *SE* = 0.005, *p* = .002, while age (*b* = -0.001, *p* = .67) and gender (*b* = 0.005, *p* = 0.89) were not significant predictors. The model explained 9% of the variance (adjusted R^2), R^2 = .13, *F*(3,70) = 3.45, *p* = .02. However, according to the Shapiro-Wilk test, the assumption of normal distribution of the residuals (*p* <.001) was violated. Therefore, bootstrapping with 5000 draws was therefore used to test the stability of the results. Again, prior triage knowledge significantly predicted triage accuracy during the scenarios, *b* = 0.02, *SE* = 0.00, 95% CI [0.01, 0.03].

4.4.5 Triage Speed and Expertise (Hypothesis 1c)

MFRs completed the triage process significantly faster than non-MFRs, *t*(72) = 1.79, *p* = .04 (Bonferroni-Holm corrected *p* = .045), *d* = 0.42. On average, MFRs required 95% of the average triage speed, whereas non-MFRs needed 5% more than the average (MFRs: *M* = 0.95, *SD* = 0.23 and non-MFRs: *M* = 1.05, *SD* = 0.23; Figure 4.2B depicts the distributions). The knowledge test score was not significantly associated with triage speed, Spearman $\rho = -0.04$, $p = .72$.

4.4.6 Information Transmission and Expertise (Hypothesis 1d)

MFRs transmitted information significantly more efficiently than non-MFRs (MFRs: *M* = 0.09, *SD* = 0.03; non-MFRs: *M* = 0.06, *SD* = 0.03), $t(68) = -4.74$, $p < .001$ (Bonferroni-Holm corrected $p < .001$), $d = 1.13$ (see Figure 4.2C depicts the distributions). However, there was no significant correlation between the knowledge test score and efficiency of information transmission, Spearman $\rho = 0.19$, $p = .11$.

Explorative analyses were conducted with information transmission scores and word counts separately. Averaged across the scenarios, MFRs did not transmit significantly more correct information (MFRs: *M* = 2.15, *SD* = 0.80; non-MFRs: *M* = 2.07, *SD* = 0.72; maximum score = 4), $t(68) = -0.44$, $p = .33$, $d = 0.11$. However, MFRs transmitted information in fewer words (MFRs: 32.34, *SD* = 20.99; non-MFRs: 46.09, *SD* = 23.68), $t(68) = 2.57$, $p = .006$, $d = 0.62$. Knowledge test scores were not significantly correlated with the number of correct information (Spearman $\rho = -0.05$, $p = .67$) or information transmission length (Spearman's $\rho = -0.13$, $p = .29$).

4.4.7 Testing the Incremental Value of Attention (Hypothesis 2)

As reported for hypothesis 1a, the attentional indicators assessed with eye tracking did not discriminate between different levels of expertise. Accordingly, testing the incremental value of visual attention indicators beyond the other performance indicators for distinguishing levels of expertise (hypothesis 2) was inapplicable.

4.4.8 Subjective Performance and Objective Performance Indicators (Hypothesis 3)

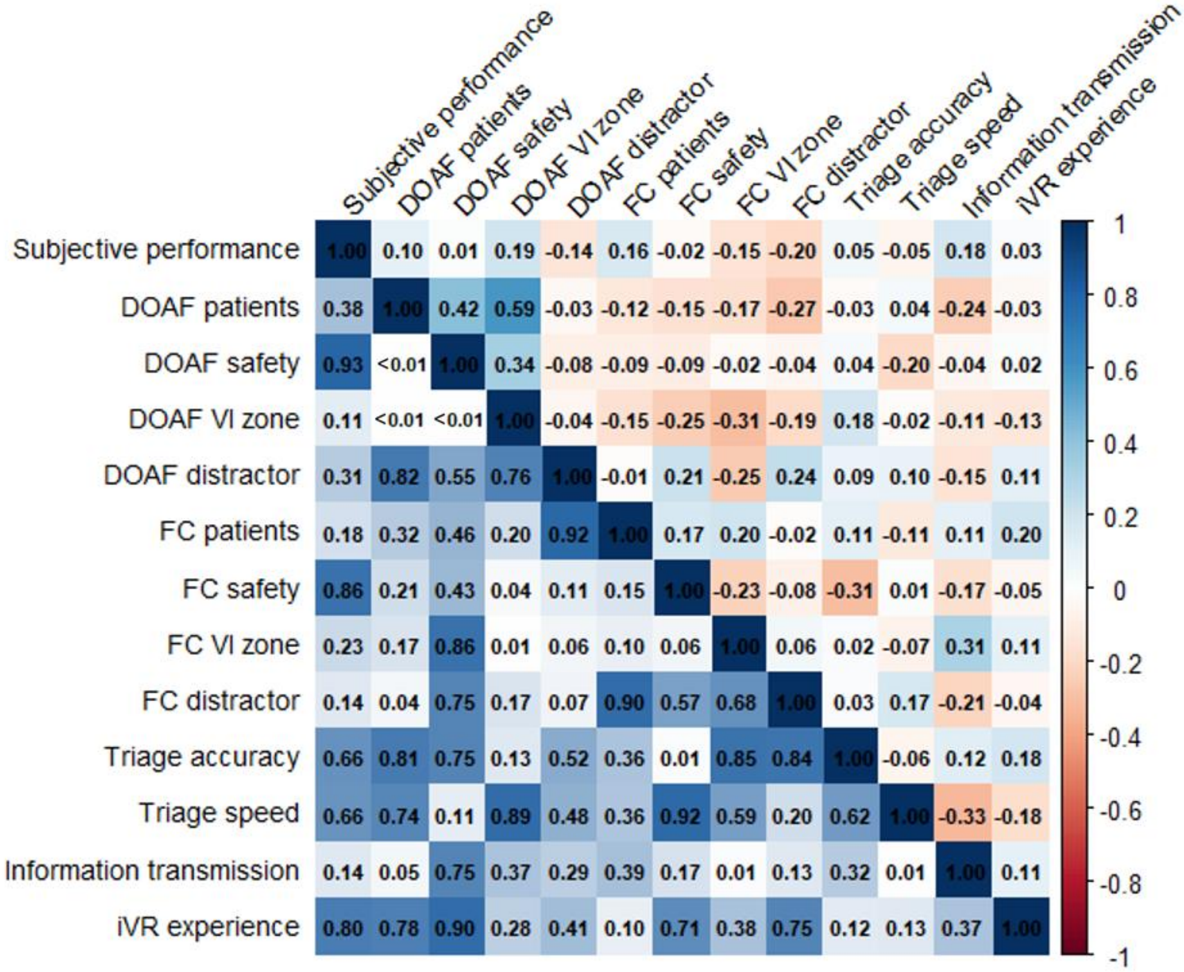
There were no significant correlations between subjective performance (*M* = 6.77, *SD* = 1.13) and any of the objective performance indicators (all $p > .05$; see Figure 4.3). Conversely, significant associations were observed between some objective performance indicators, including between greater triage speed and less efficient information transmission (rho = -.33, *p* = .01), and between less visual attention to safety cues (as measured by fixation count) and greater triage accuracy (rho = -0.31, *p* = .01).

4.4.9 Explorative Analyses

Exploratively, we tested whether prior iVR experience was associated with any of the examined performance indicators. For both MFRs and non-MFRs, the overall median value was 2 (i.e., tried iVR 1-2 times). Prior iVR experience did not correlate with any performance indicator (all $p > .05$; Figure 4.3).

Figure 4.3

Spearman correlation analysis among the performance indicators and immersive virtual reality (iVR) experience (above diagonal: ρ values, below diagonal: p values).



Note. iVR experience refers to prior iVR experience as assessed in the preliminary questionnaire, prior iVR experience was not winsorized because of the ordinal data structure. DOAF: duration of average fixation; FC fixation count; VI: Vehicle impact.

4.5 Discussion

4.5.1 Principal Findings

The prehospital triage process is a central aspect of managing MCIs, and if done correctly, leads to an efficient allocation of treatment and transport. As MCI real-life exercises are resource-intensive and therefore infrequently conducted, iVR has emerged as a new triage training method (Baetzner et al., 2022). Due to the novelty of such training, systematic evaluations of potential indicators that can be used to assess performance are still needed. This study aimed to provide such an evaluation. The objective performance indicators of triage accuracy and speed, as well as information transmission efficiency, significantly differentiated between MFRs and non-MFRs. In addition, higher triage accuracy was significantly associated with more prior triage knowledge, even when controlling for age and gender. Interestingly, however, visual attention did not differ with the level of expertise. Furthermore, in contrast to the third hypothesis, subjective performance was not correlated with any other performance indicator. Next, we discuss possible explanations as well as future directions for deriving meaningful performance indicators in iVR MCI training.

4.5.2 Visual Attention During Orientation

Visual attention represents a relatively novel potential performance indicator that has received limited attention from research in prehospital contexts. In other medical fields, studies using eye tracking have already identified differences in visual attention between experts and novices (Al-Moteri et al., 2017; Currie et al., 2019; Shinnick, 2016). In this study, we found no significant associations between visual attention and expertise, which aligns with findings based on 2D videos of an MCI (Price et al., 2022). However, the results stand in contrast to 2 studies using 2D prehospital images as stimuli (Loth et al., 2019; Prytz et al., 2018). Loth et al. showed that after START triage training, participants directed their attention more swiftly to triage information boxes. The transferability of such findings to a 3D virtual scenario without information boxes remains uncertain (Loth et al., 2019). In the second study, MFRs and novices differed significantly in their visual attention when looking at MCI photos without information boxes (Prytz et al., 2018). Several factors could explain the discrepancies. First, the former study used a purely 2D examination, which may not translate seamlessly into 3 dimensions. In contrast, iVR offers greater ecological validity for natural gaze behavior by providing a larger field of view and the ability to visually explore the environment by moving the head (Clay et al., 2019). Moreover, the stereoscopic visualization in iVR enhances depth perception through binocular cues (Lang et al., 2012). Previous research has demonstrated that 3D input can significantly alter attentional processes compared to 2D input (Li et al., 2020), although it has not yet been directly compared in the context of MCI scenarios. The additional visual cues in the iVR scenarios of this study may have enabled non-MFRs to more quickly identify relevant AOIs, thereby reducing differences between MFRs and non-MFRs. Second, it is also possible that MFRs had to exert more effort in filtering out irrelevant visual stimuli than they would with two-dimensional input, which could diminish potential differences between MFRs and non-MFRs. Furthermore, the former study (Prytz et al., 2018) included true novices, whereas the current study included non-MFRs with medical knowledge (primarily medical students) suggesting a possible power issue due to minimal effects between these groups, necessitating a larger sample size. Alternatively, it is plausible that both groups in this study possessed a minimum level of medical expertise necessary for effective visual attention, and effective visual attention remains relatively stable with additional triage knowledge.

Furthermore, for MFRs, visual attention, especially toward non-patient information, such as safety threats and distractors, may be less critical than in professions such as law enforcement or military, where constant threat awareness is an integral part of training. For instance, police officers with specialized training have exhibited superior visual search behavior for threats and overall performance in real-life exercises compared to their counterparts without such training (Murray et al., 2024). Another possibility is that differences in visual attention may only manifest when observing individual patients and their injuries, whereas our study primarily involved participants viewing a broad overview of an accident scene during eye-tracking measurements. Finally, it is worth considering that alternative visual attention indicators may offer more nuanced insights. The selection of appropriate eye-tracking measures is crucial, as certain fixation and saccade metrics or scan paths may be more applicable for specific purposes than others, depending on the task and study population (Borys & Plechawska-Wójcik, 2017). This study used 2 eye-tracking measures often used in studies on visual attention in medical contexts (Al-Moteri et al., 2017; Blondon et al., 2015), the number and length of fixations in predefined areas of interest (i.e., patients, safety aspects, vehicle impact zone, and distractor AOIs). While fixation-derived metrics can be used as a measure for visual attention and processing, saccade-derived metrics could provide further insights into visual search strategies (Al-Moteri et al., 2017). Furthermore, experts may exhibit distinct eye movement patterns and rapidly develop a holistic view of the scene, as observed in other medical domains (Al-Moteri et al., 2017).

4.5.3 Triage Accuracy and Speed

During real-life MCI exercises, triage accuracy and speed have proved to be useful as performance indicators, although usually not comparing different levels of expertise but pre-post improvements after training (Cicero et al., 2012; Cicero et al., 2017; Dittmar et al., 2018). Other studies used the indicators to compare different training methods (Ingrassia et al., 2015; Knight et al., 2010; Mills et al., 2020; Yanagawa et al., 2018). Triage accuracy and speed have also been used in iVR settings (Lowe et al., 2020; Mills et al., 2020; Vincent et al., 2008), with evidence suggesting that with more iVR training, participants become better and faster at triaging in virtual MCI situations (Vincent et al., 2008).

The results of this study suggest that accuracy is a suitable performance indicator to differentiate between different levels of expertise, with a medium effect size in terms of profession (MFRs vs non-MFRs) and a medium to large effect size in terms of triage knowledge (Cohen, 1988). These results are consistent with a previous study that found paramedic students to be as good at triage accuracy in a real-world exercise as in an iVR scenario (Mills et al., 2020). Conversely, a previous iVR study based on 360° video recordings found a less clear relationship between triage accuracy and expertise, as there were no significant differences in triage scores between residents and attendings, or between those with and without prior MCI experience (Lowe et al., 2020). The medium rather than large effect size in the comparison between MFRs and non-MFRs may be because the non-MFR group were not true novices, but also had some medical knowledge. In addition, participants were able to reread the START triage algorithms before starting the scenarios, possibly reducing the difference in triage accuracy performance.

In addition, this study suggests that speed can serve as an indicator to discriminate between MFRs and non-MFRs, as MFRs completed triage faster with a medium effect size (Cohen, 1988). However, caution should be exercised when comparing this marker to real-life exercises, as previous studies suggest that triage in virtual MCI scenarios may occur faster than in real-life scenarios (Mills et al., 2020). This difference is likely due to differences in locomotion and information retrieval. In this study, participants were able to teleport with their controllers, allowing them to quickly cover large distances. Walking, as a method of locomotion, would be closer to real-life but would require significantly larger training areas. Future studies could compare MFRs and non-MFRs in iVR systems with natural locomotion and more realistic patient information retrieval to test whether the differences between the groups are similar or greater than in this study.

4.5.4 Information Transmission

The transmission of information to the control center is crucial for the effective coordination of rescue services and thus for the management of large-scale operations. As depicted in the SSSS scheme, the rapid and clear transmission of information regarding the MCI scene, safety hazards, patient information, and the need for additional support is essential for mobilizing necessary reinforcements (Schmid et al., 2022). To our knowledge, this study was the first to record radio messages and assess information transmission efficiency as a performance indicator differentiating between different levels of expertise in iVR MCI training. MFRs transmitted information more efficiently than non-MFRs (large effect size (Cohen, 1988)), yet no correlation with the triage knowledge test score was observed. Furthermore, exploratory tests revealed that MFRs did not transmit significantly more accurate information but managed to do so using fewer words. These results suggest that a certain level of theoretical triage knowledge, that both groups seem to have had, is sufficient to facilitate MFR communication. However, the ability to quickly convey relevant information may improve with practical field experience.

The finding that MFRs transmit information more efficiently is consistent with a previous non-iVR study, which found that trained physicians were better at following a reporting scheme in a real-world simulation than their untrained counterparts (Ingrassia et al., 2013). However, the radio messages in the former study were rated by observers only in 3 categories: not performed, incomplete, and complete. In this study, the radio messages were analyzed in more detail. With the proposed method, trainees could receive more specific feedback in addition to their efficiency score, including which SSSS categories they missed or inaccurately reported. Furthermore, information transmission as a performance indicator holds promise for seamless integration into iVR, given the prevalent incorporation of microphones in iVR headsets and the increasing feasibility of automated evaluations through real-time transcription and artificial-intelligence applications. Future research may want to compare radio transmissions from iVR exercises with real-life exercises and actual MCI radio communications, providing further insight into differences in communication effectiveness and situational awareness, including the recognition and prioritization of hazards. In an Austrian study with 7 firefighters and 6 paramedics, participants underwent a 1-minute reporting phase following either an MCI scenario in iVR or non-immersive VR on a computer screen (Paletta et al., 2022). In this former study, the information given during the reporting phase was analyzed regarding situational awareness, the anticipation of consequences, and the communication of actions. There was no significant difference between the 2 training modalities. This highlights the utility of verbal information reporting as a performance indicator not only within iVR contexts but also across various training modalities.

Nevertheless, future training should consider the locally used reporting schemes. In the United Kingdom, for instance, the METHANE (a mnemonic for "major incident declared; exact location; type of incident; hazards; access; number and type of casualties; emergency services present and required") framework is commonly used as a guideline for reporting information (Lowes & Cosgrove, 2016). While the SSSS scheme greatly overlaps with METHANE, certain aspects were not particularly explored in this study, such as the exact location and access possibilities. In this study, participants started at the accident site instead of arriving there using a virtual ambulance vehicle and also received no detailed information about the site beforehand. Such aspects must be considered in studies using the METHANE scheme.

4.5.5 Subjective Performance

Self-reported performance did not correlate with the applied objective performance indicators, which contradicts previous research suggesting that self-assessed performance validity is at least low to moderate in medical education (Blanch-Hartigan, 2011; Gordon, 1991). This discrepancy may stem from the fact that individuals often base their perception of abilities on comparisons (Möller & Marsh, 2013). In terms of self-evaluation, people commonly compare themselves to others (social comparisons) or to their own past performance (temporal comparisons) (Möller & Marsh, 2013). However, in this study, the participants did not witness other trainees' performances, limiting social comparison options. In addition, comparing one's own progress in performance may have been complicated by the scarcity of MCI training, particularly because the non-MFR group lacked prior MCI training experience. It is also possible that asking about a specific behavior, such as self-rated information transmission performance, might have resulted in an estimate closer to the specific objective indicators. While we assessed global self-rated performance, future studies could test whether specific self-rated performance assessments correlate with objective indicators.

Overall, the findings underscore the significance of integrating objective indicators, providing trainees with benchmarks to assess their performance. Still, self-evaluations remain important because they help trainees reflect on their performance and enhance their self-awareness (Gow, 2013). Furthermore, comparing objective and subjective performance could potentially mitigate biases, such as the tendency for women to underestimate themselves compared to men (Blanch-Hartigan, 2011).

4.5.6 Limitations and Future Research

This study included a German sample of participants that were, on average, relatively young. Consequently, the generalizability of these findings to older populations and those from other countries should be tested in future research. For instance, previous iVR research found that cohorts over the age of 40 years had lower triage accuracy than younger cohorts (Lowe et al., 2020). This could potentially be attributed to greater familiarity with learning technologies.

This study was based on the START triage algorithm because it is well known internationally and several triage algorithms are based on it (Bazyar et al., 2019). However, in Germany, the learned triage algorithm varies depending on the rescue organization and region; thus, most of our sample was unfamiliar with START beforehand. Nevertheless, it is quite similar to the mSTART algorithm (Bazyar et al., 2019), which was the most frequently known algorithm in our sample. Future studies could program scenarios to initially allow selection of the learned algorithm, through which the scenario would then proceed, as in a previous study (Berndt, Wessel, Mentler, et al., 2018).

Furthermore, research on visual attention in prehospital settings is still in its infancy. Larger studies considering various attention and eye-tracking indicators, such as gaze movement patterns, would be beneficial in this regard. In this study, we investigated visual attention after arriving on scene and therefore only used the first thirty seconds. For standardization purposes, participants stayed at the starting position during this time. However, visual attention during the exploration of the scene or even during the whole scenario could also be an interesting indicator and more representative of behavior at real MCIs.

In this study, the ability to interact with patients was limited due to technical constraints of the software. In contrast to 360° videos, a major strength of this study was that participants were free to move around, which required them to decide how to navigate the environment, gain a full overview, and ensure that all patients were located. However, other than moving freely and using the controllers to open car doors, other actions, such as obtaining patient information, relied solely on visual and auditory cues as well as verbal information acquisition. Recent advances in iVR MCI software have introduced additional interaction features, including the use of a virtual toolkit with controllers, haptic feedback through vibrating controllers when taking a pulse, and patients responding to simple commands through automated speech recognition (Kman et al., 2023). As the need for verbal requests to obtain patient information may have reduced immersion, future training may benefit from the use of iVR MCI tools with more interactive features than those available in this study.

In addition, the assessment of expertise should be examined in a more nuanced manner. We used the participants' professions as well as a triage knowledge test as expertise indicators to cover the expertise construct in a broad manner, including theoretical knowledge and practical experience. In contrast, previous studies often only used one indicator such as years of experience (Price et al., 2022) or profession (Prytz et al., 2018). Still, measuring expertise is challenging and likely depends on numerous factors, such as years of experience (full or part-time), exposure to MCIs, and the amount of MCI training. Future studies could delve deeper into identifying measures suitable for assessing expertise and perhaps propose a composite measure encompassing multiple factors.

4.5.7 Practical Implications

This study's results highlight the suitability of iVR scenarios for MCI training and performance assessment. With its resource-efficient implementation, iVR training offers a valuable opportunity to supplement current MCI training. Although non-immersive VR on a computer monitor may be an even more affordable option, immersive training is associated with learning gains, increased enjoyment, and improved concentration (Guitiérrez et al., 2007; Mahmoud et al., 2020). While initial costs for hardware, software, and personnel occur, iVR becomes less expensive with repeated use compared to real-life exercises, which require ongoing organizational and financial resources (Mills et al., 2020). In addition, iVR software can reduce costs associated with observers by automating performance assessments, facilitating self-directed training, and allowing trainers to focus on higher-level observations. Because previous research has shown that MCI performance declines over time, MCI training should be conducted more frequently than once a year to ensure high-quality triage (Dittmar et al., 2018). In this context, iVR training and performance assessment could provide a valuable supplement to enable regular MCI training for large numbers of first responders. The integration of iVR into standard training curricula could begin with theoretical instruction (Dittmar et al., 2018), followed by regular iVR training, which could also be used to prepare MFRs for large-scale, real-life exercises.

In particular, triage accuracy, speed, and information transmission seem to be effective performance indicators for possible incorporation into iVR MCI training, complementing each other by providing insights into various aspects of overall performance. These indicators could be seamlessly integrated into future iVR learning programs to automate performance evaluation, thereby reducing the workload on trainers and allowing them to focus on higher-order evaluations, such as general procedures. The tested indicators are all suitable for individual MCI training but may also be applied in team training and could even be enhanced with novel team performance indicators, such as position and movement tracking (Wespi et al., 2023). Furthermore, the integration of real-time performance measurement into a (smart) scenario control could be valuable. On the basis of live data, the scenario difficulty could be dynamically adjusted by either the trainer or artificial intelligence to meet the individual needs of trainees.

4.5.8 Conclusions

Overall, iVR proved to be a valuable tool for assessing the performance of MFRs in MCIs scenarios. The performance indicators triage accuracy, triage speed, and information transmission can be extended to MCI training in iVR and capture multiple aspects of MCI performance. While visual attention did not function as a valid performance indicator in this study, future research might further explore visual attention as a potential indicator by examining other aspects such as gaze patterns. Overall, iVR could be integrated into current MCI training curricula to provide objective and potentially automated performance assessments and allow for more frequent practice.

Chapter 5. General Discussion

MCIs are considered particularly stressful for MFRs due to their complexity and the additional tasks involved (Chaput et al., 2007; Hugelius et al., 2020). Nevertheless, the performance of MFRs is critical to saving lives and improving patient health outcomes (Uhl, Schrom-Feiertag, et al., 2023; Zechner, Uhl, et al., 2024). To ensure adequate preparation of MFRs, MCI training must be effective despite the financial, temporal, and organizational boundaries of MFR organizations (Mills et al., 2020). Traditional MCI training is often either classroom-based without realistic, hands-on experience, or based on resource-intensive RLEs (Mills et al., 2020). To fill this gap, iVR and MR have emerged as new MCI training methods with increasing realism and feasibility, potentially allowing trainees to receive representative training, including training under pressure. However, systematic and adequate evaluation of such training is needed to ensure its effectiveness and identify areas for improvement (Wakasugi et al., 2009). This dissertation therefore focused on testing the potential of iVR and MR as training methods to prepare MFRs for MCIs.

A systematic literature review of MCI training evaluation studies indicated the potential of iVR, but revealed a lack of research on immersive virtual training, in particular MR with haptic feedback (Chapter 2). To address this research gap, MR MCI training was compared to the gold standard, high-fidelity RLEs (Chapter 3). The results indicated that MR training and RLEs elicited comparable responses on several outcomes, highlighting the potential of MR as a valuable MCI training method, despite some limitations. Furthermore, most previous studies relied on subjective measures, with less than one-third assessing behavioral performance (Chapter 2). iVR and MR, however, offer the possibility to implement objective performance assessments that are difficult to implement in RLEs. Therefore, Chapter 4 validated objective performance indicators for MCI training in immersive virtual environments that can be used in future training evaluations.

Overall, the results of this dissertation demonstrated that iVR and MR training have the potential to be a valuable complement to the preparation of MFRs for MCI response, while also identifying areas for improvement. In the following sections, the results will be discussed separately for the two research questions outlined in Chapter 1: 1) What effectiveness indicators can be used to evaluate immersive virtual MCI training? 2) Can iVR and MR training be used to prepare MFRs for MCIs? Next, strengths, limitations, and future research will be outlined, followed by practical implications for iVR and MR MCI training.

5.1 Evaluation methods for iVR/MR training

This dissertation identified and employed several evaluation measures, also providing guidance on how to evaluate future training. Previous research on iVR training has been criticized regarding its quality (Mills et al., 2020). In fact, the systematic literature review in Chapter 2 found only three quasi-experimental or experimental evaluations of iVR MCI training and none of MR with haptic feedback. Thus, before the technologies can be recommended for implementation in MFR organizations, high-quality evaluations are needed. Regarding all MCI training evaluation studies, there was a wide variety of effectiveness indicators, with more than half using written knowledge tests and/or self-report measures. However, although declarative knowledge is required, transferability to the performance of actions under stress is not necessarily given.

According to the *integrative framework of stress, attention, and human performance* (Vine et al., 2016), high levels of stress can impair both attention and performance. This effect is particularly pronounced in situations where an individual does not feel sufficiently prepared for the task and has not developed effective compensatory strategies to maintain their performance under pressure (Vine et al., 2016). As MCIs are considered particularly stressful (Hugelius et al., 2020), MCI training evaluations would greatly benefit from assessing stress levels and behavioral outcomes. However, less than a third of the studies found in Chapter 2 used behavioral performance outcomes, with most studies using composite observer ratings with non-validated instruments (Chapter 2). Overall, there was little agreement on effectiveness indicators for MCI training or specific instruments for outcome measures and little use of objective performance indicators.

Based on previous research on iVR, first responder training, as well as the theoretical models *CAMIL* (Makransky & Petersen, 2021) and the *integrative framework of stress, attention, and human performance* (Vine et al., 2016), this dissertation employed a wide variety of evaluation measures. These included perceived stress, exhaustion, MCI self-efficacy, presence, learning gain, and open-ended questions on iVR/MR training in comparison to RLEs (Chapter 3). This broad set of indicators provides a holistic overview of MR in comparison to high-fidelity RLEs. Furthermore, as performance outcomes are an essential marker for the evaluation of training success (Wakasugi et al., 2009), Chapter 4 focused on the identification of objective performance indicators for immersive virtual MCI training. iVR and MR are particularly suitable for objective performance assessments (Chapter 4). Additionally, automatic assessment of relevant performance indicators reduces biases (Salas et al., 2009) and can create opportunities for MFRs to use iVR/MR for self-training in-between training sessions with a professional trainer. Moreover, trainers could use automatically collected objective training data for their feedback (Salas et al., 2009), leaving them with more capacity to focus on higher-level processes or to identify specific training needs. The identified performance indicators in Chapter 4 have been tested in iVR MCI scenarios so that they can apply to iVR as well as MR MCI training. While the tested visual attention indicators were not found to be valid, triage accuracy, triage speed and information transmission efficiency can be recommended for integration in immersive virtual training. Additionally, the results of Chapter

4 indicate that trainees may benefit from objective performance assessment, as their self-ratings did not align with objective data.

The effectiveness indicators found and used in this dissertation relate primarily to the trained individual (Chapters 2, 3, 4). However, MFRs typically work in teams (Manser, 2009; Wespi et al., under review). While the applied indicators can be used in both individual and team training, research on incorporating (additional) measures of team dynamics and performance into MCI training remains limited (Chapter 2). Information transmission to the control center was found to be an effective performance indicator in Chapter 4, contributing to the assessment of communication skills in MCI situations. Yet, there is still a lack of indicators to assess team processes within MFR teams.

Research on healthcare teams shows the same methodological issues as found in Chapter 2: A limited use of objective performance indicators. A literature review found that the majority of previous studies used self-reports of team performance (70%) and/or were dependent on observers to assess performance (30%; Marlow et al., 2018). Raters, however, often have to assess performance subjectively, so rater training is recommended for improved rating (Marlow et al., 2018). Objective indicators could save resources and reduce errors or biases. Particularly in terms of objective medical team dynamics and team performance assessment, research is still in its infancy (Wespi et al., 2023). However, various potential indicators have been discussed, including movement and position data, gaze behavior, speech content, and voice analysis (Wespi et al., 2023). For instance, team movement tracking can be relatively easily implemented in iVR/MR training and has already been used in the context of iVR police training (Uhl, Nguyen, et al., 2023).

Ideally, evaluation measures during the training would be compared to desired outcomes in a real-life setting. In the medical context this mainly refers to patient outcomes (Marlow 2018). Such a comparison would provide stronger evidence for the validity of effectiveness indicators and the training effectiveness in general (Marlow et al., 2018). However, this approach is challenging in the context of MCIs as they are unforeseen, irregular, and encompass a wide variety of crisis situations including natural disasters, infectious disease breakouts, man-made accidents, terrorist attacks or even war (Kar, 2024). MCIs therefore pose many confounding factors for a systematic evaluation of the training prior to MCIs. An alternative approach might be to introduce iVR and MR as new training methods in a particular region and compare MCI results from periods before to after their introduction. While this method is not entirely free of confounding factors, many can be mitigated by averaging MCI data over several years. However, this requires access to reliable records of past and recent MCIs.

For more immediate effectiveness testing of iVR and MR, high-fidelity RLEs such as those applied in Chapter 3 may be used. Although they are also a training method, they are considered the current gold standard (Tin et al., 2021) and can be designed to closely mirror real MCI situations. A logical next step following this dissertation could therefore be to test the effectiveness of iVR and MR training by conducting RLEs before and after training to assess performance improvements in the performance indicators identified in Chapter 4. Relatively realistic patient outcomes could be approximated by adjusting the health status and vital parameters of simulated patients dynamically during the scenario based on real patient health status changes in the past (see e.g., Neher et al., 2024 in the context of a neurological emergency in the emergency department).

5.2 Effectiveness of iVR/MR MCI training

The results of this dissertation provide initial evidence that iVR and MR training can be used to prepare MFRs for MCIs, making them a valuable complement to existing MCI training programs (Conrad et al., 2024). Building on the results of Chapter 2, which indicated that iVR MCI training can lead to several similar outcomes as RLEs (see Mills et al., 2020), Chapter 3 provides initial evidence that MR training can be used to prepare MFRs for MCIs, as it has mostly led to similar or near-similar responses as high-fidelity RLEs. Nevertheless, Chapter 2 only included three studies using iVR and the sample size in Chapter 3 was not large enough to draw a firm conclusion about the effectiveness of MR.

While further evidence with larger sample sizes and the assessment of performance improvements are needed, the effectiveness of immersive virtual training has been demonstrated in other training contexts. The positive effects of immersive virtual training described in Chapters 2 and 3 are consistent with a recently published review study that found a positive impact of iVR on learning across different learning contexts, particularly for learning content that involved active engagement and practical application (Conrad et al., 2024). In addition, the effectiveness of immersive virtual training has been proven in other medical education contexts, such as surgical skill acquisition (Mao et al., 2021) and nursing education (Choi et al., 2022). The following sections focus on key factors of effective, representative MCI training identified in Chapters 1-3 as promising but still requiring further research attention.

5.2.1 The roles of realism and presence for iVR/MR MCI training effectiveness

A particular strength of the MR training was its realistic simulation of the MCI environment, which was comparable to RLEs in terms of stress and exhaustion levels, and close-to-similar physical presence. According to the *CAMIL* model, presence is a key factor in immersive learning. However, social presence was less pronounced in the MR training compared to the RLE training, which may explain the moderate learning gain reported by the MR group compared to the high learning gain reported by the RLE groups. Based on the *CAMIL*, representational fidelity is an important antecedent for the feeling of presence, but technical limitations affected fidelity in the context of patient interactions and communication.

Another reason for reduced social presence may have been the *uncanny valley effect* (Mori et al., 2012). This effect can be particularly strong in immersive virtual environments and describes that increased human likeness in virtual characters typically leads to increased acceptance up to point when the character looks almost but not quite real and triggers aversion such as eeriness or even disgust (Schwind et al., 2018; Stein, 2018). Several theories exist to explain this effect, with one prominent view suggesting that the abnormality of such virtual characters elicits aversion due to perceived associations with disease (Schwind et al., 2018; Stein, 2018). In Chapters 3 and 4, the virtual patients and their wounds were rendered as realistically as possible within technical constraints to maximize the training's representational fidelity. However, this may have negatively affected the perception of the patients and reduced social presence. A key challenge in iVR/MR medical training is therefore to present patient appearance, injuries, and behavior realistically enough to ensure representational fidelity, without inducing the uncanny valley effect. Future training evaluations should consider assessing if the uncanny valley effect occurs. Achieving a level of realism that overcomes the uncanny valley remains hard to achieve for now, as both appearance and interaction must be convincingly authentic (McMahan et al., 2016). Recent research indicates that the uncanny valley can already be overcome in terms of the appearance of virtual humans on a computer monitor, but realistic human behavior is more difficult to replicate (Bae et al., 2024).

5.2.2 The role of stress for iVR/MR MCI training effectiveness

Since MCIs are inherently stressful for MFRs, training programs based on representative learning design should aim to induce a certain stress level as part of the learning experience (see Chapter 1; e.g., Hutter et al., 2023). Chapter 2 highlights that previous MCI training programs have not adequately addressed the role of stress. The comparison of MR training and RLEs in Chapter 3 revealed comparable levels of stress in all groups. The RLEs demonstrated high fidelity, achieved through the professional makeup of the patient actors and the creation of a highly realistic environment featuring a damaged bus in a road tunnel. The comparable stress levels observed in the MR group suggest that MR scenarios are equally effective for representative training as RLEs, the gold standard method. Comparable stress outcomes between immersive virtual training and an RLE have also been found in police training with n = 237 Dutch police officers (Kleygrewe et al., 2024b), indicating an already high degree of realism of immersive virtual training across different first responder contexts.

Nevertheless, overall stress levels were relatively low in both the MR and RLE groups (Chapter 3), which was also the case in the police training comparison (Kleygrewe et al., 2024b). Eliciting authentic stress responses may be challenging in MCI training, regardless of the training modality, because trainees are aware that, unlike in real MCI situations, there is no danger to them or to real patients. However, the high presence values in the RLE groups, but also medium-to-high values in the MR group indicate that participants perceived the simulated situations as relatively realistic.

An alternative explanation for low stress levels in both groups might be a too low level of training difficulty. To achieve optimal learning outcomes, scenario complexity should be tailored to individual trainees or teams, seeking a balance between inducing sufficient stress without becoming overwhelming (Hill et al., 2024). The participants in Chapter 3 were predominantly MFRs with several years of professional experience, suggesting that more complex scenarios with additional stressors may have been more appropriate to effectively challenge and engage them. Nevertheless, even training under low levels of stress can be beneficial for maintaining performance under higher levels of stress at a later time (see Hill et al., 2024 for the concept of "behavioral vaccinations").

The scenarios used in Chapters 3 and 4 were highly standardized to minimize confounding factors in the conducted research. However, for training purposes, more agile scenario adaptation could be considered, either by the trainer or automatically through algorithmic rule sets (Zechner, Uhl, et al., 2024). To this end, real-time tracking of objective performance metrics, such as those explored in Chapter 4, could be used in combination with physiological and psychological stress measurement (see e.g., Zechner, Schrom-Feiertag, et al., 2024). Based on this information, the scenario could be adapted by adding or removing scenario-specific stressors.

For the preparation of the iVR scenarios in Chapter 4, scenarios with different levels of complexity were created and stressors were identified together with MFRs from European MFR organizations. These stressors included crowds of bystanders, screaming children, and MCIs at night and in fog. Furthermore, previous research has shown that in an iVR MCI scenario, a sudden, severe injury to a virtual MFR colleague after first establishing a connection through storytelling can significantly increase both physiological and self-reported stress levels (Prachyabrued et al., 2019). However, a systematic identification and evaluation of specific stressors for iVR and MR MCI scenarios, similar to efforts in iVR police training (e.g., (Zechner, Kleygrewe, et al., 2023), is still outstanding.

Altogether, immersive virtual training seems to be comparable to RLEs in several outcomes, including the degree of realism and induced stress. Moreover, trainees saw potential in the technologies, especially in terms of reduced resource requirements and increased training frequency (Chapter 3). Building on the results of this dissertation, interesting next steps include the investigation of performance improvements through iVR/MR training and mechanisms for effective scenario adaptation for training under stress. Current limitations in the interaction with virtual patients lead to the recommendation to use the technologies as a complement, not a replacement, which was also suggested by the trainees in Chapter 3. These results are consistent with a recent study that found that MFRs are highly motivated by immersive virtual training and would consider it a good addition to current MCI training programs (Vogt et al., 2023).

5.3 Strengths, Limitations, and Future Directions

Some aspects have already been discussed in Chapters 2 to 4. Therefore, this section focuses on broader strengths, limitations, and future directions.

5.3.1 Technological aspects

A notable strength of this dissertation was the integration of state-of-the-art iVR and the latest MR technologies. These technologies continue to evolve and continuously offer meaningful enhancements. Therefore, it is most likely only a matter of time before the suggested improvements identified in Chapter 3 become feasible and should be implemented in future training. For instance, automatic speech recognition and artificial intelligence may improve communication with virtual patients, and the tracking of fine motor movements and small haptic elements could allow MFRs to practice more medical procedures in immersive virtual environments.

Despite the potential of iVR and MR demonstrated in this dissertation (Chapters 2-4), a direct comparison between iVR and MR and their effectiveness in the context of MCI training is still pending. For MFRs, hands-on interaction is critical in the diagnostic process (Zechner, Uhl, et al., 2024). Furthermore, trainees in immersive medical training with virtual patients have expressed strong appreciation for the inclusion of a haptic manikin (Girau et al., 2019). Recent studies, however, present mixed results regarding differences in the perception of presence between iVR and MR. While some report no significant differences in overall presence (Elsenbast et al., 2024), others find no significant differences in physical presence but report higher social presence scores in MR compared to iVR (Uhl et al., 2024).

Future research should examine how iVR and MR compare in MCI training, particularly with regard to their impact on learning outcomes and performance improvement. Furthermore, for an even more holistic assessment of performance, future MCI training studies could identify and validate behavioral performance indicators that are supported by MR elements. Such performance indicators could include the correct application of tourniquets and the correct execution of placing unconscious patients in recovery positions. However, given that the current cost of MR training like the one used in Chapter 3, is not feasible for many MCI organizations, the performance indicators validated in Chapter 4 were tested in iVR. These indicators are applicable to both technologies and support single-user and multi-user environments.

5.3.2 Training content and scenario development

A particular strength of this dissertation is the close collaboration with MFR organizations and technology partners to ensure that the training content and scenarios in Chapters 3 and 4 are highly relevant for MFRs. Thanks to their iterative input, the simulated traffic accidents and the patients with their injuries were portrayed as realistically as possible to increase representational fidelity. The scenarios used in Chapters 3 and 4 made use of a major advantage of iVR/MR: the ability to simulate scenarios that are difficult to recreate in RLEs. In particular, bus accidents, accidents on the highway or in a road tunnel, MCIs in different weather and visibility conditions, and scenarios involving different patient and bystander groups, such as children, elderly people, or pregnant women, have enriched the scenarios. Nevertheless, for Chapter 3, the MR scenario with a bus accident in a tunnel was recreated at great expense to allow a comparison of MR and RLEs.

Despite their strengths, the scenario content also had its limitations. All the scenarios developed for chapters 3 and 4 involved road traffic accidents. This narrow scope does not reflect the full range of MCIs, which can vary considerably in terms of the safety risks to MFRs and the patient injuries. To date, no studies have compared different immersive virtual MCI scenarios to identify specific training needs based on the type of incident. Expanding research in this area could provide critical insights for tailoring MFR training to different MCI contexts.

Furthermore, the MR setup in Chapter 3 only included manikins representing adult males for haptic feedback. The lack of diversity in manikins, particularly those representing female or other demographic groups, is a known issue in medical education. Globally, most manikin manufacturers produce male-representative manikins, with 95% of manikins being flat-chested and more indicative of the male gender (Szabo et al., 2024). Accordingly, a survey of 56 training centers for cardiopulmonary resuscitation in North and Latin America found that only 6% included female-representative manikins in their training programs (Liblik et al., 2023). While the lack of female manikins has been discussed as a reason for increased hesitation and error rates among medically trained military personnel treating women (Mazzeo et al., 2021; Vaughan et al., 2024), effects of non-diverse manikins on MFR training are not sufficiently understood.

5.3.3. Measurement and Assessment

Drawing not only on MCI training research as reviewed in Chapter 2, but also on iVR/MR research, medical education research, and theoretical models such as the *CAMIL* (Makransky & Petersen, 2021) and the *integrative framework of stress, attention, and human performance* (Vine et al., 2016), this dissertation employed several multi-method evaluation approaches. These include: 1) quantitative self-rated measures such as stress, exhaustion, self-efficacy, learning gain, and presence (Chapter 3), 2) quantitative objective measures such as visual attention via eye-tracking, triage accuracy, and triage speed (Chapter 4), 3) qualitative data analysis of radio messages with two blinded and independent raters (Chapter 4), and 4) qualitative feedback from trainees after the training (Chapter 3). The variety of evaluation measures provided a holistic understanding of the potential of iVR and MR in the context of MCI training for MFRs.

Building on the findings of this dissertation, future research could further explore the assessment of learning outcomes in the context of iVR/MR training. First, the effectiveness of iVR and MR for specific learning objectives and outcomes is not sufficiently understood. Studies on iVR and learning suggest that immersive virtual training is particularly suitable for procedural knowledge and knowledge transfer (Jongbloed et al., 2024; Makransky & Petersen, 2021), but there are mixed findings on the use of iVR for factual and conceptual knowledge (Makransky & Petersen, 2021). The reasons for iVR being suitable for factual and conceptual knowledge acquisition in some training contexts and not in others are not sufficiently understood. In addition, further studies are needed to test the effect of iVR and MR on the different learning outcomes specifically in the context of MCI training. Furthermore, the retention time following iVR and MR MCI training is still not determined. Future research should examine how well MFRs can remember and apply the practiced content after weeks or months in order to determine meaningful training intervals.

5.4 Practical implications and recommendations for iVR and MR MCI training

Based on the findings of this dissertation, MFR organizations can enhance their MCI training curriculum by integrating iVR and MR technologies. As discussed in Chapters 2-4 and highlighted by participants in Chapter 3, iVR and MR provide a high degree of flexibility and require less organizational and personnel resources than RLEs. This enables increased training frequency, allowing trainees to develop routine (Chapter 3). iVR MCI training is also financially attractive compared to RLEs (Mills et al., 2020), while MR may still be too expensive for most MFR organizations at this time. The following key considerations should be taken into account when implementing iVR/MR MCI training:

5.4.1 Procedure

Training in iVR/MR, as described in Chapter 3, typically follows a common procedure (see e.g., Zechner, Kleygrewe, et al., 2023; Zechner, Uhl, et al., 2024). The training begins with a briefing to inform trainees about the training objectives and process, as well as to provide them with the iVR/MR equipment. This is followed by a tutorial or practice scenario to familiarize trainees with the technology. Once prepared, trainees proceed with the scenario(s). Between scenarios, or after all scenarios have been completed, a debriefing is held to discuss the training session. Finally, the training session should be evaluated.

5.4.2 Training preparation

As with any training program, it is important to first assess training needs and identify objectives. This can help MFR organizations to decide whether iVR/MR training, RLEs, traditional classroom training, or another training method is most appropriate (see Chapter 2 for an overview of evaluated MCI training methods). For example, basic skills such as initiating and formulating radio communications can be effectively practiced using real radios, while fine-motor, medical procedures can be practiced using manikins. Based on previous research on iVR and its impact on learning, iVR/MR technologies can be particularly effective in translating theoretical knowledge into practical skills (Makransky & Petersen, 2021). The technologies also offer unique advantages, especially for scenarios that are difficult or impossible to simulate using traditional methods (see Chapter 3). In addition, iVR/MR is well suited for individual or small-group training, where the participation of numerous MFRs, other first responders or patient actors is not necessary to achieve a realistic experience.

As discussed in Chapter 4.5, iVR/MR can even be integrated into a broader training curriculum. This integration could involve using iVR/MR after classroom-based teaching of factual knowledge and before engaging in resource-intensive, large-scale RLEs. This approach allows trainees to put theoretical knowledge into action and build confidence and skills with a potentially individualized training frequency, optimizing their preparation for high-fidelity RLEs and real MCI situations.

5.4.3 Training execution

As IVR/MR systems can differ in their handling and not all MFRs have experience in using the technologies, tutorials are essential before starting the training scenario (Heldring et al., 2024). In the MR system described in Chapter 3, it was particularly important for trainees to practice taking vital signs and applying tourniquets and triage cards, which trainees practiced together on manikins in a virtual 'waiting room'. The tutorial in Chapter 4 took place in a quiet virtual residential area, allowing trainees to practice locomotion and patient information gathering. Given the risk of technical problems during iVR/MR scenarios, such as temporary HMD failures (Berndt, Wessel, Willer, et al., 2018), it can be beneficial to have staff or trainers available to provide assistance. During the scenario, trainers can typically observe the training on an external screen and make valuable observations beyond what the system records. In addition, iVR and MR offer the ability to modify the scenarios in real time to optimize the balance between the trainees' expertise and the level of complexity. Depending on the system's capabilities, scenario adjustments can be based on a variety of factors, such as recorded performance metrics and real-time stress data sent via portable biosensors (Zechner, Uhl, et al., 2024).

5.4.4 Debriefing and training evaluation

A key benefit of iVR/MR training is its ability to enhance the debriefing process (Zechner, García Guirao, et al., 2023). The technologies can be used to create *kind learning environments* (Hogarth, 2001), in which trainees can develop and improve their intuitive expertise (Hogarth, 2001, 2010; Kahneman & Klein, 2009). Beyond providing representative scenarios, the development of intuitive expertise is substantially supported by immediate and high-quality feedback (Kahneman & Klein, 2009). Such feedback can be delivered during the post-scenario debriefing or even be implemented as an immediate response to specific actions, such as instant feedback following a triage category selection.

Furthermore, the technologies allow for richer post-scenario reflection by providing detailed recordings and analyses that assist trainers and trainees to better understand how the scenario evolved, why certain outcomes occurred, and how performance can be improved (Zechner, García Guirao, et al., 2023). Key moments can be reviewed from multiple perspectives, such as individual team members' or a bird's eye view (Kleygrewe et al., 2024a). Especially when combined with recordings of stress data and objective performance metrics, iVR/MR debriefings provide a more comprehensive evaluation than traditional training methods, supporting deeper feedback and understanding (Zechner, García Guirao, et al., 2023).

Another essential point is the prior reflection of what training success would look like according to the learning objective and which effectiveness indicators are suitable (see Chapter 2-4). For instance, improvements in relevant outcome variables can be assessed through pre- and post-tests. According to Chapter 4, MCI performance indicators such as triage accuracy, triage speed, and information transfer efficiency are particularly recommended. Although self-assessed performance did not correlate with objective performance in Chapter 4, it remains valuable for enhancing reflective skills (Gow, 2013). In addition, lack of improvement in relevant competencies, or insights from trainee evaluations, as collected in Chapter 3, can provide critical feedback for ongoing training program improvement.

5.5 Conclusion

Regardless of whether they are man-made or caused by natural disasters, MCIs cause a large number of injured patients and fatalities each year. To effectively manage these high-stress situations, MFRs require frequent, high-quality training. While classroom-based training lacks the realism and complexity of real MCIs, RLEs are considered the gold standard, but are too resource-intensive for frequent training. This dissertation demonstrates that iVR and MR can help bridge this gap and provide a valuable addition to current MCI training curricula. Immersive virtual training mostly produces responses that are comparable to those in RLEs, highlighting its great potential. Although further research is needed to assess performance improvements through iVR/MR training, this dissertation lays a necessary foundation by validating relevant performance indicators. In particular, iVR and MR stand out for their flexibility, which can lead to increased training frequency, and for their ability to create scenarios that are difficult to simulate in RLEs.

References

Aghababaeian, H., Sedaghat, S., Tahery, N., Moghaddam, A. S., Maniei, M., Bahrami, N., & Ahvazi, L. A. (2013). A comparative study of the effect of triage training by role-playing and educational video on the knowledge and performance of emergency medical service staffs in Iran. *Prehospital and Disaster Medicine*, *28*(6), 605-609. https://doi.org/10.1017/s1049023x13008911

Al-Moteri, M. O., Symmons, M., Plummer, V., & Cooper, S. (2017). Eye tracking to investigate cue processing in medical decision-making: A scoping review. *Computers in Human Behavior*, *66*, 52-66. https://doi.org/10.1016/j.chb.2016.09.022

Alenyo, A. N., Smith, W. P., McCaul, M., & Van Hoving, D. J. (2018). A comparison between differently skilled prehospital emergency care providers in major-incident triage in South Africa. *Prehospital and Disaster Medicine*, *33*(6), 575-580. https://doi.org/10.1017/s1049023x18000699

Alharbi, A. A. (2018). Effect of Mass Casualty Training Program on Prehospital Care Staff in Kuwait. *The Egyptian Journal of Hospital Medicine*, *71*(6), 3393-3397. https://doi.org/10.12816/0047293

Alim, S., Kawabata, M., & Nakazawa, M. (2015). Evaluation of disaster preparedness training and disaster drill for nursing students. *Nurse Education Today*, *35*(1), 25-31. https://doi.org/10.1016/j.nedt.2014.04.016

Almukhlifi, Y., Crowfoot, G., Wilson, A., & Hutton, A. (2021). Emergency healthcare workers' preparedness for disaster management: An integrative review. *Journal of Clinical Nursing*. https://doi.org/10.1111/jocn.15965

Altmann, T. (2019). Training. In M. A. Wirtz (Ed.), *Dorsch Lexikon der Psychologie*. Hogrefe. https://dorsch.hogrefe.com/stichwort/training

Aluisio, A. R., Teicher, C., Wiskel, T., Guy, A., & Levine, A. (2016). Focused training for humanitarian responders in regional anesthesia techniques for a planned randomized controlled trial in a disaster setting. *PLoS Currents*, *8*. https://doi.org/10.1371/currents.dis.e75f9f9d977ac8adededb381e3948a04

Andreatta, P., Klotz, J., Madsen, J., Hurst, C., & Talbot, T. (2014). Assessment instrument validation for critical clinical competencies: pediatric-neonatal intubation and cholinergic crisis management. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL.

Andreatta, P., Klotz, J. J., Madsen, J. M., Hurst, C. G., & Talbot, T. B. (2015). Outcomes From Two Forms of Training for First-Responder Competency in Cholinergic Crisis Management. *Military Medicine*, *180*(4), 468-474. https://doi.org/10.7205/milmed-d-14-00290

Andreatta, P. B., Maslowski, E., Petty, S., Shim, W., Marsh, M., Hall, T., Stern, S., & Frankel, J. (2010). Virtual Reality Triage Training Provides a Viable Solution for Disaster-preparedness. *Academic Emergency Medicine*, *17*(8), 870-876. https://doi.org/10.1111/j.1553-2712.2010.00728.x

Anthes, C., García-Hernández, R. J., Wiedemann, M., & Kranzlmüller, D. (2016, March 5-12). State of the art of virtual reality technology. 2016 IEEE Aerospace Conference, USA. https://doi.org/10.1109/AERO.2016.7500674

Bae, S., Jung, T., Cho, J., & Kwon, O. (2024). Effects of meta-human characteristics on user acceptance: from the perspective of uncanny valley theory. *Behaviour & Information Technology*. https://doi.org/10.1080/0144929X.2024.2338408

Baetzner, A. S., Hill, Y., Roszipal, B., Gerwann, S., Beutel, M., Birrenbach, T., Karlseder, M., Mohr, S., Salg, G., Schrom-Feiertag, H., Frenkel, M. O., & Wrzus, C. (2023). *Preregistration - Triage Training in Immersive Virtual Reality: A Quasi-Experimental*

Evaluation of Multimethod Performance Indicators https://doi.org/10.17605/OSF.IO/AMGK8

Baetzner, A. S., Hill, Y., Roszipal, B., Gerwann, S., Beutel, M., Birrenbach, T., Karlseder, M., Mohr, S., Salg, G., Schrom-Feiertag, H., Frenkel, M. O., & Wrzus, C. (2025). Triage Training in Immersive Virtual Reality: Quasi-Experimental Evaluation of Performance Indicators. *Journal of Medical Internet Research (JMIR)*, *27*, e63241. https://doi.org/10.2196/63241

Baetzner, A. S., Wespi, R., Hill, Y., Gyllencreutz, L., Sauter, T. C., Saveman, B.-I., Mohr, S., Regal, G., Wrzus, C., & Frenkel, M. O. (2022). Preparing medical first responders for crises: a systematic literature review of disaster training programs and their effectiveness. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *30*(1), 76. https://doi.org/10.1186/s13049-022-01056-8

Bajow, N., Djalali, A., Ingrassia, P. L., Ragazzoni, L., Ageely, H., Bani, I., & Corte, F. D. (2016). Evaluation of a new community-based curriculum in disaster medicine for undergraduates. *BMC Medical Education*, *16*(1), 225. https://doi.org/10.1186/s12909-016-0746-6

Bandura, A. (1977). Self-Efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review*, *84*, 191-215.

Barnsley, L., Lyon, P. M., Ralston, S. J., Hibbert, E. J., Cunningham, I., Gordon, F. C., & Field, M. J. (2004). Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. *Medical Education*, *38*(4), 358-367. https://doi.org/10.1046/j.1365-2923.2004.01773.x

Barrie, M., Socha, J. J., Mansour, L., & Patterson, E. S. (2019). Mixed Reality in Medical Education: A Narrative Literature Review. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *8*(1), 28-32. https://doi.org/10.1177/2327857919081006

Bazyar, J., Farrokhi, M., & Khankeh, H. (2019). Triage systems in mass casualty incidents and disasters: a review study with a worldwide approach. *Open Access Macedonian Journal of Medical Sciences*, *7*(3), 482. https://doi.org/10.3889/oamjms.2019.119

Bazyar, J., Farrokhi, M., Salari, A., & Khankeh, H. R. (2020). The principles of triage in emergencies and disasters: a systematic review. *Prehospital and Disaster Medicine*, *35*(3), 305-313. https://doi.org/10.1017/S1049023X20000291

Berndt, H., Wessel, D., Mentler, T., & Herczeg, M. (2018). Human-centered design of a virtual reality training simulation for mass casualty incidents. 2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games). https://doi.org/10.1109/VS-Games.2018.8493427

Berndt, H., Wessel, D., Willer, L., Herczeg, M., & Mentler, T. (2018). Immersion and Presence in Virtual Reality Training for Mass Casualty Incidents. Proceedings from the 15th Information Systems for Crisis Response and Management Conference, Rochester, NY, USA.

Betka, A. A., Bergren, M. D., & Rowen, J. L. (2021). Improving rural disaster response preparedness. *Public Health Nursing*, *38*(5), 856-861. https://doi.org/10.1111/phn.12924

Beyramijam, M., Farrokhi, M., Ebadi, A., Masoumi, G., & Khankeh, H. R. (2021). Disaster preparedness in emergency medical service agencies: A systematic review. *Journal of Education and Health Promotion*, *10*, Article 258. https://doi.org/10.4103/jehp.jehp_1280_20

Bilek, N., Feldhofer, A., & Moser, T. (2021). Virtual reality based mass disaster triage training for emergency medical services. 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). https://doi.org/10.1109/VRW52623.2021.00135

Bischof, W. F., Anderson, N. C., & Kingstone, A. (2024). A tutorial: Analyzing eye and head movements in virtual reality. *Behavior Research Methods*, *56*, 8396–8421. https://doi.org/10.3758/s13428-024-02482-5

Blanch-Hartigan, D. (2011). Medical students' self-assessment of performance: results from three meta-analyses. *Patient Education and Counseling*, *84*(1), 3-9. https://doi.org/10.1016/j.pec.2010.06.037

Blondon, K., Wipfli, R., & Lovis, C. (2015). Use of eye-tracking technology in clinical reasoning: a systematic review. *Digital Healthcare Empowering Europeans*, 90-94. https://doi.org/10.3233/978-1-61499-512-8-90

Borys, M., & Plechawska-Wójcik, M. (2017). Eye-tracking metrics in perception and visual attention research. *European Journal of Medical Technologies*, *3*, 11-23.

Bowyer, M. W., Streete, K. A., Muniz, G. M., & Liu, A. V. (2008). Immersive virtual environments for medical training. *Seminars in Colon and Rectal Surgery*, *19*(2), 90-97. https://doi.org/10.1053/j.scrs.2008.02.005

Brunyé, T. T., Drew, T., Weaver, D. L., & Elmore, J. G. (2019). A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications*, *4*, 7. https://doi.org/10.1186/s41235-019-0159-2

Cardós-Alonso, M. C., Inzunza, M., Gyllencreutz, L., Espinosa, S., Vázquez, T., Fernandez, M. A., Blanco, A., & Cintora-Sanz, A. M. (2024). Use of Self-Efficacy Scale in Mass Casualty Incidents During Drill Exercises. *BMC Health Services Research*, *24*, 745. https://doi.org/10.1186/s12913-024-11175-w

Centre for Research on the Epidemiology of Disasters. (2020). *Natural Disasters 2019*. https://cred.be/sites/default/files/adsr_2019.pdf

Chan, S. S. S., Chan, W. S., Cheng, Y. J., Fung, O. W. M., Lai, T. K. H., Leung, A. W. K., Leung, K. L. K., Li, S. J., Yip, A. L. K., & Pang, S. M. C. (2010). Development and evaluation of an undergraduate training course for developing International council of nurses disaster nursing competencies in China. *Journal of Nursing Scholarship*, *42*(4), 405-413. https://doi.org/10.1111/j.1547-5069.2010.01363.x

Chandra, A., Kim, J., Pieters, H. C., Tang, J., McCreary, M., Schreiber, M., & Wells, K. (2014). Implementing psychological first-aid training for medical reserve corps volunteers. *Disaster Medicine and Public Health Preparedness*, *8*(1), 95-100. https://doi.org/10.1017/dmp.2013.112

Chaput, C. J., Deluhery, M. R., Stake, C. E., Martens, K. A., & Cichon, M. E. (2007). Disaster training for prehospital providers. *Prehospital Emergency Care*, *11*(4), 458-465. https://doi.org/10.1080/00207450701537076

Choi, J., Thompson, C. E., Choi, J., Waddill, C. B., & Choi, S. (2022). Effectiveness of Immersive Virtual Reality in Nursing Education: Systematic Review. *Nurse Educator*, *47*(3). https://doi.org/10.1097/NNE.0000000000001117

Chou, W. K., Cheng, M. T., Lin, C. H., & Shih, F. Y. (2021). The effectiveness of functional exercises for teaching method disaster medicine to medical students. *Cureus*, *13*(5), e15151. https://doi.org/10.7759/cureus.15151

Cicero, M. X., Auerbach, M. A., Zigmont, J., Riera, A., Ching, K., & Baum, C. R. (2012). Simulation training with structured debriefing improves residents' pediatric disaster triage performance. *Prehospital and Disaster Medicine*, *27*(3), 239-244. https://doi.org/10.1017/s1049023x12000775

Cicero, M. X., Whitfill, T., Overly, F., Baird, J., Walsh, B., Yarzebski, J., Riera, A., Adelgais, K., Meckler, G. D., Baum, C., Cone, D. C., & Auerbach, M. (2017). Pediatric Disaster Triage: Multiple Simulation Curriculum Improves Prehospital Care Providers'

Assessment Skills. *Prehospital Emergency Care*, *21*(2), 201-208. https://doi.org/10.1080/10903127.2016.1235239

Clay, V., König, P., & Koenig, S. (2019). Eye tracking in virtual reality. *Journal of Eye Movement Research*, *12*(1). https://doi.org/10.16910/jemr.12.1.3

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Conrad, M., Kablitz, D., & Schumann, S. (2024). Learning effectiveness of immersive virtual reality in education and training: A systematic review of findings. *Computers & Education: X Reality*, *4*, 100053. https://doi.org/10.1016/j.cexr.2024.100053

Cowling, L., Swartzberg, K., & Groenewald, A. (2021). Knowledge retention and usefulness of simulation exercises for disaster medicine-what do specialty trainees know and think? *African Journal of Emergency Medicine*, *11*(3), 356-360. https://doi.org/10.1016/j.afjem.2021.05.001

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*(4), 174-203.

Currie, J., Bond, R. R., McCullagh, P., Black, P., Finlay, D. D., Gallagher, S., Kearney, P., Peace, A., Stoyanov, D., & Bicknell, C. D. (2019). Wearable technology-based metrics for predicting operator performance during cardiac catheterisation. *International Journal of Computer Assisted Radiology and Surgery*, *14*, 645-657. https://doi.org/10.1007/s11548-019-01918-0

Cuttance, G., Dansie, K., & Rayner, T. (2017). Paramedic application of a triage sieve: A paper-based exercise. *Prehospital and Disaster Medicine*, *32*(1), 3-13. https://doi.org/10.1017/s1049023x16001163

Davis, S., Nesbitt, K., & Nalivaiko, E. (2014). A systematic review of cybersickness. Proceedings of the 2014 conference on interactive entertainment. https://doi.org/10.1145/2677758.2677780

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*(5910), 66-69. https://doi.org/10.1126/science.1167311

Dittmar, M. S., Wolf, P., Bigalke, M., Graf, B. M., & Birkholz, T. (2018). Primary mass casualty incident triage: evidence for the benefit of yearly brief re-training from a simulation study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 26*(1), Article 35. https://doi.org/10.1186/s13049-018-0501-6

Dong, H., Schafer, J., & Ganz, A. (2016). Augmented reality based mass casualty incident training system. 2016 IEEE Symposium on Technologies for Homeland Security (HST). https://doi.org/10.1109/THS.2016.7568946

Dresing, T., & Pehl, T. (2018). *Praxisbuch Interview, Transkription & Analyse. Anleitungen und Regelsysteme für qualitativ Forschende.* (8 ed.). Eigenverlag.

Düking, P., Holmberg, H.-C., & Sperlich, B. (2018). The potential usefulness of virtual reality systems for athletes: a short SWOT analysis. *Frontiers in Physiology, 9*, Article 128. https://doi.org/10.3389/fphys.2018.00128

Dwivedi, Y. K., Hughes, L., Baabdullah, A. M., Ribeiro-Navarrete, S., Giannakis, M., Al-Debei, M. M., Dennehy, D., Metri, B., Buhalis, D., & Cheung, C. M. (2022). Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *66*, 102542. https://doi.org/10.1016/j.ijinfomgt.2022.102542

Edinger, Z. S., Powers, K. A., Jordan, K. S., & Callaway, D. W. (2019). Evaluation of an online educational intervention to increase knowledge and self-efficacy in disaster responders and critical care transporters caring for individuals with developmental disabilities. *Disaster Medicine and Public Health Preparedness*, *13*(4), 677-681. https://doi.org/10.1017/dmp.2018.129

Ellebrecht, N. (2013). The reality of triage: Results of a questionnaire on triage training and MCI experience of emergency physicians and paramedics. *Notfall + Rettungsmedizin*, *16*, 369-376. https://doi.org/10.1007/s10049-013-1726-6

Elsenbast, C., Dahlmann, P., & Schnier, D. (2024). Virtual team training with Mixed Reality and Virtual Reality–benefits and limitations illustrated on the example of two paramedic classes. *Multimedia Tools and Applications*, *83*, 63871–63895. https://doi.org/10.1007/s11042-023-17878-2

Evans, D. (2003). Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, *12*(1), 77-84. https://doi.org/10.1046/j.1365-2702.2003.00662.x

Farnsworth, B. (2024). *Eye Tracking: The Complete Pocket Guide*. Retrieved 4th Oct 2024 from https://imotions.com/blog/learning/best-practice/eye-tracking/

Farra, S., Miller, E., Timm, N., & Schafer, J. (2013). Improved Training for Disasters Using 3-D Virtual Reality Simulation. *Western Journal of Nursing Research*, *35*(5), 655-671. https://doi.org/10.1177/0193945912471735

Felnhofer, A., Kothgassner, O. D., Beutl, L., Hlavacs, H., & Kryspin-Exner, I. (2012). Is virtual reality made for men only? Exploring gender differences in the sense of presence. *Proceedings of the International Society on Presence Research*, 103-112.

Fernandez-Pacheco, A. N., Rodriguez, L. J., Price, M. F., Perez, A. B. G., Alonso, N. P., & Rios, M. P. (2017). Drones at the service for training on mass casualty incident: A simulation study. *Medicine*, *96*(26), Article e7159. https://doi.org/10.1097/md.0000000000007159

Ferrandini Price, M., Escribano Tortosa, D., Nieto Fernandez-Pacheco, A., Perez Alonso, N., Cerón Madrigal, J. J., Melendreras-Ruiz, R., García-Collado Á, J., Pardo Rios, M., & Juguera Rodriguez, L. (2018). Comparative study of a simulated incident with multiple victims and immersive virtual reality. *Nurse Education Today*, *71*, 48-53. https://doi.org/10.1016/j.nedt.2018.09.006

Finch, H. (2006). Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated. *Methodology*, *1*(1), 27-38. https://doi.org/10.1027/1614-1881.1.1.27

Foronda, C. L., Shubeck, K., Swoboda, S. M., Hudson, K. W., Budhathoki, C., Sullivan, N., & Hu, X. G. (2016). Impact of Virtual Simulation to Teach Concepts of Disaster Triage. *Clinical Simulation in Nursing*, *12*(4), 137-144. https://doi.org/10.1016/j.ecns.2016.02.004

Franc, J. M., Verde, M., Gallardo, A. R., Carenzo, L., & Ingrassia, P. L. (2017). An Italian version of the Ottawa Crisis Resource Management Global Rating Scale: a reliable and valid tool for assessment of simulation performance. *Internal and Emergency Medicine*, *12*(5), 651-656. https://doi.org/10.1007/s11739-016-1486-7

Furseth, P. A., Taylor, B., & Kim, S. C. (2016). Impact of interprofessional education among nursing and paramedic students. *Nurse Educator*, *41*(2), 75-79. https://doi.org/10.1097/nne.0000000000000219

Fusco, A., & Tieri, G. (2022). Challenges and perspectives for clinical applications of immersive and non-immersive virtual reality. *Journal of Clinical Medicine*, *11*(15), 4540. https://doi.org/10.3390/jcm11154540

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, *23*, 523-552. https://doi.org/10.1007/s10648-011-9174-7

Gerwann, S., Baetzner, A. S., & Hill, Y. (2024). Immersive virtual reality and augmented virtuality in sport and performance psychology: opportunities, current limitations and practical guidelines. *Sport, Exercise, and Performance Psychology*. https://doi.org/10.1037/spy0000367

Giessing, L. (2021). The Potential of Virtual Reality for Police Training Under Stress: A SWOT Analysis. In E. Arble & B. Arnetz (Eds.), *Interventions, Training, and Technologies for Improved Police Well-Being and Performance* (pp. 102-124). IGI Global. https://doi.org/10.4018/978-1-7998-6820-0.ch006

Giessing, L., & Frenkel, M. O. (2022). Virtuelle Realität als vielversprechende Ergänzung im polizeilichen Einsatztraining – Chancen, Grenzen und Implementierungsmöglichkeiten. In M. S. Staller & S. Körner (Eds.), *Handbuch polizeiliches Einsatztraining: Professionelles Konfliktmanagement - Theorie, Trainingskonzepte und Praxiserfahrungen* (pp. 677-692). Springer Gabler.

Giessing, L., Frenkel, M. O., Zinner, C., Rummel, J., Nieuwenhuys, A., Kasperk, C., Brune, M., Engel, F. A., & Plessner, H. (2019). Effects of coping-related traits and psychophysiological stress responses on police recruits' shooting behavior in reality-based scenarios. *Frontiers in Psychology*, *10*, Article 1523. https://doi.org/10.3389/fpsyg.2019.01523

Giessing, L., Oudejans, R. R., Hutter, V., Plessner, H., Strahler, J., & Frenkel, M. O. (2020). Acute and chronic stress in daily police service: A three-week N-of-1 study. *Psychoneuroendocrinology*, *122*, Article 104865. https://doi.org/10.1016/j.psyneuen.2020.104865

Girau, E., Mura, F., Bazurro, S., Casadio, M., Chirico, M., Solari, F., & Chessa, M. (2019). A mixed reality system for the simulation of emergency and first-aid scenarios. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany. https://doi.org/10.1109/EMBC.2019.8856777

Goniewicz, K., Goniewicz, M., Burkle, F. M., & Khorram-Manesh, A. (2021). Cohort research analysis of disaster experience, preparedness, and competency-based training among nurses. *Plos One*, *16*(1), Article e0244488. https://doi.org/10.1371/journal.pone.0244488

Gordon, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine*, *66*(12), 762-769.

Gow, K. W. (2013). Self-evaluation: how well do surgery residents judge performance on a rotation? *The American Journal of Surgery*, *205*(5), 557-562. https://doi.org/10.1016/j.amjsurg.2013.01.010

Greco, S., Lewis, E. J., Sanford, J., Sawin, E. M., & Ames, A. (2019). Ethical reasoning debriefing in disaster simulations. *Journal of Professional Nursing*, *35*(2), 124-132. https://doi.org/10.1016/j.profnurs.2018.09.004

Guitiérrez, F., Pierce, J., Vergara, V. M., Coulter, R., Saland, L., Caudell, T. P., Goldsmith, T. E., & Alverson, D. C. (2007). The effect of degree of immersion upon learning performance in virtual reality simulations for medical education. In J. D. Westwood (Ed.), *Medicine Meets Virtual Reality* (Vol. 15, pp. 155-160). IOS Press.

Heldring, S., Jirwe, M., Wihlborg, J., Berg, L., & Lindström, V. (2024). Using High-Fidelity Virtual Reality for Mass-Casualty Incident Training by First Responders–A Systematic Review of the Literature. *Prehospital and Disaster Medicine*, *39*(1), 94-105. https://doi.org/10.1017/S1049023X24000049

Hill, Y., Kiefer, A. W., Oudejans, R. R., Baetzner, A. S., & Den Hartigh, R. J. (2024). Adaptation to stressors: Hormesis as a framework for human performance. *New Ideas in Psychology*, *73*, Article 101073. https://doi.org/10.1016/j.newideapsych.2024.101073

Hill, Y., Kiefer, A. W., Silva, P. L., Van Yperen, N. W., Meijer, R. R., Fischer, N., & Den Hartigh, R. J. R. (2020). Antifragility in Climbing: Determining Optimal Stress Loads for Athletic Performance Training. *Frontiers in Psychology*, *11*, Article 272. https://doi.org/10.3389/fpsyg.2020.00272

Hill, Y., Van Yperen, N. W., & Den Hartigh, R. J. (2021). Facing repeated stressors in a motor task: does it enhance or diminish resilience? *Journal of Motor Behavior*, *53*(6), 717-726. https://doi.org/10.1080/00222895.2020.1852155

Hogarth, R. M. (2001). *Educating Intuition*. University of Chicago Press.

Hogarth, R. M. (2010). Intuition: A challenge for psychological research on decision making. *Psychological Inquiry*, *21*(4), 338-353. https://doi.org/10.1080/1047840X.2010.520260

Hugelius, K., Becker, J., & Adolfsson, A. (2020). Five challenges when managing mass casualty or disaster situations: a review study. *International Journal of Environmental Research and Public Health*, *17*(9), Article 3068. https://doi.org/10.3390/ijerph17093068

Huh, S. S., & Kang, H. Y. (2019). Effects of an educational program on disaster nursing competency. *Public Health Nursing*, *36*(1), 28-35. https://doi.org/10.1111/phn.12557

Hutchins, S. G., & Timmons, R. P. (2006). Radio Interoperability: Addressing the Real Reason We Don't Communicate Well During Emergencies. *Calhoun: The NPS Institutional Archive*.

Hutchinson, S. W., Haynes, S., Parker, P., Dennis, B., McLin, C., & Welldaregay, W. (2011). Implementing a multidisciplinary disaster simulation for undergraduate nursing students. *Nurs Educ Perspect*, *32*(4), 240-243. https://doi.org/10.5480/1536-5026-32.4.240

Hutter, R. I., Renden, P. G., Kok, M., Oudejans, R., Koedijk, M., & Kleygrewe, L. (2023). Criteria for the High Quality Training of Police Officers. In M. S. Staller, S. Koerner, & B. Zaiser (Eds.), *Police Conflict Management, Volume II: Training and Education* (pp. 7-32). Springer. https://doi.org/10.1007/978-3-031-41100-7_2

Ignacio, J., Dolmans, D., Scherpbier, A., Rethans, J.-J., Chan, S., & Liaw, S. Y. (2016). Stress and anxiety management strategies in health professions' simulation training: a review of the literature. *BMJ Simulation & Technology Enhanced Learning*, *2*(2), 42-46. https://doi.org/10.1136/bmjstel-2015-000097

iMotions. (2022). *iMotions A/S, (Version 9.3), Copenhagen, Denmark.* https://imotions.com/

Ingrassia, P. L., Colombo, D., Barra, F. L., Carenzo, L., Franc, J., & Della Corte, F. (2013). Impact of training in medical disaster management: a pilot study using a new tool for live simulation. *Emergencias*, *25*(459), 459-466.

Ingrassia, P. L., Foletti, M., Djalali, A., Scarone, P., Ragazzoni, L., Della Corte, F., Kaptan, K., Lupescu, O., Arculeo, C., von Arnim, G., Friedl, T., Ashkenazi, M., Heselmann, D., Hreckovski, B., Khorrram-Manesh, A., Komadina, R., Lechner, K., Patru, C., Burkle, F. M., & Fisher, P. (2014). Education and training initiatives for crisis management in the European Union: A web-based analysis of available programs. *Prehospital and Disaster Medicine*, *29*(2), 115-126. https://doi.org/10.1017/s1049023x14000235

Ingrassia, P. L., Ragazzoni, L., Carenzo, L., Colombo, D., Gallardo, A. R., & Della Corte, F. (2015). Virtual reality and live simulation: a comparison between two simulation tools for assessing mass casualty triage skills. *European Journal of Emergency Medicine*, *22*(2), 121-127. https://doi.org/10.1097/mej.0000000000000132

Ingrassia, P. L., Ragazzoni, L., Tengattini, M., Carenzo, L., & Della Corte, F. (2014). Nationwide Program of Education for Undergraduates in the Field of Disaster Medicine: Development of a Core Curriculum Centered on Blended Learning and

Simulation Tools. *Prehospital and Disaster Medicine*, *29*(5), 508-515. https://doi.org/10.1017/s1049023x14000831

James, L. S., Williams, M. L., Camel, S. P., & Slagle, P. (2021). Nursing student's attitudes toward teams in an undergraduate interprofessional mass casualty simulation. *Nursing Forum*, *56*(3), 500-512. https://doi.org/10.1111/nuf.12570

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, *23*, 1515-1529. https://doi.org/10.1007/s10639-017-9676-0

Jones, J., Kue, R., Mitchell, P., Eblan, G., & Dyer, K. S. (2014). Emergency medical services response to active shooter incidents: provider comfort level and attitudes before and after participation in a focused response training program. *Prehospital and Disaster Medicine*, *29*(4), 350-357. https://doi.org/10.1017/s1049023x14000648

Jongbloed, J., Chaker, R., & Lavoué, E. (2024). Immersive procedural training in virtual reality: A systematic literature review. *Computers & Education*, *221*, Article 105124. https://doi.org/10.1016/j.compedu.2024.105124

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review*, *87*(4), 329-354.

Kahneman, D., & Klein, G. (2009). Conditions for Intuitive Expertise: A Failure to Disagree. *American psychologist*, *64*(6), 515-526. https://doi.org/10.1037/a0016755

Kar, N. (2024). Impact of disasters on the disaster responders: A review of stress, coping, resilience, and supportive methods. *Odisha Journal of Psychiatry*, *19*(2), 37-47. https://doi.org/10.4103/OJP.OJP_23_23

Kent, J. A., & Hughes, C. E. (2022). Law enforcement training using simulation for locally customized encounters. *Frontiers in Virtual Reality*, *3*, 960146. https://doi.org/10.3389/frvir.2022.960146

Kiefer, A. W., Silva, P. L., Harrison, H. S., & Araújo, D. (2018). Antifragility in sport: Leveraging adversity to enhance performance. *Sport, Exercise, and Performance Psychology*, *7*(4), 342-350. https://doi.org/10.1037/spy0000130

Kim, J., & Lee, O. (2020). Effects of a simulation-based education program for nursing students responding to mass casualty incidents: A pre-post intervention study. *Nurse Education Today*, *85*, Article 104297. https://doi.org/10.1016/j.nedt.2019.104297

Kleygrewe, L., Hutter, R. I. V., Koedijk, M., & Oudejans, R. R. D. (2024a). Changing perspectives: enhancing learning efficacy with the after-action review in virtual reality training for police. *Ergonomics*, *67*(5), 628-637. https://doi.org/10.1080/00140139.2023.2236819

Kleygrewe, L., Hutter, R. V., Koedijk, M., & Oudejans, R. R. (2024b). Virtual reality training for police officers: A comparison of training responses in VR and real-life training. *Police Practice and Research*, *25*(1), 18-37.

Kman, N. E., Price, A., Berezina-Blackburn, V., Patterson, J., Maicher, K., Way, D. P., McGrath, J., Panchal, A. R., Luu, K., & Oliszewski, A. (2023). First responder virtual reality simulator to train and assess emergency personnel for mass casualty response. *Journal of the American College of Emergency Physicians Open*, *4*(1), Article e12903. https://doi.org/10.1002/emp2.12903

Knight, J. F., Carley, S., Tregunna, B., Jarvis, S., Smithies, R., de Freitas, S., Dunwell, I., & Mackway-Jones, K. (2010). Serious gaming technology in major incident triage training: a pragmatic controlled trial. *Resuscitation*, *81*(9), 1175-1179. https://doi.org/10.1016/j.resuscitation.2010.03.042

Koca, B., & Arkan, G. (2020). The effect of the disaster management training program among nursing students. *Public Health Nursing*, *37*(5), 769-777. https://doi.org/10.1111/phn.12760

Koutitas, G., Smith, S., & Lawrence, G. (2021). Performance evaluation of AR/VR training technologies for EMS first responders. *Virtual Reality*, *25*(1), 83-94. https://doi.org/10.1007/s10055-020-00436-8

Kuhls, D. A., Chestovich, P. J., Coule, P., Carrison, D. M., Chua, C. M., Wora-Urai, N., & Kanchanarin, T. (2017). Basic disaster life support (BDLS) training improves first responder confidence to face mass-casualty incidents in Thailand. *Prehospital and Disaster Medicine*, *32*(5), 492-500. https://doi.org/10.1017/s1049023x17006550

Lampi, M., Vikstrom, T., & Jonson, C. O. (2013). Triage performance of Swedish physicians using the ATLS algorithm in a simulated mass casualty incident: a prospective cross-sectional survey. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *21*, Article 90. https://doi.org/10.1186/1757-7241-21-90

Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., & Yan, S. (2012). Depth matters: Influence of depth cues on visual saliency. Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12. https://doi.org/10.1007/978-3-642-33709-3_8

Lateef, F. (2010). Simulation-based learning: Just like the real thing. *Journal of Emergencies, Trauma, and Shock*, *3*(4), 348-352. https://doi.org/10.4103/0974-2700.70743

LeBlanc, V. R. (2009). The effects of acute stress on performance: implications for health professions education. *Academic Medicine*, *84*(10), 25-33. https://doi.org/10.1097/ACM.0b013e3181b37b8f

Lennquist Montán, K., Örtenwall, P., & Lennquist, S. (2015). Assessment of the accuracy of the Medical Response to Major Incidents (MRMI) course for interactive training of the response to major incidents and disasters. *American Journal of Emergency Medicine*, *10*(2), 93-107. https://doi.org/10.5055/ajdm.2015.0194

Li, G., Anguera, J. A., Javed, S. V., Khan, M. A., Wang, G., & Gazzaley, A. (2020). Enhanced attention using head-mounted virtual reality. *Journal of Cognitive Neuroscience*, *32*(8), 1438-1454. https://doi.org/10.1162/jocn_a_01560

Liaw, S. Y., Scherpbier, A., Rethans, J.-J., & Klainin-Yobas, P. (2012). Assessment for simulation learning outcomes: a comparison of knowledge and self-reported confidence with observed clinical performance. *Nurse Education Today*, *32*(6), 35-39. https://doi.org/10.1016/j.nedt.2011.10.006

Liblik, K., Byun, J., Lloyd-Kuzik, A., Farina, J. M., Burgos, L. M., Howes, D., & Baranchuk, A. (2023). The DIVERSE Study: Determining the Importance of Various gEnders, Races, and Body Shapes for CPR Education Using Manikins. *Current Problems in Cardiology*, *48*(7), 101159. https://doi.org/10.1016/j.cpcardiol.2022.101159

Lim, G. H., Lim, B. L., & Vasu, A. (2013). Survey of Factors Affecting Health Care Workers' Perception Towards Institutional and Individual Disaster Preparedness. *Prehospital and Disaster Medicine*, *28*(4), 353-358. https://doi.org/10.1017/s1049023x1300349x

Lomaglio, L., Ansaloni, L., Catena, F., Sartelli, M., & Coccolini, F. (2020). Mass casualty incident: definitions and current reality. In Y. Kluger, F. Coccolini, F. Catena, & L. Ansaloni (Eds.), *WSES Handbook of Mass Casualties Incidents Management*. Springer. https://doi.org/10.1007/978-3-319-92345-1_1

Loth, S., Cote, A. C., Shaafi Kabiri, N., Bhangu, J. S., Zumwalt, A., Moss, M., & Thomas, K. (2019). Improving triage accuracy in first responders: measurement of short structured protocols to improve identification of salient triage features. *World Medical & Health Policy*, *11*(2), 163-176. https://doi.org/10.1002/wmh3.306

Lowe, J., Peng, C., Winstead-Derlega, C., & Curtis, H. (2020). 360 virtual reality pediatric mass casualty incident: A cross sectional observational study of triage and out-of-hospital intervention accuracy at a national conference. *Journal of the American*

College of Emergency Physicians Open, *1*(5), 974-980. https://doi.org/10.1002/emp2.12214

Lowes, A., & Cosgrove, J. (2016). Prehospital organization and management of a mass casualty incident. *BJA Education*, *16*(10), 323-328. https://doi.org/10.1093/bjaed/mkw005

Ma, D. H., Shi, Y. X., Zhang, G., & Zhang, J. (2021). Does theme game-based teaching promote better learning about disaster nursing than scenario simulation: A randomized controlled trial. *Nurse Education Today*, *103*, Article 104923. https://doi.org/10.1016/j.nedt.2021.104923

Mahmoud, K., Harris, I., Yassin, H., Hurkxkens, T. J., Matar, O. K., Bhatia, N., & Kalkanis, I. (2020). Does immersive VR increase learning gain when compared to a non-immersive VR learning experience? In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies. Human and Technology Ecosystems. HCII 2020. Lecture Notes in Computer Science* (Vol. 12206, pp. 480-498). Springer. https://doi.org/10.1007/978-3-030-50506-6_33

Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior*, *72*, 276-285. https://doi.org/10.1016/j.chb.2017.02.066

Makransky, G., & Petersen, G. B. (2021). The cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, *33*(3), 937-958. https://doi.org/10.1007/s10648-020-09586-2

Manser, T. (2009). Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. *Acta Anaesthesiologica Scandinavica*, *53*(2), 143-151. https://doi.org/10.1111/j.1399-6576.2008.01717.x

Mao, R. Q., Lan, L., Kay, J., Lohre, R., Ayeni, O. R., Goel, D. P., & Sa, D. d. (2021). Immersive Virtual Reality for Surgical Training: A Systematic Review. *Journal of Surgical Research*, *268*, 40-58. https://doi.org/10.1016/j.jss.2021.06.045

Markou-Pappas, N., Lamine, H., Ragazzoni, L., & Caviglia, M. (2024). Key performance indicators in pre-hospital response to disasters and mass casualty incidents: a scoping review. *European Journal of Trauma and Emergency Surgery*, *50*, 2029–2037. https://doi.org/10.1007/s00068-024-02533-8

Marlow, S., Bisbey, T., Lacerenza, C., & Salas, E. (2018). Performance measures for health care teams: A review. *Small Group Research*, *49*(3), 306-356. https://doi.org/10.1177/1046496417748196

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to Qualitative Research in Mathematics Education. Advances in Mathematics Education* (pp. 365-380). Springer. https://doi.org/10.1007/978-94-017-9181-6_13

Mayring, P. (2022). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (13th ed. ed.). Beltz. https://www.content-select.com/index.php?id=bib_view&ean=9783407258991

Mazzeo, M., Chewning-Kulick, M., Pike, W., Cartwright, J., Rovira, E., & Thomson, R. (2021). Detecting Hesitation During Battlefield Wound Treatment on Female Soldiers. In J. Kalra, N. J. Lightner, & R. Taiar (Eds.), *Advances in Human Factors and Ergonomics in Healthcare and Medical Devices* (pp. 329-336). Springer International Publishing. https://doi.org/10.1007/978-3-030-80744-3_41

McEwan, D., Ruissen, G. R., Eys, M. A., Zumbo, B. D., & Beauchamp, M. R. (2017). The effectiveness of teamwork training on teamwork behaviors and team performance: a systematic review and meta-analysis of controlled interventions. *Plos One*, *12*(1), Article e0169604. https://doi.org/10.1371/journal.pone.0169604

McMahan, R. P., Lai, C., & Pal, S. K. (2016, 2016//). Interaction Fidelity: The Uncanny Valley of Virtual Reality Interactions. Virtual, Augmented and Mixed Reality, Cham.

Merlin, M. A., Moon, J., Krimmel, J., Liu, J., & Marques-Baptista, A. (2010). Improving medical students' understanding of prehospital care through a fourth year emergency medicine clerkship. *Emergency Medicine Journal*, *27*(2), 147-150. https://doi.org/10.1136/emj.2008.066654

Mikropoulos, T. A., & Natsis, A. (2011). Educational virtual environments: A ten-year review of empirical research (1999–2009). *Computers & Education*, *56*(3), 769-780. https://doi.org/10.1016/j.compedu.2010.10.020

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, *77*(12), 1321-1329.

Mills, B., Dyks, P., Hansen, S., Miles, A., Rankin, T., Hopper, L., Brook, L., & Bartlett, D. (2020). Virtual reality triage training can provide comparable simulation efficacy for paramedicine students compared to live simulation-based scenarios. *Prehospital Emergency Care*, *24*(4), 525-536. https://doi.org/10.1080/10903127.2019.1676345

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1). https://doi.org/10.1186/2046-4053-4-1

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review*, *120*(3), 544-560. https://doi.org/10.1037/a0032459

Moore, J. W. (2016). What is the sense of agency and why does it matter? *Frontiers in Psychology*, *7*, 1272. https://doi.org/10.3389/fpsyg.2016.01272

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98-100. https://doi.org/10.1109/MRA.2012.2192811

Motola, I., Burns, W. A., Brotons, A. A., Withum, K. F., Rodriguez, R. D., Hernandez, S., Rivera, H. F., Issenberg, S. B., & Schulman, C. I. (2015). Just-in-time learning is effective in helping first responders manage weapons of mass destruction events. *Journal of Trauma and Acute Care Surgery*, *79*(4), 152-156. https://doi.org/10.1097/ta.0000000000000570

Murray, N. P., Lewinski, W., Sandri Heidner, G., Lawton, J., & Horn, R. (2024). Gaze Control and Tactical Decision-Making Under Stress in Active-Duty Police Officers During a Live Use-of-Force Response. *Journal of Motor Behavior*, *56*(1), 30-41. https://doi.org/10.1080/00222895.2023.2229946

Murtinger, M., Jaspaert, E., Schrom-Feiertag, H., & Egger-Lampl, S. (2021). CBRNe training in virtual environments: SWOT analysis & practical guidelines. *International Journal of Safety and Security Engineering*, *11*(4), 295-303. https://doi.org/10.18280/ijsse.110402

Neher, A. N., Wespi, R., Rapphold, B. D., Sauter, T. C., Kämmer, J. E., & Birrenbach, T. (2024). Interprofessional Team Training With Virtual Reality: Acceptance, Learning Outcome, and Feasibility Evaluation Study. *JMIR Serious Games*, *12*, Article e57117. https://doi.org/10.2196/57117

Nguyen, H., & Bednarz, T. (2020, 2020 Nov 25-27). User Experience in Collaborative Extended Reality: Overview Study. In P. Bourdot, V. Interrante, R. Kopper, A.-H. Olivier, H. Saito, & G. Zachmann, *Virtual Reality and Augmented Reality* 17th EuroVR International Conference, Valencia, Spain. https://doi.org/10.1007/978-3-030-62655-6_3 (Virtual Reality and Augmented Reality)

Nguyen, T.-H. D., El-Nasr, M. S., & Isaacowitz, D. M. (2017). Interactive visualization for understanding of attention patterns. In H.-C. Hege, D. Hoffman, C. R. Johnson, K. Polthier, & M. Rump (Eds.), *Eye Tracking and Visualization: Foundations, Techniques, and Applications.* (pp. 23-39). Springer. https://doi.org/10.1007/978-3-319-47024-5_2

Nizam, S. M., Abidin, R. Z., Hashim, N. C., Lam, M. C., Arshad, H., & Majid, N. (2018). A Review of Multimodal Interaction Technique in Augmented Reality Environment. *International Journal of Advanced Science and Engineering Information Technology*, *8*(4-2), 1460-1469.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan - a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), Article 210. https://doi.org/10.1186/s13643-016-0384-4

Paddock, M. T., Bailitz, J., Horowitz, R., Khishfe, B., Cosby, K., & Sergel, M. J. (2015). Disaster response team FAST skills training with a portable ultrasound simulator compared to traditional training: pilot study. *Western Journal of Emergency Medicine*, *16*(2), 325-330. https://doi.org/10.5811/westjem.2015.1.23720

Pagano, R. R. (2013). *Understanding Statistics in the Behavioural Sciences* (10 ed.). Cengage Learning.

Paletta, L., Pszeida, M., Schneeberger, M., Dini, A., Reim, L., & Kallus, K. W. (2022). Cognitive-emotional Stress and Risk Stratification of Situational Awareness in Immersive First Responder Training. 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). https://doi.org/10.1109/BHI56158.2022.9926805

Parra, M. A., & Kaplan, R. I. (2019). Predictors of performance in real and virtual scenarios across age. *Experimental Aging Research*, *45*(2), 180-198. https://doi.org/10.1080/0361073X.2019.1586106

Perry, O., Jaffe, E., & Bitan, Y. (2021). Dynamic Communication Quantification Model for Measuring Information Management During Mass-Casualty Incident Simulations. *Human Factors*, *64*(1), 228-249. https://doi.org/10.1177/00187208211018880

Phattharapornjaroen, P., Glantz, V., Carlstrom, E., Holmqvist, L. D., & Khorram-Manesh, A. (2020). Alternative leadership in flexible surge capacity-the perceived impact of tabletop simulation exercises on Thai emergency physicians capability to manage a major incident. *Sustainability*, *12*(15), Article 6216. https://doi.org/10.3390/su12156216

Pinder, R. A., Davids, K., Renshaw, I., & Araújo, D. (2011). Representative learning design and functionality of research and practice in sport. *Journal of Sport and Exercise Psychology*, *33*(1), 146-155. https://doi.org/10.1123/jsep.33.1.146

Pollard, K. A., Bachmann, D. J., Greer, M., Way, D. P., & Kman, N. E. (2015). Development of a disaster preparedness curriculum for medical students: a pilot study of incorporating local events into training opportunities. *American Journal of Emergency Medicine*, *10*(1), 51-59. https://doi.org/10.5055/ajdm.2015.0188

Pottle, J. (2019). Virtual reality and the transformation of medical education. *Future Healthcare Journal*, *6*(3), 181-185. https://doi.org/10.7861/fhj.2019-0036

Pouraghaei, M., Sadegh Tabrizi, J., Moharamzadeh, P., Rajaei Ghafori, R., Rahmani, F., & Najafi Mirfakhraei, B. (2017). The effect of start triage education on knowledge and practice of emergency medical technicians in disasters. *J Caring Sci*, *6*(2), 119-125. https://doi.org/10.15171/jcs.2017.012

Prachyabrued, M., Wattanadhirach, D., Dudrow, R. B., Krairojananan, N., & Fuengfoo, P. (2019). Toward virtual stress inoculation training of prehospital healthcare personnel: A stress-inducing environment design and investigation of an emotional connection factor. 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). https://doi.org/10.1109/VR.2019.8797705

Price, M. F., Serrano, A. M., Fernández-Pacheco, A. N., Ruiz, R. M., Collado, Á. J. G., & Ríos, M. P. (2022). Lack of a standard visual attention pattern for use when health professionals triage mass casualty incidents is revealed by Tobbi eye tracking technology. *Emergencias*, *34*, 230-237.

Prytz, E. G., Norén, C., & Jonson, C.-O. (2018). Fixation differences in visual search of accident scenes by novices and expert emergency responders. *Human Factors*, *60*(8), 1219-1227. https://doi.org/10.1177/00187208187881

Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, *46*, 175-208.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510-532. https://doi.org/10.1037/0033-2909.114.3.510

Ricci, S., Calandrino, A., Borgonovo, G., Chirico, M., & Casadio, M. (2022). Virtual and augmented reality in basic and advanced life support training. *JMIR Serious Games*, *10*(1), Article e28595. https://doi.org/10.2196/28595

Ripoll-Gallardo, A., Ragazzoni, L., Mazzanti, E., Meneghetti, G., Franc, J. M., Costa, A., & della Corte, F. (2020). Residents working with Medecins Sans Frontieres: training and pilot evaluation. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *28*(1), Article 86. https://doi.org/10.1186/s13049-020-00778-x

Rivkind, A. I., Faroja, M., Mintz, Y., Pikarsky, A. J., Zamir, G., Elazary, R., Abu-Gazala, M., & Bala, M. (2015). Combating terror: A new paradigm in student trauma education. *Journal of Trauma and Acute Care Surgery*, *78*(2), 415-421. https://doi.org/10.1097/ta.0000000000000508

Rodgers, D. L., Bhanji, F., & McKee, B. R. (2010). Written evaluation is not a predictor for skills performance in an Advanced Cardiovascular Life Support course. *Resuscitation*, *81*(4), 453-456. https://doi.org/10.1016/j.resuscitation.2009.12.018

Saiboon, I. M., Zahari, F., Isa, H. M., Sabardin, D. M., & Robertson, C. E. (2021). E-Learning in teaching emergency disaster response among undergraduate medical students in Malaysia. *Front Public Health*, *9*, 628178. https://doi.org/10.3389/fpubh.2021.628178

Salas, E., Rosen, M. A., Held, J. D., & Weissmuller, J. J. (2009). Performance measurement in simulation-based training: A review and best practices. *Simulation & Gaming*, *40*(3). https://doi.org/10.1177/1046878108326734

Scarfe, P., & Glennerster, A. (2019). The science behind virtual reality displays. *Annual Review of Vision Science*, *5*(1), 529-547. https://doi.org/10.1146/annurev-vision-091718-014942

Schmid, B., Sauer, F., & Busch, H.-J. (2022). Initial preclinical assessment on-site. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, *65*, 979–986. https://doi.org/10.1007/s00103-022-03582-3

Schmidt, H. G., & Boshuizen, H. P. (1993). On acquiring expertise in medicine. *Educational Psychology Review*, *5*, 205-221. https://doi.org/10.1007/BF01323044

Schöne, B., Kisker, J., Lange, L., Gruber, T., Sylvester, S., & Osinsky, R. (2023). The Reality of Virtual Reality. *Frontiers in Psychology*, *14*, 1093014. https://doi.org/10.3389/fpsyg.2023.1093014

Schwind, V., Wolf, K., & Henze, N. (2018). Avoiding the uncanny valley in virtual character design. *Interactions*, *25*(5), 45-49. https://doi.org/10.1145/3236673

Scott, J. A., Miller, G. T., Barry Issenberg, S., Brotons, A. A., Gordon, D. L., Gordon, M. S., McGaghie, W. C., & Petrusa, E. R. (2006). Skill improvement during emergency response to terrorism training. *Prehospital Emergency Care*, *10*(4), 507-514. https://doi.org/10.1080/10903120600887072

Scott, L. A., Carson, D. S., & Greenwell, I. B. (2010). Disaster 101: a novel approach to disaster medicine training for health professionals. *Journal of Emergency Medicine*, *39*(2), 220-226. https://doi.org/10.1016/j.jemermed.2009.08.064

Sena, A., Forde, F., Yu, C., Sule, H., & Masters, M. M. (2021). Disaster preparedness training for emergency medicine residents using a tabletop exercise. *MedEdPORTAL*, *17*, Article 11119. https://doi.org/10.15766/mep_2374-8265.11119

Servotte, J.-C., Goosse, M., Campbell, S. H., Dardenne, N., Pilote, B., Simoneau, I. L., Guillaume, M., Bragard, I., & Ghuysen, A. (2020). Virtual reality experience: Immersion, sense of presence, and cybersickness. *Clinical Simulation in Nursing*, *38*, 35-43. https://doi.org/10.1016/j.ecns.2019.09.006

Shinnick, M. A. (2016). Validating eye tracking as an objective assessment tool in simulation. *Clinical Simulation in Nursing*, *12*(10), 438-446. https://doi.org/10.1016/j.ecns.2016.06.001

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British journal of psychology*, *109*(3), 431-433. https://doi.org/10.1111/bjop.12305

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, *3*, 74. https://doi.org/10.3389/frobt.2016.00074

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *6*(6), 603-616. https://doi.org/10.1162/pres.1997.6.6.603

Smith, S., Farra, S., Dempsey, A., & Arms, D. (2015). Preparing nursing students for leadership using a disaster-related simulation. *Nurse Educator*, *40*(4), 212-216. https://doi.org/10.1097/nne.0000000000000143

Soola, A. H., Mehri, S., & Azizpour, I. (2022). Evaluation of the factors affecting triage decision-making among emergency department nurses and emergency medical technicians in Iran: a study based on Benner’s theory. *BMC Emergency Medicine*, *22*(1), Article 174. https://doi.org/10.1186/s12873-022-00729-y

Sorani, M., Tourani, S., Khankeh, H. R., & Panahi, S. (2018). Prehospital Emergency Medical Services Challenges in Disaster; a Qualitative Study. *Emergency*, *6*(1), Article e26.

Srivastava, K. (2010). Disaster: Challenges and perspectives. *Industrial Psychiatry Journal*, *19*(1). https://doi.org/10.4103/0972-6748.77623

Stein, C. (2018). Uncanny Valley in Virtual Reality. In N. Lee (Ed.), *Encyclopedia of Computer Graphics and Games* (pp. 1-3). Springer International Publishing. https://doi.org/10.1007/978-3-319-08234-9_177-1

Stratton, S. J. (2019). Quasi-experimental design (pre-test and post-test studies) in prehospital and disaster research. *Prehospital and Disaster Medicine*, *34*(6), 573-574. https://doi.org/10.1017/S1049023X19005053

Super, G., Groth, S., & Hook, R. (1994). START: simple triage and rapid treatment plan. *Newport Beach, CA: Hoag Memorial Presbyterian Hospital*, *199*.

Szabo, R. A., Forrest, K., Morley, P., Barwick, S., Bajaj, K., Britt, K., Yong, S. A., Park-Ross, J., Story, D., & Stokes-Parish, J. (2024). CPR training as a gender and rights-based healthcare issue. *Health Promotion International*, *39*(6), Article 156. https://doi.org/10.1093/heapro/daae156

Thompson, S. (2023). Mass casualty triage: using virtual reality in hazardous area response teams training. *Journal of Paramedic Practice*, *15*(10), 418-427. https://doi.org/10.12968/jpar.2023.15.10.418

Tin, D., Hertelendy, A. J., & Ciottone, G. R. (2021). Disaster medicine training: The case for virtual reality. *American Journal of Emergency Medicine*, *48*, 370-371. https://doi.org/10.1016/j.ajem.2021.01.085

Tufanaru, C., Munn, Z., Aromataris, E., Campbell, J., & Hopp, L. (2020). Chapter 3: Systematic reviews of effectiveness. In E. Aromataris & Z. Munn (Eds.), *JBI Manual for Evidence Synthesis*. JBI. https://doi.org/10.46658/JBIMES-20-04

Uhl, J. C., Gutierrez, R., Regal, G., Schrom-Feiertag, H., Schuster, B., & Tscheligi, M. (2024). Choosing the right reality: A comparative analysis of tangibility in immersive trauma simulations. Proceedings of the CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3613904.3641912

Uhl, J. C., Neundlinger, K., & Regal, G. (2023). Social presence as a training resource: comparing VR and traditional training simulations. *Research in Learning Technology*, *31*. https://doi.org/10.25304/rlt.v31.2827

Uhl, J. C., Nguyen, Q., Hill, Y., Murtinger, M., & Tscheligi, M. (2023). xHits: An Automatic Team Performance Metric for VR Police Training. 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE). https://doi.org/10.1109/MetroXRAINE58569.2023.10405600

Uhl, J. C., Schrom-Feiertag, H., Regal, G., Gallhuber, K., & Tscheligi, M. (2023). Tangible immersive trauma simulation: is mixed reality the next level of medical skills training? Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3544548.3581292

Ullén, F., Hambrick, D. Z., & Mosing, M. A. (2016). Rethinking expertise: A multifactorial gene–environment interaction model of expert performance. *Psychological Bulletin*, *142*(4), 427-446. https://doi.org/10.1037/bul0000033

Unver, V., Basak, T., Tastan, S., Kok, G., Guvenc, G., Demirtas, A., Ayhan, H., Köse, G., Iyigun, E., & Tosune, N. (2018). Analysis of the effects of high-fidelity simulation on nursing students' perceptions of their preparedness for disasters. *International Emergency Nursing*, *38*, 3-9. https://doi.org/10.1016/j.ienj.2018.03.002

Vaughan, S. R., Ballard, T., Ward-Demo, P., Vojta, L., Ahmed, A. E., & Costello, A. (2024). Evaluation of Gender Disparity in Tactical Combat Casualty Care. *Military Medicine*, *189*(9-10), 2114-2119. https://doi.org/10.1093/milmed/usad455

Vincent, D. S., Sherstyuk, A., Burgess, L., & Connolly, K. K. (2008). Teaching mass casualty triage skills using immersive three-dimensional virtual reality. *Academic Emergency Medicine*, *15*(11), 1160-1165. https://doi.org/10.1111/j.1553-2712.2008.00191.x

Vine, S. J., Moore, L. J., & Wilson, M. R. (2016). An integrative framework of stress, attention, and visuomotor performance. *Frontiers in Psychology*, *7*, Article 1671. https://doi.org/10.3389/fpsyg.2016.01671

Vogt, P., Boer, R., de Boer, M., Prins, H., Smit, J., Tuinstra, D., Degens, N., Hettinga, M., & Paans, W. (2023). Designing and evaluating a Virtual Reality training for paramedics to practice triage in complex situations. International Conference on Human-Computer Interaction. https://doi.org/10.1007/978-3-031-35634-6_36

Voigt, L., & Frenkel, M. O. (2023). How Officers Perform and Grow under Stress: Police Training in Virtual Reality. In M. S. Staller, S. Koerner, & B. Zaiser (Eds.), *Police Conflict Management, Volume II: Training and Education* (pp. 187-211). Springer. https://doi.org/10.1007/978-3-031-41100-7_9

Wakasugi, M., Nilsson, H., Hornwall, J., Vikström, T., & Rüter, A. (2009). Can performance indicators be used for pedagogic purposes in disaster medicine training? *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *17*, Article 15. https://doi.org/10.1186/1757-7241-17-15

Wespi, R., Birrenbach, T., Schauber, S. K., Manser, T., Sauter, T. C., & Kämmer, J. E. (2023). Exploring objective measures for assessing team performance in healthcare:

an interview study. *Frontiers in Psychology*, *14*, Article 1232628. https://doi.org/10.3389/fpsyg.2023.1232628

Wespi, R., Neher, A. N., Birrenbach, T., Schauber, S. K., Frenkel, M. O., Schrom-Feiertag, H., Sauter, T. C., & Kämmer, J. E. (under review). Physiological Team Dynamics Explored: Physiological Synchrony in Medical Simulation Training. *Advances of Simulation*.

Whetzel, E., Walker-Cillo, G., Chan, G. K., & Trivett, J. (2013). Emergency nurse perceptions of individual and facility emergency preparedness. *Journal of Emergency Nursing*, *39*(1), 46-52. https://doi.org/10.1016/j.jen.2011.08.005

Wiese, L. K., Love, T., & Goodman, R. (2021). Responding to a simulated disaster in the virtual or live classroom: Is there a difference in BSN student learning? *Nurse Education in Practice*, *55*, 103170. https://doi.org/10.1016/j.nepr.2021.103170

Williams, J., Nocera, M., & Casteel, C. (2008). The effectiveness of disaster training for health care workers: a systematic review. *Annals of Emergency Medicine*, *52*(3), 211-222. https://doi.org/10.1016/j.annemergmed.2007.09.030

World Health Organization. (2007). *Mass casualty management systems: strategies and guidelines for building health sector capacity*. https://www.who.int/publications/i/item/9789241596053

Wrzus, C., Baetzner, A. S., Hill, Y., Frenkel, M. O., & Gerwann, S. (2024). *Med1stMR VR study - Supplementary Study Material* https://doi.org/10.17605/OSF.IO/VHWTP

Wrzus, C., Schöne, B., & Frenkel, M. O. (2024). Current Opportunities and Challenges of Immersive Virtual Reality for Psychological Research and Application. *Acta Psychologica*, *249*, Article 104485. https://doi.org/10.1016/j.actpsy.2024.104485

Xia, R., Li, S. J., Chen, B. B., Jin, Q., & Zhang, Z. P. (2020). Evaluating the effectiveness of a disaster preparedness nursing education program in Chengdu, China. *Public Health Nursing*, *37*(2), 287-294. https://doi.org/10.1111/phn.12685

Xu, C., Oberman, T., Aletta, F., Tong, H., & Kang, J. (2020). Ecological validity of immersive virtual reality (IVR) techniques for the perception of urban sound environments. In *Acoustics* (Vol. 3, pp. 11-24). MDPI. https://doi.org/10.3390/acoustics3010003

Yafe, E., Walker, B. B., Amram, O., Schuurman, N., Randall, E., Friger, M., & Adini, B. (2019). Volunteer first responders for optimizing management of mass casualty incidents. *Disaster Med Public Health Prep*, *13*(2), 287-294. https://doi.org/10.1017/dmp.2018.56

Yanagawa, Y., Omori, K., Ishikawa, K., Takeuchi, I., Jitsuiki, K., Yoshizawa, T., Sato, J., Matsumoto, H., Tsuchiya, M., & Osaka, H. (2018). Difference in first aid activity during mass casualty training based on having taken an educational course. *Disaster Med Public Health Prep*, *12*(4), 437-440. https://doi.org/10.1017/dmp.2017.99

Zechner, O., García Guirao, D., Schrom-Feiertag, H., Regal, G., Uhl, J. C., Gyllencreutz, L., Sjöberg, D., & Tscheligi, M. (2023). NextGen Training for Medical First Responders: Advancing Mass-Casualty Incident Preparedness through Mixed Reality Technology. *Multimodal Technologies and Interaction*, *7*(12), Article 113. https://doi.org/10.3390/mti7120113

Zechner, O., Kleygrewe, L., Jaspaert, E., Schrom-Feiertag, H., Hutter, R. V., & Tscheligi, M. (2023). Enhancing operational police training in high stress situations with virtual reality: experiences, tools and guidelines. *Multimodal Technologies and Interaction*, *7*(2), Article 14. https://doi.org/10.3390/mti7020014

Zechner, O., Schrom-Feiertag, H., Wespi, R., Pretolesi, D., Nguyen, Q., & Tscheligi, M. (2024). Enhancing Mixed Reality Simulation Training Technology with Real-Time Performance Visualization: Mixed Methods Study with Medical First

Responders. *JMIR XR and Spatial Computing (JMXR)*. https://doi.org/10.2196/57655

Zechner, O., Uhl, J. C., Baetzner, A. S., Birrenbach, T., Egger-Lampl, S., Schrom-Feiertag, H., & Tscheligi, M. (2024). Mixed Reality Training for Medical First Responders: System Evaluation and Technical Guidelines. *Research Square*. https://doi.org/10.21203/rs.3.rs-4383276/v1

Zhang, B., Yan, Z., Wang, J., Luo, Y., Yang, S., & Fei, Z. (2018, 23-27 July 2018). An Audio-Visual Quality Assessment Methodology in Virtual Reality Environment. 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). https://doi.org/10.1109/ICMEW.2018.8551522

Zhang, D., Liao, H., Jia, Y., Yang, W., He, P., Wang, D., Chen, Y., Yang, W., & Zhang, Y. P. (2021). Effect of virtual reality simulation training on the response capability of public health emergency reserve nurses in China: a quasiexperimental study. *BMJ Open*, *11*(9), Article e048611. https://doi.org/10.1136/bmjopen-2021-048611

Zheng, Z., Yuan, S., Huang, M., Liao, J., Cai, R., Zhan, H., Yang, Z., & Xiong, Y. (2020). Flipped classroom approach used in the training of mass casualty triage for medical undergraduate students. *Disaster Med Public Health Prep*, *16*(1), 94-101. https://doi.org/10.1017/dmp.2020.162

Ziegler, M., & Bühner, M. (2012). *Grundlagen der psychologischen Diagnostik*. Springer-Verlag.

Zimmer, P., Buttlar, B., Halbeisen, G., Walther, E., & Domes, G. (2019). Virtually stressed? A refined virtual reality adaptation of the Trier Social Stress Test (TSST) induces robust endocrine responses. *Psychoneuroendocrinology*, *101*, 186-192. https://doi.org/10.1016/j.psyneuen.2018.11.010

Appendix

Additional File 2.1

Search String

Pubmed

(("Health Personnel"[MeSH Terms] OR "Emergency Medical Technicians"[MeSH Terms] OR "Emergency Medical Services"[MeSH Terms] OR "Emergency Medicine"[MeSH Terms] OR "first responder*"[Title/Abstract] OR "emergency medical technician*"[Title/Abstract] OR "ambulance"[Title/Abstract] OR "paramedic*"[Title/Abstract] OR "prehospital"[Title/Abstract] OR "nurse*"[Title/Abstract] OR "nursing student*"[Title/Abstract] OR "medical student*"[Title/Abstract] OR "physician*"[Title/Abstract] OR "health care"[Title/Abstract] OR "healthcare"[Title/Abstract] OR "emergency service*"[Title/Abstract])
AND ("Teaching"[MeSH Terms] OR "Education"[MeSH Terms] OR "Simulation Training"[MeSH Terms] OR "training*"[Title/Abstract] OR "practice*"[Title/Abstract] OR "exercise*"[Title/Abstract] OR "education*"[Title/Abstract] OR "teaching*"[Title/Abstract] OR "simulation*"[Title/Abstract])
AND ("Disasters"[MeSH Terms] OR "disaster medicine/education"[MeSH Terms] OR "Mass Casualty Incidents"[MeSH Terms] OR "mass casualt*"[Title/Abstract] OR "disaster*"[Title/Abstract] OR "major incident*"[Title/Abstract] OR "major accident*"[Title/Abstract] OR "catastrophe*"[Title/Abstract])
AND ("Comparative Study" [Publication Type] OR "Evaluation Study"[Publication Type] OR "Outcome Assessment, Health Care"[MeSH Terms] OR "intervention*"[Title/Abstract] OR "effective*"[Title/Abstract] OR "compar*"[Title/Abstract] OR "evaluat*"[Title/Abstract] OR "measure*"[Title/Abstract] OR "assess*"[Title/Abstract])
AND (2010/1:2021/9[pdat]) AND (english[Filter]))

Web of Science

TS=((health personnel OR emergency medical service* OR emergency service* OR emergency medicine OR first responder* OR emergency medical technician* OR ambulance OR paramedic* OR prehospital OR nurse* OR nursing student* OR medical student* OR physician* OR health care OR healthcare)
AND (training* OR practice* OR exercise* OR education* OR teaching* OR simulation*)
AND (mass casualt* OR disaster* OR major incident* OR major accident* OR catastrophe*)
AND (compar* OR evaluat* OR intervention* OR effective* OR measure* OR assess*))
Language: English; Time Span: 2010-01-01 – 2021-09-30

Supplementary Table 2.1

Table with additional information about studies.

First author, year	Pre- or pre- and in-hospital	Location	Mean age (SD)	Age range	Gender (% female)	Training Scenarios	Results (long)
Aghababaeian et al. (2013)	Prehospital	Iran	31.32 (3.9)	*NA*	*NA*	Reconstructed accident	There was no significant difference in knowledge increase between groups directly or 15 days after the training. Groups did not significantly differ in their mean performance increase directly after their trainings but the IG performed slightly better than the CG after 15 days (p =.02)
Alenyo et al. (2018)	Prehospital	South Africa	33.14 (7.77)	*NA*	49.6	NA	After the training, the overall correct triage score (pre: 53.9%, post: 63.6%) and overtriage rate (pre: 31.4%, post: 17.9%) improved significantly, while undertriage rates increased (pre: 13.8%, post: 17.8%; for all three no overlap in confidence intervals: p <.05).
Alim et al. (2015)	Both	Indonesia	20.66 *(NA)*	18-29	87.70	Earthquake	The mean test scores for undergraduate students (pre: 9.84, post: 14.46) and diploma students (pre: 10.83, post: 14.68) improved significantly (max. range: 0-20; both p = 0.001).
Aluisio et al. (2016)	Prehospital	USA	*NA*	*NA*	83.3	NA	The mean knowledge score improved significantly (pre: 79.2%, post: 88.4%, p < 0.001).
Andreatta et al. (2010)	Both	USA	*NA*	*NA*	*NA*	Explosion in office building	After training, the CG had a greater knowledge improvement than the IG (Cohen's d = 0.63). The IG demonstrated a better triage performance (Cohen's d = 0.25; effect sizes only).
Andreatta et al. (2015)	Prehospital	USA	*NA*	*NA*	*NA*	Exposure to nerve agents	There were significant increases in scores for knowledge, performance, self-efficacy, and affect after training for both groups (all p < 0.001). There were no significant differences in any of the post-training outcomes for the two groups.
Bajow et al. (2016)	Both	Italy	23.6 (1.9)	*NA*	48.28	Building collapse, fire on boats at a seaport	The mean knowledge score improved significantly (pre: 41.0 %, post: 67.7 %, p < 0.0001).
Betka et al. (2021)	Both	USA	*NA*	*NA*	*NA*	Agriculture disaster simulation	*Separate analyses for relevant subsamples;* After training, nursing students reported significantly increased interprofessional collaborative competencies (t(9)=7.673, p<.001), disaster management competencies (t(9)=8.938, p<.001), and self-confidence (t(8)=5.893, p<.001). There was no significant improvement of medical students' interprofessional collaborative competencies or self-confidence but they reported increased disaster management competencies (t(6)=6.075, p =.001)

Chan et al. (2010)	Prehospital	China	*NA*	NA	88.2	NA	The Wilcoxon signed rank test revealed that the mean competency score was significantly improved after training (Z=-9.02, p < .001, Cohen's d = 2.79).
Chandra et al. (2014)	Both	USA	*NA*	18 - 64	71	NA	From pre- to post-training, self-reported capability increased from 71% to 90% (p < .01). There was no significant increase of knowledge.
Chou et al. (2021)	Prehospital	Taiwan	*NA*	*NA*	54.17	Earthquake	Four knowledge domains were tested with possible scores between 0 and 2. The mean knowledge score increased significantly in the safety domain (pre: 1.22, post: 1.91, p < .0001) and decreased in disaster patient care (pre: 1.83, post: 1.44, p = .0005). There were no significant differences in the mean scores for the communication or resource management domains. There was no significant improvement of willingness to pursue further training or interest in disaster training exercises.
Cicero et al. (2012)	Prehospital	USA	*NA*	*NA*	*NA*	School shooting, playground violence, school bus crash	Mean triage performance improved from 6.9 before training to 8.0 out of 10 patients accurately triaged one week after the training (p < .0001). Five months later, there was maintenance of triage improvement, with a mean triage score of 7.8 (n = 42; p < .0001).
Cicero et al. (2017)	Prehospital	USA	*NA*	*NA*	*NA*	Multi-family house fire, school shooting, school bus crash	The median score of triage accuracy improved significantly between baseline and the posttest two weeks after the training (pre: 80%, post: 90%, p < 0.001). There was no significant difference between the posttest and follow-up 6 months later.
Cowling et al. (2021)	Prehospital	South Africa	*NA*	*NA*	*NA*	Structural collapse	There was no significant improvement in knowledge test scores after training. After the training, participants reported increased self-reported knowledge (p < 0.001) and confidence (p < 0.001).
Cuttance et al. (2017)	Prehospital	Australia	*NA*	*NA*	46.92	Traffic accident of a minibus	Compared to the CG, all other groups had a significantly greater number of correctly triaged cases (p < .001) and lower under-triaging rates (p < .001). The CG had an accuracy rate of 47%. The provision of either an educational refresher lecture or aide-memoir significantly increased the accuracy rate to 77% and 90%, respectively. Participants who received both the lecture and aide-memoir had an overall accuracy rate of 89% which is significantly higher than the rate of those only receiving the lecture (p = .02). Over-triage rates were found not to differ significantly across any of the groups.

Dittmar et al. (2018)	Prehospital	Germany	*NA*	*NA*	*NA*	*NA*	*Separate analyses for subgroup;* One year after the first training and before the re-training, paramedics' overall performance score was 90%. After the re-training, it improved to 97% (p < .05). While the overall performance components accuracy, sensitivity, (critical) under-triage, airway and bleeding management significantly improved, the components specificity, (critical) over-triage, and time requirement did not.
Edinger et al. (2019)	Prehospital	USA	*NA*	*NA*	*NA*	*NA*	Mean knowledge improved significantly (overall pre: 66%, post: 81%, for 4/14 items p < .05). Self-efficacy also improved (for all 10 items p < .05).
Farra et al. (2013)	Prehospital	USA	*NA*	18 - 57	91	Radioactive and explosive events	Overall, the main effect of being in the IG on knowledge was significant (p < .0001). Although the two groups already differed before training (IG: 13.5, CG: 11.3, max. Range: 0-20, p = .023; Cohen's d = .964), the use of Generalized Estimating Equations controlled for these differences. Both groups showed a similar improvement following the training (posttest; IG: 17.68, CG: 16.24). After two months, the IG's knowledge scores demonstrated stability over time while the CG showed significant decay (follow-up; IG: 16.95, CG: 14.10).
Fernandez-Pacheco et al. (2017)	Prehospital	Spain	29 (5)	*NA*	57	*NA*	After watching the video of the drill, they participated in, 80% of the students modified their self-perception (p = .001). The number of behaviors and moments that the students were able to describe increased (behaviors: increase of 14%, p = .031; moments: increase of 40%, p = .033). Scores in the other variables (thoughts, feelings, strengths, and weaknesses) did not change significantly.
Foronda et al. (2016)	Prehospital	USA	*NA*	18 - 39	100	Earthquake	There was no statistically significant difference between the group's pre- and post-test performance (p = .168).
Furseth et al. (2016)	Both	USA	nursing students: 21 *(NA)*, para-medic students: 26 *(NA)*	19 - 45	67.2	Large outbreak of food poisoning aboard a cruise ship, an explosion/bombing, a bus crash	Nursing students in the IG had a greater change in attitudes (both p = .001), self-confidence (p < .001) and satisfaction (p < .001) than in the CG. Regarding paramedic students' attitudes, self-confidence and satisfaction, there was no significant difference between the two trainings.
Greco et al. (2019)	Prehospital	USA	*NA*	*NA*	*NA*	Toxic chemical spill caused by a train derailment	After training, perceived importance of ethical reasoning (t(89)=-2.832, p = .006) and confidence (t(89)=−6.609, p < 0.001) increased significantly.

Huh and Kang (2019)	Both	South Korea	23.12 (5.19)	*NA*	83.3	Earthquake, explosion at a marathon tournament (both part of in-class learning)	After training, the IG had a greater improvement in disaster nursing knowledge ($t(58) = 14.37$, $p < 0.001$), triage knowledge ($t(58) = 7.90$, $p = 0.002$) and disaster readiness ($t(58) = 10.82$, $p < 0.001$) than the CG.
Hutchinson et al. (2011)	Both	USA	*NA*	*NA*	*NA*	Explosion in a chemistry lab (after posttest)	Mean knowledge scores for two out of three groups of nursing students were significantly higher after training than before. The sophomore students' mean knowledge increased from 60.97 to 95.60 ($n = 26$; $p < .05$) and the senior nursing students' from 49.97 to 85.70 ($n = 24$; $p < .05$). Junior-level students' knowledge did not increase significantly (from 54.81 to 93.52; $n = 31$).
Ingrassia et al. (2015)	Prehospital	Italy	*NA*	*NA*	*NA*	Car accident	On day 1, the group A live scenario triage accuracy was 58% and the average time to assess all patients was 4'28 min per participant. For group B, the overall virtual scenario triage accuracy was 52% and the average time to complete the assessment was 5'18. There was no statistical difference between the two groups. On day 3, the overall triage accuracy for group A in the virtual simulation was 92% and the average time was 3'53. In live exercise, group B performed the triage with an overall accuracy of 84% in 3'25. Again, there was no statistical difference between the two groups. However, there was an equivalent significant improvement between the pre- and the postintervention triage scores (day 1 vs. Day 3, $p < 0.001$). The time to complete each scenario decreased from day 1 to day 3 in both groups ($p < 0.05$). There was a significant improvement between the day 1 and day 3 life-saving treatment scores in both groups (day 1 vs. Day 3, $p < 0.01$).
Ingrassia, Ragazzoni, et al. (2014)	Both	Italy	*NA*	*NA*	61.5	Car accident	After training, the mean knowledge score increased from 3.95 to 8.29 out of 10 ($p < .01$). Triage accuracy improved from 45% to 78% ($p < .01$).
James et al. (2021)	Prehospital	USA	25.06 (7.74)	*NA*	94.10	House contamination/ bioterrorism, explosion at a fertilizer plant, camping chaos post-tornado	After training, participants reported more positive attitudes toward teamwork in training ($t(33)=-4.25$, $p < .01$).
Jones et al. (2014)	Prehospital	USA	*NA*	*NA*	22	Active shooter incident	After training, more participants felt prepared to respond to an active shooter incident (pre: 41%, post: 89%) and comfortable working jointly on rescue operations with law enforcement personnel in response to an active shooter incident (pre: 61%, post: 93%; only descriptive statistics reported).
Kim and Lee (2020)	Both	South Korea	22.82 (1.38)	*NA*	85.3	Gas explosion	After training, participants reported improved response attitudes ($t(33)=16.31$, $p < .001$).

Knight et al. (2010)	Prehospital	United Kingdom	*NA*	*NA*	28.57	Bomb explosion in a busy urban street; domestic outdoor gas explosion accident (scenario in evaluation)	After training, triage accuracy was significantly higher in the IG than in the CG ($Chi^2 = 13.126$, $p = 0.02$). 72% of the IG and 55% of the CG correctly triaged all 8 patients. Groups did not differ in the number of patients that received correctly followed procedure ($p >$ .05). However, more trainees of the IG followed correct procedure for all eight patients ($Chi^2 = 5.45$, $p = 0.0196$). There was no significant difference in time to triage all casualties ($p = 0.155$).
Koca and Arkan (2020)	Both	Turkey	IG: 21.16 (1.23) CG: 20.81 (2.47)	*NA*	77.45	Earthquake, fires and combination of both	After training, the IG had significantly higher preparedness and self-efficacy scores than the CG. Training explained 33.1% ($R^2 = .331$) of the increase in disaster preparedness and 31.7% ($R^2 = 0.317$) of the increase in disaster response self-efficacy.
Koutitas et al. (2021)	Prehospital	USA	*NA*	*NA*	NA	NA	Compared to the CG, the VR solution helped trainees to increase their skills by a factor of 46% in terms of number of errors, 29% in terms of speed and 36% as an overall performance (a metric that normalizes the error and speed metric in one formula). The AR solution improved their skills by a factor of 34.5% in terms of number of errors, 10% in terms of speed and 17% as an overall performance (only descriptive statistics reported).
Kuhls et al. (2017)	Prehospital	Thailand	Nurses: 40 (*NA*) physicians: 38 (*NA*)	nurses: 24-63 physicians: 25–61	54.94	NA	Separate analyses for relevant subsamples; After the training, all occupational groups reported a significant confidence increase in each confidence category surveyed (all $p <$ .001). Physicians' and nurses' median changes in the different confidence areas were between 1 and 2 points (scale: 1-5).
Lampi et al. (2013)	Prehospital	Sweden	*NA*	*NA*	*NA*	Bus crash (scenario in evaluation)	There was no significant difference between pre and post-course test results ($p >$.05).
Lennquist Montán et al. (2015)	Both	NA	*NA*	*NA*	*NA*	NA	Separate analyses for subsample; A significant increase between pre- and post-course assessment was registered for all items of the self-reported knowledge and skills questionnaire ($p < 0.001$). The average increase for prehospital staff was 74%.
Ma et al. (2021)	Prehospital	China	IG: 19.22 (0.76), CG: 19.17 (0.80)	*NA*	83.7	Earthquake disaster as the background of the IG game	After training, the IG rated their competency significantly higher than the CG ($t(102)=3.114$, $p =.002$).

Merlin et al. (2010)	Both	USA	*NA*	*NA*	*NA*	*NA*	The mean value of the 4 self-reported knowledge items increased from 3.11 before training to 4.35 after training (scale from 1 to 5; 3 out of 4 items $p > .0001$). Greater than 35% opinion change was found in several areas, including education differences in prehospital providers, general teaching of prehospital care (both $p < 0.0001$).
Mills et al. (2020)	Prehospital	Australia	*NA*	*NA*	*NA*	Police car chase and shooting	Triage accuracy did not differ in the two tested training methods. Participants needed more time to triage live patients compared to virtual patients (p<0.001). Average heart rate, heart rate increase, and maximum heart rate were significantly higher during the live simulation compared to the VR simulation (all p<0.001). Participants reported a significantly higher immersion level during the live simulations (p<0.001) which seems to be caused by the subscale physical demand (p<001, all other subscales p>.05). There were no significant differences in learning satisfaction.
Motola et al. (2015)	Prehospital	USA	IG: 37.3 (*NA*) CG: 36.9 (*NA*)	*NA*	IG: 4 CG: 7	Nerve agent, explosives, radiologic event	After training, the IG had a significantly greater improvement in knowledge than the CG (IG: 53.3% to 63.4%, CG: 55.5% to 59.3%, $p = 0.001$). The IG performed better in the explosives and chemical nerve agent scenarios than the CG (both $p < 0.01$) but there was no significant difference between the groups in the radiologic scenario ($p = 0.51$).
Paddock et al. (2015)	Prehospital	USA	IG_1: 43 (13) CG: 50 (11) IG_2: 43 (9)	*NA*	33.33	*NA*	Within each training group, there was a statistically significant improvement in the mean pre- and post-course knowledge test scores (all $p < 0.001$, pretest: 14.3%-20.8%, posttest: 53%-54.3%). There were no significant differences between the three training groups' post-course knowledge gains ($p > 0.05$). After the training the groups did not differ in their mean image acquisition scores or their image interpretation scores.
Phattharapornjaroen et al. (2020)	Both	Thailand	*NA*	26–35	69.32	Terror attack along with a bomb explosion, riot, and shooting; Building fire	After training, self-reported knowledge increased in all domains (all $p < .01$) with the greatest improvement in safety issues (pre: 40%, post: 96%) and the lowest in treatment (pre: 54%, post: 71%).
Pollard et al. (2015)	Both	USA	*NA*	*NA*	*NA*	Plane crash on a soccer field near the airport, carrying terrorists; operating room explosion; patient trapped in debris field resulted from a tornado	The training led to a significant increase in knowledge. This result was found in the subsample that provided complete data ($t(7)=-2.35$, $p = 0.05$) as well as in the imputed dataset ($t(40)=-11.72$, $p < 0.001$).

Pouraghaei et al. (2017)	Prehospital	Azerbaijan	34.97 (6.42)	23 - 50	*NA*	*NA*	After training, the mean triage knowledge score (pre: 11.47, post: 13.63, max. Range: 0-15, $p < .05$) and the knowledge score in the performance section increased significantly (pre: 10.73, post: 14.93, max. Range: 0-19, $p < .05$). The number of participants able to perform the jaw trust airway maneuver increased as well (pre: 21.9%, post: 88.3%, $p < 0.001$).
Ripoll-Gallardo et al. (2020)	Both	Italy	*NA*	*NA*	37.5	*NA*	After training, there was a significant improvement in knowledge scores (10.4 points, max. Range: 0-30; $p = .001$) as well as in performance scores (3 points, scale: 1-7; $p = .000001$). Attitudes did not change significantly.
Rivkind et al. (2015)	Both	Israel	*NA*	*NA*	NA	*NA*	After training, the mean knowledge score increased significantly (pre: 54%, post: 68%, $p < 0.001$).
Saiboon et al. (2021)	Both	Malaysia	*NA*	*NA*	80.4	*NA*	After training, the mean knowledge score increased significantly (pre: 6.99, post: 13.31, max. Range 0-20, $p < 0.001$).
Scott et al. (2010)	Prehospital	USA	*NA*	*NA*	*NA*	Hazardous materials/chemical spill scenario (overturned truck on a rural two-lane highway with a possible chemical exposure and multiple victims)	In 2008, the average of knowledge scores improved from 39% to 58% ($n = 30$) and in 2009 from 47% to 57% ($n = 31$). Self-reported knowledge increased as well (2008: 3.76 to 7.64 out of 10; 2009: 2.52 to 3.76 out of 5; only descriptive statistics reported).
Sena et al. (2021)	Both	USA	*NA*	*NA*	*NA*	Explosion at a major sporting event with blast injuries	After training, there was no significant increase in knowledge or perceived importance of disaster medicine training but a significant increase in confidence (pre: 2, post: 4, scale: 1-5; $p = .011$).
Smith et al. (2015)	Both	USA	*NA*	*NA*	*NA*	Radiological bomb explosion in front of a local courthouse	After training, participants reported increased self-efficacy ($t(64)=8.45$; $p < .001$).
Unver et al. (2018)	Both	Turkey	21.95 (0.26)	*NA*	100	Earthquake	The Wilcoxon signed rank test revealed that there was a significant increase in preparedness after training ($Z = -7.572$, $p = .001$)
Wiese et al. (2021)	Prehospital	USA	26.5 (3.49)	16 - 37	70.1	Bus encountering a tornado	After training (but before the cross-over), the IG had a higher mean knowledge score than the CG (IG: 20.55, CG: 15.93, possible range: 0-25; only descriptive statistics reported). Furthermore, participants reported significant learning gains, regardless of whether the training was live or virtual (9 out of 12 items: $p < .05$).

Xia et al. (2020)	both	China	IG: 21.46 (1.34) CG:34.38 (1.35)	*NA*	84.13 IG: 83.87 CG: 84.37	Earthquake	After training, the IG displayed greater theoretical knowledge scores in the three tested domains (all $p < .01$) and greater skill-related knowledge in two out of three domains (both $p < .05$) than the CG. One month after the training, the IG had higher scores in theoretical and skill-related knowledge in one of the three domains. The groups did not differ in their attitude at any point.
Yanagawa et al. (2018)	both	Japan	*NA*	*NA*	NA	Collision between a minibus and a common automobile	Teams that included a chief EMT who attended the training as well as teams that included staff members who attended the training did not perform significantly better than teams without (both total performance scores $p > .05$).
Zhang et al. (2021)	both	China	*NA*	*NA*	90	Covid-19 patients	After training, the IG demonstrated greater improvement in knowledge ($t(58)=4.783$, $p < .001$), performance ($t(58)=4.416$, $p < .001$), technical skills ($t(58)=2.708$, $p = .008$) and disaster preparedness ($t(58)=5.295$, $p < .001$) than the CG.
Zheng et al. (2020)	Prehospital	China	IG: 24.55 (1.14) CG: 24.52 (1.0)	*NA*	46.60	Traffic accident	After training, the IG scored significantly better in the knowledge test (difference of 3 points, max. Range:0-50; $p < 0.001$). There were no significant differences in performance between groups. The IG reported a higher satisfaction with the course (5 out of 8 items, $p < .05$) but also perceived it as more work and burden (both, $p < .05$)

Note. IG = intervention group, CG = control group, NA = not available

Risk of Bias

JBI CRITICAL APPRAISAL CHECKLIST FOR RANDOMIZED CONTROLLED TRIALS

1. Was true randomization used for assignment of participants to treatment groups?
2. Was allocation to treatment groups concealed?
3. Were treatment groups similar at the baseline?
4. Were participants blind to treatment assignment?
5. Were those delivering treatment blind to treatment assignment?
6. Were outcomes assessors blind to treatment assignment?
7. Were treatment groups treated identically other than the intervention of interest?
8. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analyzed?
9. Were participants analyzed in the groups to which they were randomized?
10. Were outcomes measured in the same way for treatment groups?
11. Were outcomes measured in a reliable way?
12. Was appropriate statistical analysis used?
13. Was the trial design appropriate, and any deviations from the standard RCT design (individual randomization, parallel groups) accounted for in the conduct and analysis of the trial?

Supplementary Table 2.2

Risk of bias of the experimental studies.

First author, year	1	2	3	4*	5*	6*	7	8	9	10	11	12	13
Andreatta, 2010	no	yes	yes	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes	yes
Cuttance, 2017	yes	yes	yes	not appl.	not appl.	not appl.	yes	not appl. [1]	yes	yes	yes	yes	yes
Farra, 2013	yes	yes	no	not appl.	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes
Huh, 2019	yes	yes	yes	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes	yes
Ingrassia, 2015	yes	yes	un-clear	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes	yes
Koca, 2020	yes	yes	yes	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes	yes
Koutitas, 2021	yes	yes	un-clear	not appl.	not appl.	not appl.	yes	not appl. [1]	yes	yes	yes	yes	yes
Ma, 2021	yes	yes	yes	not appl.	not appl.	not appl.	yes	yes	yes	yes	no	yes	yes
Mills, 2020	yes	no	yes	not appl.	not appl.	not appl.	yes	not appl. [1]	yes	yes	yes	yes	yes
Motola, 2015	yes	yes	yes	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes	yes
Paddock, 2015	no	yes	yes	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes	yes
Xia, 2020	yes	yes	yes	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes	yes
Zhang, 2021	yes	yes	yes	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes	yes
Zheng, 2020	yes	yes	yes	not appl.	not appl.	not appl.	yes	yes	yes	yes	yes	yes	yes

Notes. *yes* indicates a lower risk of bias, *no* indicates a higher risk of bias; *not appl.* = not applicable; questions 4-6 refer to blinding. Because blinding is hardly or not at all feasible in training evaluation studies, the items were answered with not applicable; [1]RCT studies with only a post-test.

JBI CRITICAL APPRAISAL CHECKLIST FOR QUASI-EXPERIMENTAL STUDIES

1. Is it clear in the study what is the 'cause' and what is the 'effect' (i.e. there is no confusion about which variable comes first)?
2. Were the participants included in any comparisons similar?
3. Were the participants included in any comparisons receiving similar treatment/care, other than the exposure or intervention of interest?
4. Was there a control group?
5. Were there multiple measurements of the outcome both pre and post the intervention/exposure?
6. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analyzed?
7. Were the outcomes of participants included in any comparisons measured in the same way?
8. Were outcomes measured in a reliable way?
9. Was appropriate statistical analysis used?

Supplementary Table 2.3

Risk of bias of the quasi-experimental studies.

First author, year	1	2	3	4	5	6	7	8	9
Aghababaeian, 2013	yes	yes	yes	yes	yes	yes	yes	yes	yes
Alenyo, 2018	yes	yes	yes	no	yes	yes	yes	yes	yes
Alim, 2015	yes	yes	yes	no	yes	yes	yes	yes	yes
Aluisio, 2016	yes	yes	yes	no	yes	yes	yes	yes	yes
Andreatta, 2015	yes	yes	yes	yes	yes	yes	yes	yes	yes
Bajow, 2016	yes	yes	yes	no	yes	yes	yes	yes	yes
Betka, 2021	yes	yes	yes	no	yes	yes	yes	yes	yes
Chan, 2010	yes	yes	yes	no	yes	yes	yes	yes	yes
Chandra, 2014	yes	yes	yes	no	yes	yes	yes	yes	yes
Chou, 2021	yes	yes	yes	no	yes	no	yes	yes	yes
Cicero, 2012	yes	yes	yes	no	yes	no	yes	yes	yes
Cicero, 2017	yes	yes	yes	no	yes	no	yes	yes	yes
Cowling, 2021	yes	yes	yes	no	yes	yes	yes	yes	yes
Dittmar, 2018	yes	yes	yes	no	yes	no	yes	yes	yes
Edinger, 2019	yes	yes	yes	no	yes	no	yes	yes	yes
Fernandez-Pacheco, 2017	yes	yes	yes	no	yes	yes	yes	yes	yes
Foronda, 2016	yes	yes	yes	no	yes	yes	yes	yes	yes
Furseth, 2016	yes	yes	yes	yes	yes	yes	yes	yes	yes
Greco, 2019	yes	yes	yes	no	yes	yes	yes	yes	yes
Hutchinson, 2011	yes	yes	no	no	yes	yes	yes	yes	yes
Ingrassia, 2014	yes	yes	yes	no	yes	yes	yes	yes	yes
James, 2021	yes	yes	yes	no	yes	yes	yes	yes	yes
Jones, 2014	yes	yes	yes	no	yes	yes	yes	yes	no
Kim, 2020	yes	yes	yes	no	yes	yes	yes	yes	yes
Knight, 2010	yes	yes	yes	yes	no	not appl.[1]	yes	yes	yes
Kuhls, 2017	yes	yes	yes	no	yes	yes	yes	yes	yes

Lampi, 2013	yes	yes	yes	no	yes	yes	yes	yes	yes
Merlin, 2010	yes	yes	yes	no	yes	yes	yes	no[2]	yes
Lennquist Montán, 2015	yes	yes	yes	no	yes	yes	yes	no[2]	yes
Phattharapornjaroen, 2020	yes	yes	yes	no	yes	yes	yes	no[2]	yes
Pollard, 2015	yes	yes	yes	no	yes	yes	yes	yes	yes
Pouraghaei, 2017	yes	yes	yes	no	yes	no	yes	yes	yes
Ripoll-Gallardo, 2020	yes	yes	yes	no	yes	yes	yes	yes	yes
Rivkind, 2015	yes	yes	yes	no	yes	no	yes	yes	yes
Saiboon, 2021	yes	yes	yes	no	yes	no	yes	yes	yes
Scott, 2010	yes	yes	yes	no	yes	yes	yes	yes	yes
Sena, 2021	yes	no	yes	no	yes	no	yes	yes	no
Smith, 2015	yes	yes	yes	no	yes	yes	yes	yes	yes
Unver, 2018	yes	yes	yes	no	yes	yes	yes	yes	yes
Wiese, 2021	yes	yes	yes	yes	yes	yes	yes	yes	yes
Yanagawa, 2018	yes	unclear	yes	yes	no	not appl.[1]	yes	yes	yes

Note. [1]only a post-test; [2]called the outcome “knowledge”/”skills” but only measured self-rated knowledge/self-rated skills.

Additional File 4.1

Scenario Descriptions

Triage color categories: Green = minor injuries; yellow = patient's transport can be delayed; red = immediate intervention and transport needed; black = deceased (see the START triage algorithm for further information)

Scenario 1 (Very Low Difficulty)

- Accident on countryside road (clear vision, middle of the day)
- Two cars with four people involved in frontal crash (one car left its lane and crossed into incoming traffic)
- The cars have minor damage (clearly visible brake marks indicate the cars collided at low speed)
- No bystanders or other vehicles around
- Traffic was stopped by the police
- Spilled oil
- Triage
    - Four injured in total
    - Three individuals are green, one yellow

Scenario 2 (Low Difficulty)

- Accident on a motorway (clear vision, cloudy day)
- A car and a van were involved in a frontal crash (one car left its lane and crossed into oncoming traffic). Additionally, a motor cyclist crashed into one of the cars and fell far from the site
- The vehicles show larger degrees of damage; light smoke emanates from the crash zone
- The accident site has already been secured; traffic passes by slowly but quite close to the accident scene
- Spilled oil around the motorcycle
- One uninvolved car with three bystanders (a father with his son and a toddler daughter left in her car)
    - The father and the boy attend to two injured people while the toddler is continuously screaming (distractor)
- Triage
    - Five injured in total (two from each car + a motor cyclist)
    - One person from the cars is red (removed from the vehicle, attended to by the bystanders)
    - The other person from the car is yellow (also sitting outside the car)
    - In the van, one person is yellow and the other passenger green
    - The motor cyclist (far away from the site) is yellow

Scenario 3 (Medium Difficulty)

- One car has been rear-ended (with an angle) in a busy inner-city area. The hit car skipped onto the sidewalk and collided with pedestrians, one of which is stuck under the car
- The street is narrow and many bystanders stand around. High levels of noise come from the crowd. One bystander accuses the MFR of working too slowly; another bystander films the scene with her smartphone (distractors)
- The police stopped the traffic
- Triage
    - Six injured in total
    - The person under the car is red
    - The person who caused the accident was thrown from their SUV (no seatbelt worn), also red. The driver's daughter is screaming for help (green). Another passenger in the vehicle is yellow.
    - The driver of the other car is yellow
    - One pedestrian is yellow

Scenario 4 (Greater Difficulty)

- Express way accident involving four vehicles in total
- Twilight
    - One van with only a driver
    - Three cars: one with three passengers (one adult male, one adult female who is pregnant, one child), one with two passengers (one adult, one child), another one with two passengers (two adults)
- (Almost) all individuals were removed from the cars
- Heavy damage to the van
- Many bystanders around; a high level of noise comes from the crowd and traffic; a bystander toddler was left in one of the bystander cars and cries loudly (distractors).
- A large dog is walking around unleashed and barking
- Oil spill, broken glass on the floor, van is smoking at the front (light smoke)
- Triage
    - Eight injured in total
    - Van driver black
    - Family: driver (male) yellow, pregnant woman yellow, child green
    - One of the other cars: adult driver red and child yellow
    - Other car: both adults red

Scenario 5 (Highest Difficulty)

- Express way accident involving a bus and truck
- Night-time and fog
- Oil spilled and dark smoke emanates, with broken glass on the pavement
- The accident site is already secured; traffic passes by relatively close to the accident scene; the fire brigade has already set up floodlights
- Some people are screaming (distractor)
- Triage (18 people involved)
    - Truck driver black
    - Patients already outside of the bus: two black, two red, four green
    - Patients inside the bus: one black, six yellow, two red

Performance per Scenario

Supplementary Table 4.1

Means and standard deviations of the performance indicators per scenario.

Indicator	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
DOAF patients	372.85 (138.62)	457.26 (237.30)	451.71 (193.69)	382.55 (214.41)	336.89 (113.96)
DOAF Safety	316.46 (200.61)	338.82 (138.54)	Not appl.	Not appl.	306.31 (95.20)
DOAF VI zone	271.77 (97.65)	302.97 (118.61)	280.13 (92.32)	374.25 (267.63)	341.27 (164.07)
DOAF distractor	Not appl.	Not appl.	233.33 (124.27)	Not appl.	Not appl.
FC patients	13.82 (7.81)	5.81 (4.43)	12.60 (5.96)	8.25 (5.58)	23.72 (11.98)
FC safety	1.66 (0.81)	12.71 (8.77)	Not appl.	Not appl.	10.83 (5.42)
FC VI zone	12.39 (9.86)	6.11 (4.21)	11.43 (5.67)	4.40 (3.13)	5.83 (3.47)
FC distractor	Not appl.	Not appl.	3.35 (2.34)	Not appl.	Not appl.
Triage accuracy st.	0.78 (0.28)	0.80 (0.26)	0.83 (0.19)	0.85 (0.19)	0.78 (0.18)
Triage accuracy non-st.	3.11 (1.11)	3.99 (1.30)	4.97 (1.15)	6.84 (1.52)	14.00 (3.25)
Triage speed st.	1.00 (0.37)	1.00 (0.26)	1.00 (0.27)	1.00 (0.32)	1.00 (0.33)
Triage speed non-st.	2.56 (0.94)	3.99 (1.03)	3.89 (1.04)	5.62 (1.78)	9.02 (2.96)
Info. transm. efficiency	0.11 (0.06)	0.07 (0.05)	0.07 (0.04)	0.07 (0.03)	0.05 (0.03)
Info. transm. word count	27.40 (21.80)	39.61 (30.70)	38.35 (29.52)	40.76 (29.87)	47.37 (30.10)
Info. transm. scores	2.16 (0.85)	2.05 (0.89)	2.10 (0.88)	2.18 (0.97)	2.02 (0.89)
Subj. performance	7.40 (1.40)	7.07 (1.52)	7.07 (1.55)	6.46 (1.94)	5.87 (1.67)

Note. Not appl. = Not applicable; DOAF = duration of average fixation, FC = fixation count, VI zone = vehicle impact zone, st. = standardized, non-st. = non-standardized, Info. trans. = information transmission, Subj. performance = subjective performance;
Variables are not winsorized.

FAKULTÄT FÜR VERHALTENS- UND EMPIRISCHE KULTURWISSENSCHAFTEN

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Promotionsausschuss der Fakultät für Verhaltens- und Empirische Kulturwissenschaften der Ruprecht-Karls-Universität Heidelberg / Doctoral Committee of the Faculty of Behavioural and Cultural Studies of Heidelberg University

Erklärung gemäß § 8 (1) c) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften / Declaration in accordance to § 8 (1) c) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies

Ich erkläre, dass ich die vorgelegte Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe. / I declare that I have made the submitted dissertation independently, using only the specified tools and have correctly marked all quotations.

Erklärung gemäß § 8 (1) d) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften / Declaration in accordance to § 8 (1) d) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies

Ich erkläre, dass ich die vorgelegte Dissertation in dieser oder einer anderen Form nicht anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe. / I declare that I did not use the submitted dissertation in this or any other form as an examination paper until now and that I did not submit it in another faculty.

Vorname Nachname / First name Family name	Anke Sabine Baetzner
Datum / Date	06.02.2025
Unterschrift / Signature	Dem Dekanat der Fakultät für Verhaltens- und Empirische Kulturwissenschaften liegt eine unterschriebene Version dieser Erklärung vom 06.02.2025 vor.

Acknowledgements

Die Promotionszeit war für mich ein sehr spannender, abwechslungsreicher und lehrreicher Lebensabschnitt. Ich möchte diese Gelegenheit nutzen, mich bei all jenen zu bedanken, die mich auf diesem Weg begleitet haben.

Mein besonderer Dank gilt zunächst meinen beiden Betreuer*innen. Prof. Dr. Henning Plessner danke ich für seine wegweisende Beratung, seine unermüdliche Unterstützung und seinen Rückhalt. Die von ihm geschaffene herzliche Arbeitsatmosphäre und sein stets offenes Ohr, sei es im Institut oder beim Joggen, waren von unschätzbarem Wert. Prof. Dr. Cornelia Wrzus danke ich für ihre exzellente fachliche Betreuung im Rahmen des EU-Projektes und für die gelungene Balance zwischen gezielter Förderung und dem Vertrauen in meine selbständige Arbeit.

Besonderer Dank gilt auch dem EU-Horizon-2020-Projekt *MED1stMR* (Grant-Agreement-Nummer: 101021775). Die im Rahmen des Projekts bereitgestellten Ressourcen haben diese Forschung überhaupt erst ermöglicht. Ich bedanke mich außerdem bei dem MED1stMR-Konsortium. Es war sehr bereichernd, Teil eines so großen, interdisziplinären Teams zu sein.

Ein ganz persönlicher Dank geht an meine Kolleg*innen am ISSW. Die formellen, informellen und sportlichen Veranstaltungen haben diese Zeit zu etwas ganz Besonderem gemacht. Ich hätte nie gedacht, dass ich neben meiner wissenschaftlichen Expertise auch meine Jogging-, Volleyball-, Tennis- und Tanzfähigkeiten erweitern könnte. Insbesondere danke ich meiner Doktorschwester Solène Gerwann, die mich durch alle Berg- und Talfahrten begleitet hat. Dr. Yannick Hill danke ich für sein wertvolles Mentoring und sein Vertrauen in meine Fähigkeiten. Beiden danke ich außerdem für das bereichernde Feedback während der Finalisierung dieser Dissertation. Matthias Beutel gilt mein Dank für die wunderbare Zusammenarbeit und die ebenso erfolgreichen wie unterhaltsamen Dienstreisen. Becky Smith danke ich für ihr sorgfältiges Korrekturlesen meines Mantelteils. Unseren vielen studentischen Hilfskräften der letzten Jahre danke ich für die gute Zusammenarbeit und die Unterstützung bei Datenerhebungen. Ich danke außerdem Karin Baum, Frederik Borkenhagen, Carolin Rohde und Andreas Klewe, die für die MED1stMR-Projektverwaltung und IT unentbehrlich waren und uns stets freundlich und geduldig unterstützt haben. Meinen Kolleg*innen in der Abteilung für Sportpsychologie, Dr. Geoffrey Schweizer und Dr. Lisa-Marie Schütz, sowie allen Doktorand*innen aus dem Keller danke ich für die schöne gemeinsame Zeit und den spannenden Austausch.

Von ganzem Herzen möchte ich meiner Familie, meinen Freund*innen und ganz besonders meinem Mann Jan dafür danken, dass sie mir stets liebevoll und unterstützend zur Seite standen. Die geteilte Freude über Erfolge und der Rückhalt in herausforderndenen Momenten haben maßgeblich zum Gelingen dieser Arbeit beigetragen.