Original Paper

Knowledge Gain and the Impact of Stress in a Fully Immersive Virtual Reality–Based Medical Emergencies Training With Automated Feedback: Randomized Controlled Trial

Marco Lindner; Tobias Leutritz, PhD; Joy Backhaus, MSc; Sarah König, MD; Tobias Mühling, MD

Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Würzburg, Germany

Corresponding Author:

Tobias Mühling, MD
Institute of Medical Teaching and Medical Education Research
University Hospital Würzburg
Joseph-Schneider-Strasse 2
Würzburg, 97070
Germany
Phone: 49 931-201-70627
Email: muehling_t@ukw.de

Abstract

Background: A significant gap exists in the knowledge and procedural skills of medical graduates when it comes to managing emergencies. In response, highly immersive virtual reality (VR)–based learning environments have been developed to train clinical competencies. However, robust evidence on how VR-based methods affect both short- and long-term learning outcomes, as well as physiological and perceived stress, remains limited.

Objective: This study aimed to assess the effectiveness of VR-based simulation training, augmented with automated feedback, compared with video seminars at improving emergency medical competency among medical students. Furthermore, the study investigated the relationship between learning outcomes and physiological stress markers. The evaluation of participants' perceived stress and estimated learning success was also performed to provide a more comprehensive insight into VR's potential role in emergency training.

Methods: In total, 72 senior medical students underwent VR-based emergency training (intervention) or viewed video seminars (control) on 2 topics (acute myocardial infarction and exacerbated chronic obstructive pulmonary disease) in an intraindividual crossover design. Levels of applied knowledge were assessed objectively by open-response tests pre- and postintervention and after 30 days. In addition, 2 electrodermal activity markers representing physiological stress response were measured during VR sessions using a wearable sensor. Participants also rated their estimated learning success and perceived stress. They also completed self-ratings of perceived stress and estimated learning success.

Results: Short-term knowledge gains were comparable between the VR (mean 26.6%, SD 15.3%) and control (mean 27.2%, SD 16%) condition. However, VR training produced significantly higher long-term knowledge gains (VR: mean 17.8%, SD 15.1% vs control: mean 11.9%, SD 18%; difference: –5.9, 95% CI –11.5 to –0.4). Overall retention scores were likewise higher for VR (mean 75.4%, SD 12.5%) than for video-based learning (mean 69.0%, SD 14.5%), a difference that was more pronounced in the myocardial infarction scenario. Participants rated the VR format as significantly more effective (mean 4.83, SD 0.41, on a 5-point scale) than the video seminar (mean 3.44, SD 1.00). While physiological stress markers increased during VR sessions, their correlation with knowledge gains was weak and negative. No significant relationship was detectable between perceived stress and objective learning outcomes.

Conclusion: VR-based simulation training with automated feedback may offer long-term learning advantages over a traditional video seminar in emergency-medicine education. Given the time constraints and resource limitations of clinical education, self-moderated VR-based learning may represent a valuable addition to conventional training methods. Future research could investigate the learning effects of VR scenarios regarding the retention of practical skills, as well as the impact of repeated or team-based scenarios.

Introduction

Although emergency medicine constitutes a fundamental component of most medical education curricula, existing evidence highlights deficiencies in differential diagnosis, initial treatment, and verbal handover during medical emergencies among graduates [1-3]. In response to such deficits, highly immersive virtual reality (VR) learning environments have been developed to train resuscitation [4], clinical reasoning [5], team communication among paramedics [6], and the use of specific medical equipment [7]. A significant advantage of these software-based simulations is the potential incorporation of automated feedback (eg, insights into beneficial or detrimental actions and end-of-session evaluations), which can enhance learning even in the absence of a tutor. The acceptance and educational achievement of VR-based training programs are generally rated highly by participants [8-10], fostering the belief that they will play a crucial role in the future of emergency training. So far, most studies focus on subjective parameters such as usability, satisfaction, and self-reported learning success [5,6,11-13]. However, self-reported learning outcomes are criticized as being prone to errors [14,15] and in the context of VR-based simulations in emergency medicine, objective learning outcomes have not been systematically investigated to meet the quality criteria of meta-analyses [9,16].

When evaluating the learning outcomes of VR-based training, it is important to consider that immersive environments may influence the learning process in various ways. Evidence strongly supports the notion that VR increases learner motivation and improves the contextualization of the learning experience [17]. In a recent research project of ours, we explored different types of perceived stress encountered during VR-based training in medical emergencies [13]. We identified two key stressors within the training scenario, exerting influence on participants' estimated learning success: (1) the control over the clinical situation and (2) the challenge of making decisions based on knowledge and skills. The first exhibited an inverse U-shaped relationship with estimated learning success, meaning that moderate control was beneficial, while too little or too much reduced estimated learning success. For the second, a higher perceived demand for knowledge and skills showed a negative correlation with estimated learning success. These observations align with those of similar studies, in which excessive stress was found to undermine cognitive abilities [18-20].

Stress, undeniably, encompasses both psychological and physiological components. A more comprehensive understanding of stress responses could thus be achieved through the investigation of perceived stress along with physiological stress markers [21]. Electrodermal activity (EDA) as a measure of sympathetic nervous system activity is frequently used to monitor psychophysiological activation in stressful conditions continuously [22]. EDA includes various parameters, among which the skin conductance level (SCL) and the frequency of nonspecific skin conductance responses (nsSCR) are particularly suitable for assessing background psychophysiological activation in the absence of specific stimuli [23]. Both physiological stress markers can be recorded by wearable sensors (eg, wristbands), causing minimal or no disruption to the simulation [24,25].

To evaluate objectively the learning outcomes of self-moderated VR-based emergency training with automated feedback and simultaneously to gain insight into the impact of different stress dimensions on the learning process, we aimed to answer the following questions. First, whether VR-based emergency simulation training is superior to video seminars for short- and long-term learning outcomes. Second, how perceived stress and physiological stress markers correlate during VR training sessions. Finally, to what extent perceived stress, physiological stress markers, and learning outcomes differ across various demographic groups and experience levels.

Methods

Fully Immersive VR-Based Simulation Training and Automated Feedback

STEP-VR (Simulation-based Training of Medical Emergencies for Physicians Using Virtual Reality; version 0.11b; ThreeDee GmbH, Munich, Germany) was used as the VR simulation of complex emergencies, which represents a joint development between a 3D visualization company and the University Hospital of Würzburg. The previously described hardware setup [13] comprised a Schenker XMG Core 15 Laptop (Schenker Technologies GmbH; chipset: Intel Core i7-9750H, 6×2.6 GHz; graphics adapter: Nvidia GeForce GTX 1650, 4 GB GDDR6 VRAM) and an Oculus Rift S (Lenovo Technologies and Oculus VR) VR head-mounted display.

For the intervention, 2 VR-based case scenarios from the subject of internal emergency medicine were used: (1) chest pain resulting from acute myocardial infarction and (2) dyspnea caused by exacerbation of chronic obstructive pulmonary disease. Both scenarios were developed by board-certified specialists in internal medicine and emergency medicine with additional qualifications in medical education, based on guidelines from the relevant professional societies [26,27]. The medical content of the scenarios and their evaluation have been described in detail in studies by Rickenbacher-Frey et al [12] and Mühling et al [13].

Furthermore, for this study, the software was complemented with an automated feedback system, providing

recommendations based on medical guidelines. This system included various feedback modalities:

1. Positive notifications for correctly executed actions.
2. Warnings for actions that could potentially be harmful.
3. Prompts for missed time-critical actions using a "scaffolding" approach, in which additional cues are provided to guide the user without directly specifying the required action.
4. Access to the results of diagnostics (eg, computed tomography scans) through a virtual representation of a patient documentation system.

5. A final evaluation presented in a checklist format, indicating the completion of relevant actions.

For points 1-4, brief messages, accompanied by an acoustic signal, were displayed immediately following the action. Users also had the option of accessing more detailed feedback through the virtual computer menu. Feedback components in the adapted version of STEP-VR are illustrated in Figure 1 and in real time during the use of the software are illustrated in Multimedia Appendix 1.

Figure 1. Feedback components in the adapted software. (A) Positive notifications (green) for correctly executed actions, (B) results of diagnostics in the virtual computer menu or (C) as direct output from medical devices (parameters calculated by dynamic physiology), and (D) final evaluation in checklist format.
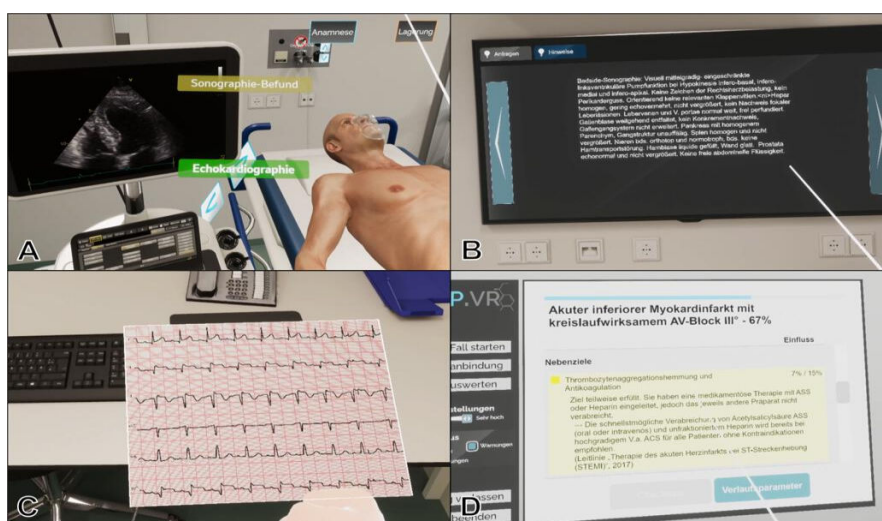
Study Design and Data Collection

The study was conducted at the skills lab of a German medical school with a 6-year medicine degree curriculum between June 2022 and May 2023. In order to align with the study design standards of recent meta-analyses, we adopted a randomized controlled design. To ensure a uniform baseline knowledge level, only medical students who had passed the written examination in internal medicine (regularly at the end of the eighth semester) were eligible to enroll. In addition, students with known epilepsy or severe simulation sickness were not considered eligible for this study. Recruitment was through semester mailing lists, with participants being offered a €15 (approximately US $16.98) voucher as an incentive. The simulation training itself took place outside regular teaching hours in the form of individual sessions. To avoid motivation bias among participants in a potential control group, we used an intraindividual cross-over design. This design required all participants to complete 2 scenarios, each lasting 35 minutes—one using VR (intervention condition) and the other in the form of a video seminar (control condition).

At the start of the intervention condition, all participants completed a 5-minute VR tutorial, during which they could familiarize themselves with the functionality of the empty virtual emergency room while listening to explanations via an audio commentary. Following the tutorial, participants engaged in 20 minutes of a self-directed learning activity within the scenario, during which automated feedback was provided (learning phase [LP]). During this self-moderated period, users decided which measures they wanted to perform. Afterward, participants reviewed the final evaluation checklist in the virtual doctor's office. To enhance content retention, they were given an additional 10 minutes to revisit the scenario and practice the medical procedures correctly (repetition phase [RP]). Throughout the whole session, a student tutor was available for technical support. As a control condition, video seminars for scenarios A and B lasting 35 minutes each were created comprising the same learning content in the form of a slide-based presentation. The lecturer interspersed content delivery with questions designed to foster reflection among the learners. Similar to the intervention, during the last 10 minutes participants were free to memorize the final summary or review specific slides from the presentation. All participants were randomly assigned to one of 2 groups. The first started with case A as VR scenario and then continued with case B as video seminar, while the second started with case A as video seminar and continued with case B as VR scenario. The random allocation sequence of whether case A or B was completed as an intervention or control condition was conducted by ML using computer-generated random numbers, who assigned participants to the intervention. Blocks of size 4 were used, ensuring that

the groups remained comparable throughout the recruitment process.

Students were asked to take applied knowledge tests with open-ended questions on the 2 case scenarios, both before and after the intervention or control session, following a pre- and posttest design. In addition, students completed a web-based survey, in which they rated their estimated learning success (after both the intervention and control) and perceived stress (only following the intervention), along with providing person-specific data. To assess knowledge retention, a retention test took place in a separate in-person session 30 days later under the supervision of the raters. Figure 2 provides an overview of the study design and data collection process.

Assuming a moderate effect size, a significance level of $P=.05$, and a power of 80%, the estimated sample size was approximately 63 participants. To account for a projected dropout rate of 10% before retention, we adjusted the total number of participants to 72.

Figure 2. Study design and data collection process for knowledge tests and estimated learning success, as well as physiological stress markers (skin conductance level and nonspecific skin conductance response) and perceived stress during intervention and control. Participants completed one scenario in VR and another as a video seminar. Est: estimated; nsSCR: nonspecific skin conductance response; SCL: skin conductance level; VR: virtual reality.
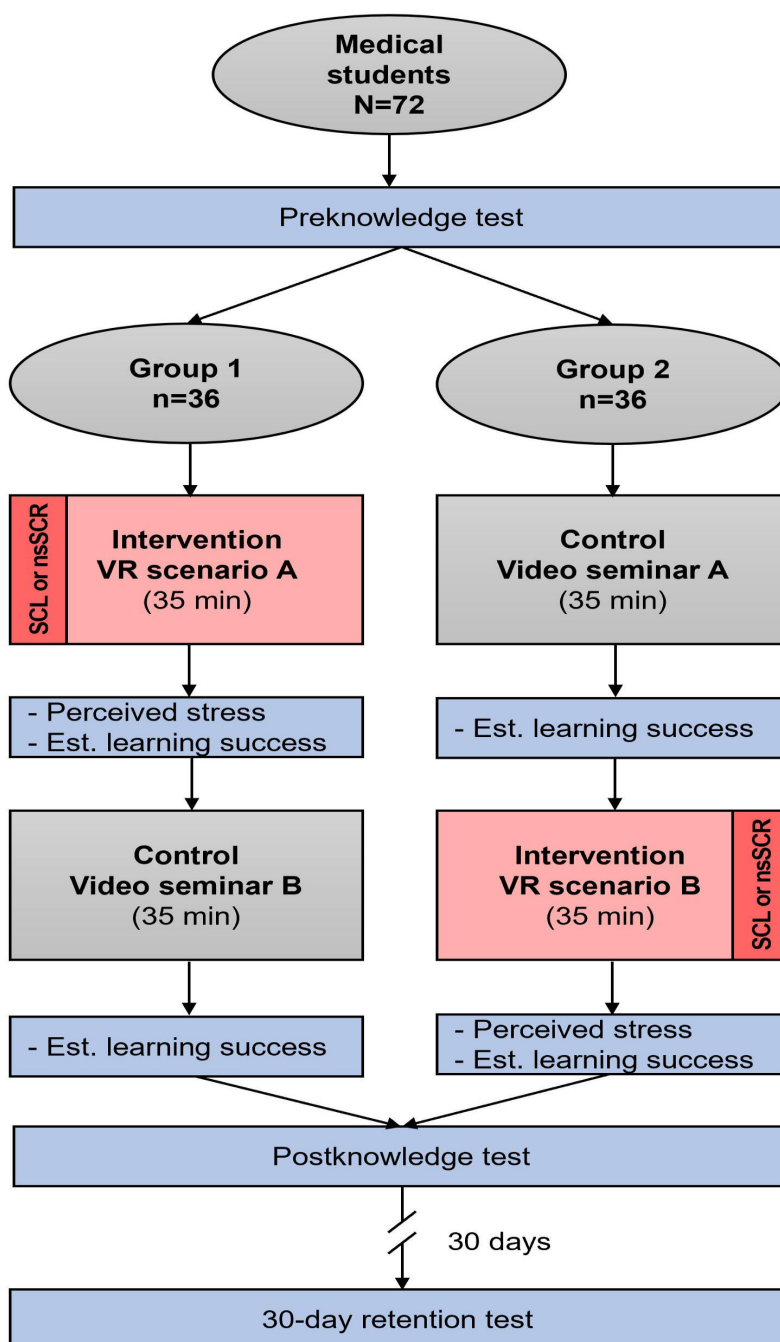
Questionnaires and Measures

Knowledge Tests

To assess the objective learning outcome, 8 open-ended questions were created based on the scenario content, adhering to the guidelines provided by professional societies [26,27]. These questions focused on applied knowledge related to diagnosing and initiating treatment in the 2 emergency scenarios. Questions covered topics that were taught in both the intervention and control conditions. A corresponding rubric was developed (Multimedia Appendix 2) and tested during a pilot phase with a small number (n=8) of participants. It was subsequently refined to achieve optimal discriminative power. Participants could score a maximum of 1-3 points for correct answers, depending on the difficulty and complexity of the content. Student performance was rated according to the percentage of correctly answered questions. The pre-, post-, and 30-day retention tests of knowledge all featured identical open-ended questions. Correct answers were never disclosed to participants. The responses were evaluated by 2 independent reviewers (a board-certified specialist in internal medicine with additional specialization in medical education [TM], and a medical student and trained skills laboratory tutor with 3 years of experience in rating of Objective Structured Clinical Examinations [ML]) in a blinded manner. The raters were unaware of which training modality (intervention or control) was delivered and at which time point (pre, post, or retention) the knowledge test was performed. In cases of disagreement between the raters, the respective answers were considered correct.

Short- and long-term knowledge gains were calculated as the differences between the posttest or 30-day retention test and the pretest, respectively.

Self-Assessment Questionnaires

Perceived Stress

In a previous study [13], we identified 3 main stress factors in VR-based emergency training: "sense of control" (ability to impact patient outcomes), "challenge to perform" (difficulty in making correct decisions using specific knowledge), and "social interaction to learn" (feeling of being watched and judged by others). Owing to time constraints and practical considerations, this study assessed these factors with just 1 relevant item each, based on their strong performance in the previous exploratory factor analysis. Students rated items on a 5-point Likert scale with 1 representing the minimum and 5 the maximum value. A separate overall rating for "total perceived stress" was evaluated based on a single item using the same scale. It should be noted that the questionnaire was used exclusively for the intervention (VR).

For the correlation analysis, the average of all 4 items (3 factors+total perceived stress) was calculated and labeled as "average perceived stress."

Estimated Learning Success

A similar approach was taken for the student self-assessment of estimated learning success represented by 2 factors previously identified as "didactic value" (participants' perceived general effectiveness of VR simulation) and "individual learning benefit" (the positive impact on the theoretical or practical skills of each participant). In this study, each factor was also evaluated with only 1 single item on the described 5-point Likert scale. Survey questionnaires with their items are listed in Multimedia Appendix 3. This questionnaire was used for both the intervention and the control condition.

To calculate the correlations, the average of the 2 items was taken and labeled as "average estimated learning."

Demographic Information

Demographic data were collected following the questionnaires on perceived stress and estimated learning success. It included questions on gender, academic semester, and previous experience in emergency medicine (eg, voluntary service and clinical rotation). In addition, participants were asked about the frequency of 3D-application use and their cumulative experience with VR applications.

Measurement of Physiological Stress Markers

EDA was measured using the Empatica E4 wristband, following the manufacturer's guide and good practice recommendations [23]. As there were no defined stimuli present during the VR scenarios, we opted to measure the continuous parameters SCL and nsSCR, suited to capture spontaneous fluctuations in EDA in the presence of an ongoing sustained stimulus over a period of time [28]. During completion of the tutorial, 3 minutes of baseline data were recorded for each participant. Recording during the intervention phase was divided according to the study design into the LP and RP. SCL and temperatures of all recordings were within normal ranges of 0.05-60 μs and 30-40°C, respectively [29]. Data were processed with R (Version 4.3.1, RRID:SCR_001905; R Core Team) using the wearables package [30]. We calculated the median values of SCL and nsSCR per minute during the respective observation phases (baseline, LP, and RP), resulting in 6 parameters.

For correlation analysis, we wanted to account for the interindividual variability of SCL and nsSCR, as well as for the level of stress generated using the VR environment itself (without emergency scenario). Thus, the values of the individual baseline recording were subtracted from LP and RP data for every participant, yielding 4 parameters (2 baseline-corrected physiological stress markers combined with 2 observation phases): SCL (LP), SCL (RP), nsSCR (LP), and nsSCR (RP).

Statistical Analysis

All data were analyzed by using descriptive statistics such as counts (both absolute and percentages), means, and SDs. Regarding primary outcome data from knowledge tests, differences and 95% CIs were also reported. As most of the parameters were not distributed normally (eg, knowledge tests and measures of EDA), repeated measures ANOVA was

not an option. Likewise, the Friedman test was not suitable for a 2×3 design (2 groups at 3 time points). Therefore, the Wilcoxon rank-sum test was used for comparisons between the intervention and control conditions. Spearman rank test was used to calculate correlation coefficients (ϱ). Curvilinear relationships were calculated using quadratic regression analyses, and the fit of the models was expressed by R^2. The interrater reliability was calculated using Cohen $\varkappa$ to quantify the agreement between the assessments of the independent reviewers. For all calculations excluding EDA data, GraphPad Prism (version 10.1.2, RRID:SCR_002798; GraphPad Software Inc) was used. Reporting was primarily conducted in accordance with the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) recommendations [31]. The corresponding checklist is provided in Checklist 1.

Ethical Considerations

Human Participant Ethics Review Approval

The local institutional review and ethics board judged the project as not representing medical or epidemiological research on human participants and as such adopted a simplified assessment protocol. The project was approved without any reservation under the proposal 20210917-01.

Trial Registration

This study was not registered as a clinical trial as it examined an educational intervention (VR-based training) rather than a clinical or therapeutic measure. The primary outcomes focused on knowledge acquisition and learning effectiveness rather than patient-related health outcomes.

Informed Consent

Students were informed about the study, and their participation was voluntary. Written informed consent was obtained in printed form from all participants, who were also provided with information on data processing for the analysis and the publication of results. Contact details were supplied for participants wishing to withdraw their consent to data processing. The decision to participate or not had no consequences on the student's academic progress.

Privacy and Confidentiality

Survey data from the questionnaires were collected anonymously using the EvaSys platform (Evasys GmbH). Data were processed and stored in accordance with local data protection laws.

Compensation Details

A €15 (approximately US $16.98) book voucher was handed to the participants upon completion of the retention test.

Results

Participants, Knowledge Tests, and Estimated Learning Success

A total of 72 participants were recruited for the study. All participants also completed the 30-day retention test, resulting in no dropouts. Person-specific factors are listed in Table 1. The gender distribution (46/72, 64% female participants) was representative of students in the degree program in medicine. Overall, there was only little previous experience with VR and 3D applications.

Table 1. Person-specific factors of participants.

Characteristics and experiences and subgroup	Value (N=72), n (%)
Gender	
Female	46 (64)
Male	25 (35)
Diverse	1 (1)
Academic semester	
8	20 (28)
9	23 (32)
10	26 (36)
>10	3 (4)
Experience in emergency medicine (eg, voluntary service and clinical rotation)	
Yes	32 (44)
No	40 (56)
Frequency of 3D-application use	
Never	49 (68)
Rarely	12 (17)
Sometimes	3 (4)

| Characteristics and experiences and subgroup	Value (N=72), n (%)
Frequent	7 (10)
Daily	1 (1)
Cumulative experience with VR[a]-applications	
None	30 (42)
<1 h	26 (36)
1-5 h	13 (18)
6-10 h	1 (1)
>10 h	2 (3)

[a]VR: virtual reality.

Interrater reliability was consistently high in the knowledge test with a Cohen ϰ of 0.927 averaged across all questions. Student performance is depicted in Figure 3. The previous knowledge of participants was similar in both the intervention and control condition for the pretest (intervention: mean 57.8%, SD 16.1%; control: mean 57.4%, SD 16.4%; difference: –0.4, 95% CI –5.8 to 4.9). Both conditions exhibited a similar short-term increase immediately following the respective learning sessions (posttest intervention: mean 84.4%, SD 9.6%; control: mean 84.6%, SD 11%; difference: 0.2, 95% CI –3.2 to 3.6). However, the 30-day retention test revealed differences; knowledge acquired in the VR scenario was retained significantly better than that acquired from the video seminar (intervention: mean 75.4%, SD 12.5%; control: mean 69.0%, SD 14.5%; difference: –6.4, 95% CI –10.9 to –1.9). At the level of scenarios, the difference in the 30-day retention test was more noticeable for the case of myocardial infarction (intervention: mean 75.2%, SD 15.3%; control: mean 67.4%, SD 13.9%; difference: –7.8, 95% CI –14.8 to –0.8). Conversely, for chronic obstructive pulmonary disease, the difference was slightly less pronounced (intervention: mean 75.6%, SD 9%; control: mean 70.5%, SD 15.3%; difference: –5.1, 95% CI –11.0 to 0.9). To account for previous knowledge, short-term gains (intervention: mean 26.6%, SD 15.3%; control: mean 27.2%, SD 16%; difference: 0.6, 95% CI –4.5 to 5.8) and long-term gains (intervention: mean 17.8%, SD 15.1%; control: mean 11.9%, SD 18%; difference: –5.9, 95% CI –11.5 to –0.4) were calculated.

On self-assessment by participants, the 2 items of estimated learning success in the VR scenarios approached the maximum possible value of 5 (didactic value: mean 4.83, SD 0.41; individual learning benefit: mean 4.85, SD 0.40). In contrast, these items in the control condition had lower scores, only slightly above average (didactic value: mean 3.44, SD 1.00; individual learning benefit: mean 3.15, SD 1.15). Furthermore, the differences between the conditions were statistically significant (each P<.001).

Figure 3. Knowledge test results as percentages of correct responses before (pre), immediately after (post), and 30 days after the intervention. Data are given in mean (box) with corresponding error bars (SD). The larger symbols and lines indicate overall results, while the smaller symbols detail the results for the individual scenarios. COPD: chronic obstructive pulmonary disease; MI: myocardial infarction; VR: virtual reality.
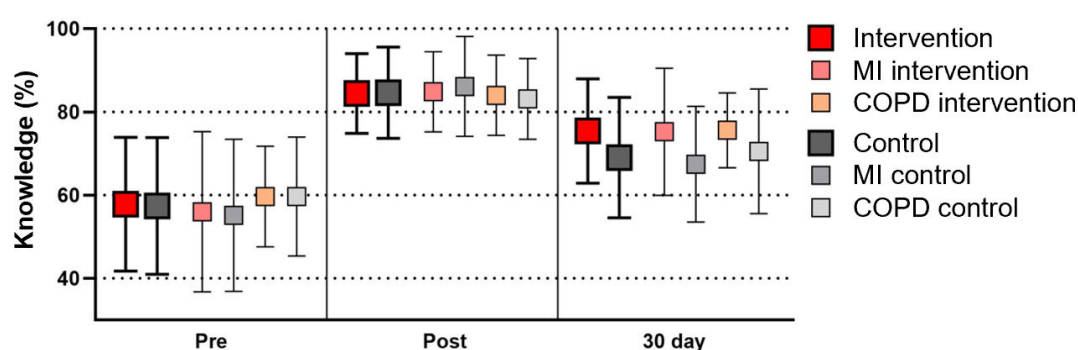
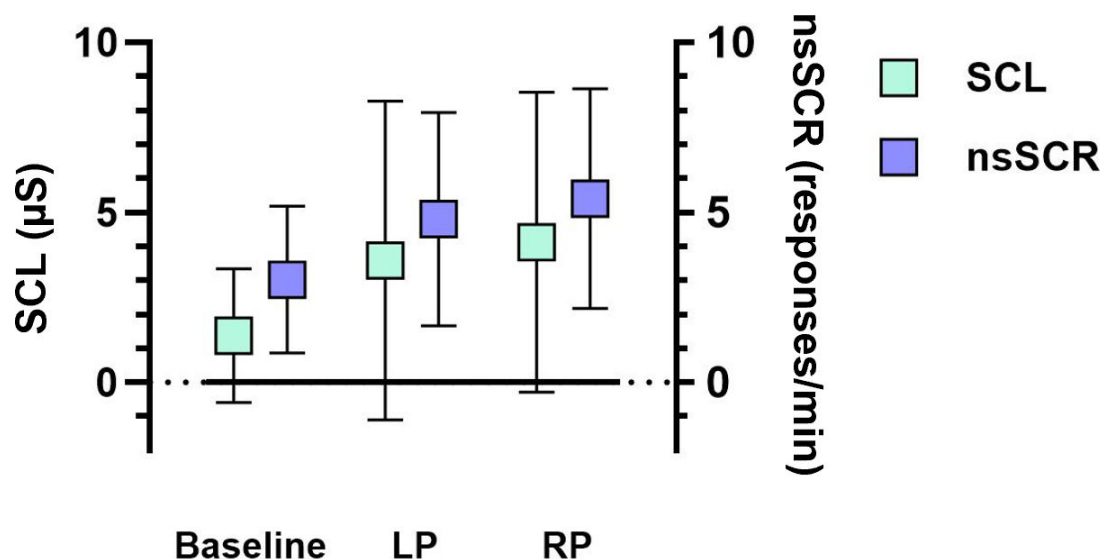## Physiological Stress Markers and Perceived Stress During VR Scenarios

For all participants, parameters of EDA were recorded during the completion of the VR scenarios, as illustrated in Figure 4. SCL and nsSCR exhibited increases from baseline values (mean 1.39, SD 1.98 μs and mean 3.05, SD 2.16 per min, respectively) to higher levels during the LP (mean 3.63, SD 4.71 μs and mean 4.81, SD 3.16 per min, respectively), further rising to their peak values during the RP (mean 4.16, SD 4.43 μs and mean 5.38, SD 3.26 per min, respectively).

The differences between the average values of LP and RP for both SCL and nsSCR and their respective baseline values were highly significant (both P<.001). Throughout the study, there was considerable variability in the values, as indicated by relatively high SDs.

Within the perceived stress questionnaire used in the VR scenarios, the factor "challenge to perform" received average ratings (mean 3.00, SD 1.08) among participants, followed by the "sense of control" (mean 2.51, SD 0.95) and low ratings for "social interaction to learn" (mean 1.37, SD 0.80).

Accordingly, moderate levels were reported for the "total perceived stress" item (mean 2.86, SD 0.94).

Figure 4. The progression of physiological stress markers SCL (μs) and nsSCR (responses/minute) during the completion of the virtual reality scenarios. Data are given in mean (box) with corresponding error bars (SD). LP: learning phase; nsSCR: nonspecific skin conductance response; RP: repetition phase; SCL: skin conductance level.


Analysis of Correlations

To investigate any potential influence of the baseline-corrected physiological stress markers (SCL and nsSCR) on short or long-term knowledge gains and students' self-assessments (average perceived stress and average learning gain), correlations were calculated (Figure 5). The parameters within the same modality (knowledge gains and stress markers) correlated well with each other. However, weak and nonsignificant negative correlations were observed between short- and long-term knowledge gains and baseline-corrected physiological stress markers recorded during the LP. The respective correlations were even less pronounced during the RP. Subsequently, quadratic regression analyses were performed to explore further the relationships between knowledge gains and average perceived stress as well as baseline-corrected physiological stress markers, respectively. While some plots visually suggested the presence of inverted U-shaped trendlines, all R^2 values fell below 0.2, indicating only weak correlation (data not illustrated).

Average perceived stress and average estimated learning exhibited only minimal correlation with knowledge gains and baseline-corrected physiological stress markers (all $\varrho < 0.2$). Notably, there was no correlation between average perceived stress and average estimated learning ($\varrho = -0.09$). All ϱ values are depicted in Figure 5 and corresponding P values are listed in Multimedia Appendix 4.

To evaluate the potential influence of (nominally and ordinally distributed) person-specific factors, group comparisons were conducted. No significant differences were found between gender, academic semester, previous experience level in emergency medicine, or experience with 3D or VR applications in relation to short- and long-term knowledge gain or baseline-corrected physiological stress markers.

Figure 5. Correlation matrix between stress markers, knowledge gains, average perceived stress, and average estimated learning. The heatmap scale (right) indicates the color coding for positive (blue) and negative (red) correlations. The more intense the color, the higher the correlation coefficient. Corresponding significance values are listed in Multimedia Appendix 4. LP: learning phase; nsSCR: nonspecific skin conductance response; RP: repetition phase, SCL: skin conductance level.
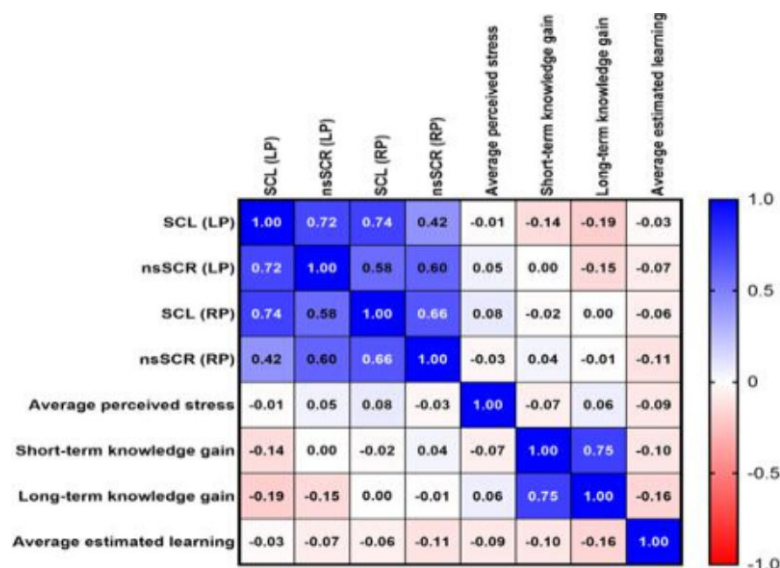
Discussion

Principal Findings

This study, conducted within the curricular framework of a degree course in medicine, used a randomized, controlled design to assess short- and long-term knowledge gains, as well as estimated learning success resulting from self-moderated learning in a complex VR-based medical emergency training. Furthermore, we investigated the influence of physiological stress markers and perceived stress among students.

Focusing on our primary objective, we found significantly greater knowledge retention after 30 days following a 35-minute self-moderated VR learning session with automated feedback, when compared with a video seminar of equivalent duration and content. The moderate effect size must also be considered in the context of the high previous knowledge (almost 60%) and the very explicit presentation of medical facts in the control condition. While VR-based programs are increasingly integrated into medical curricula, small sample sizes and heterogeneous study designs hinder clear conclusions on the effectiveness of learning with such programs [32]. To date, the number of randomized, controlled trials is limited, and results are partly contradictory. For VR-based resuscitation training, 2 studies reported different results. The VR-training group had either better or worse no-flow times, dependent on which control group was chosen (web-based versus "classic" basic life support course) [4,33]. Another study examining the effectiveness of VR-based pediatric intubation training reported similar scores compared with learning with a checklist [34]. Although it is known that context-specific knowledge tends to be anchored in memory better, we found no studies that included long-term outcomes by measuring knowledge or skill retention in VR-based emergency training. However, only training

methods that demonstrate long-term effects are likely to help close the knowledge and skill gaps observed in newly graduated physicians. By allowing learners to engage with highly immersive case scenarios—complete with automated feedback and the option for repetitive, self-directed practice—VR-based training addresses these deficits in ways that traditional methods alone often cannot [10,35]. Such simulations allow repeated practice of high-stakes actions and decision-making without endangering patients, which is especially relevant for complex or infrequent emergencies where hands-on exposure in real clinical settings can be hard to achieve [36]. Furthermore, time constraints in medical curricula frequently limit comprehensive hands-on training. In this regard, self-moderated VR sessions offer a resource-efficient alternative; learners can work independently, and tutors or faculty need only intervene selectively. This flexibility may reduce the strain on educators and infrastructure, making it feasible to revisit core competencies regularly in the future.

We could not observe a relevant correlation between knowledge gains and estimated learning success in this study. Even though subjective learning outcomes are frequently used in educational research [37,38] (including our own work [13]), a discrepancy between subjective and objective measures has been observed in various medical disciplines. Both medical students and junior doctors often face challenges in accurately self-assessing their own competence in various clinical skills [14,15]. Research from cognitive science suggests that self-assessment of learning is influenced more by affective factors, such as enthusiasm for a subject or a teaching method, rather than by objective knowledge gain [39]. This seems particularly plausible in the context of engaging and innovative teaching methods that predominantly elicit positive emotional responses among learners. We conclude that self-reported learning outcomes—especially

those from such methods—should ideally be supplemented or even substituted by objective assessments of competence.

The second objective of our study was to quantify physiological stress responses during VR-based learning and correlate them with levels of perceived stress. We successfully measured 2 physiological stress markers, SCL and nsSCR, in all participants. Both parameters exhibited similar increases during the scenarios, consistent with literature benchmarks [23]. Current evidence suggests that autonomic nervous system responses may be less pronounced in VR settings compared with real-life situations [40], yet more prominent than those observed in similar 2D screen settings [41]. Notably, SCLs were higher than average in stress-inducing VR-based disaster scenarios [42] and from VR environments designed specifically to trigger stress [43], indicating an authentic representation of the VR case scenarios. It is surprising that both physiological stress markers were lower during the LP (presumably the more stressful activity) than during the RP (where learning content was supposed to be reproduced). This is plausible for SCL, which can drift toward higher values over the recording period under certain circumstances (eg, change in skin temperature), as reported [23]. However, we also observed a similar increase in nsSCR during the RP, indicating that the rise from LP to RP was not merely an artifact of SCL. One possible explanation is that participants might have found it challenging to concentrate on reproducing the medical procedures during the RP, especially after just receiving (potentially poor) results from the final evaluation. This type of stress during the RP might differ qualitatively from that experienced during the LP, potentially explaining the differing correlation levels observed between the 2 phases (described further in this study). Interestingly, we found no correlation between the physiological stress markers and the identified items for perceived stress and the average perceived stress, respectively. This finding contrasts with a recent review, which reported that the majority (approximately 87%) of studies across various contexts found a positive association between EDA and perceived stress [44]. However, the reported degree of correlation varies strongly even within the specific context of simulation training, from no correlation in a simulated airway emergency [45] to a very high correlation during surgical skills training [46]. These findings underscore the need for further research into different measures of stress responses in VR settings (eg, heart rate variability or salivary cortisol) to improve our understanding and contextualize our results.

Finally, we explored the relationship between physiological stress markers and perceived stress in relation to knowledge outcomes. Our findings revealed a weak negative correlation between physiological stress markers recorded during the LP and long-term knowledge gain, while correlations with stress markers recorded during the RP were largely absent. It is generally recognized that affects and emotions play a pivotal role in influencing learning and memory performance [28]. Nonetheless, the exact relationship between physiological stress parameters and learning or performance outcomes remains somewhat ambiguous. For instance, in similar training sessions designed to establish

intraosseous access [47], conduct pediatric intubation [48], or execute simulated assessments [49], key physiological stress markers like salivary cortisol and heart rate variability displayed no correlation with performance outcomes. Contrastingly, broader data from simulation contexts generally suggest that stress adversely affects learning outcomes [20,50] and that stress management training can improve performance during critical simulated situations [30]. The lack of correlation between average perceived stress and average estimated learning success aligns with the divergent objective outcomes (knowledge and physiological stress response). Mathematically, this was expected given that subjective and objective measures within the same domain (knowledge or stress) did not correlate in our study. The reason we were unable to reproduce the correlations between the perceived stress and estimated learning success found in our previous study [13] may lie, in the simplified questionnaire structure, which replaced all items of a factor (eg, sense of control) with a single question. To elucidate the discrepancies among subjective and objective measures of stress and their relation to knowledge outcomes further, a future research direction could involve a more detailed qualitative approach, such as the repeated assessment of subjective stress dimensions throughout the VR scenario. This method could provide deeper insight into how stress dynamics uniquely affect individual learning trajectories in immersive virtual environments.

Strengths

The study benefited primarily from a relatively large sample size, which was representative of the medical student population at the study location. Knowledge was assessed at multiple time points and results largely corresponded across 2 scenarios with distinct content, thereby increasing the generalizability of the study. Established physiological stress markers, such as SCL and nsSCR, were recorded using validated tools alongside a baseline measurement, augmenting the robustness of the findings. The greatest strength, however, is the randomized, controlled, and single-blinded study design in which these measures were conducted.

Limitations

Several factors limit the generalizability of this study, however. First, the sample consisted of medical students from a single academic institution, which may not reflect broader demographics. Second, the learning sessions for participants were relatively short, with a duration of only 35 minutes. It would be interesting to explore the (dose-dependent) effects of longer or repeated learning sessions. Third, the participants in this study were not blinded regarding whether they were currently receiving the intervention or control condition, which may have influenced their performance. We aimed to minimize at least the motivational bias by using an intraindividual cross-over design, where each participant received both the intervention and control in different case scenarios. Finally, our study focused solely on applied knowledge. Given the simplified representation of invasive procedures in the program and the lack of haptic feedback from VR controllers, we chose not to assess practical skills. Future

research should investigate whether VR can effectively support long-term skill-based learning outcomes.

Conclusion

This study provides new insights into the durability and effectiveness of learning outcomes achievable through VR technology in emergency simulation training implemented in a medical degree program. The results suggest that the immersive, practical learning experiences offered by VR-based scenarios may yield long-term educational benefits, although research on longer retention periods is still lacking. Given the time constraints faced by clinical educators and the heterogeneity of curricula regarding complex emergencies, self-moderated VR-based learning emerges as a valuable complement to medical education.

While real-time measurement of physiological stress markers in VR scenarios would be desirable to improve the learning experience (eg, by dynamically adjusting the level of difficulty based on stress data), this study could not establish a correlation between measured stress and learning outcomes. More sophisticated methods may be required to clarify this relationship.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request. The training software STEP-VR (Simulation-based Training of Medical Emergencies for Physicians Using Virtual Reality) used in this study is licensed by ThreeDee GmbH. Requests regarding its use for research purposes should be directed at the company.

Authors' Contributions

TM was responsible for the conceptualization of the study, acquiring funding, supervising the execution, analyzing the knowledge and electrodermal activity (EDA) data, and writing the original draft. ML contributed to the conceptualization, conducted the investigation, and performed analysis of knowledge and EDA data. JB assisted in conceptualization of the study and performed statistical analysis. TL was involved in EDA data curation, formal analysis, and the review, and editing of the manuscript. SK also contributed to the conceptualization and participated in writing, review, and editing of the manuscript.

Conflicts of Interest

TM was involved in the software development process of STEP-VR (Simulation-based Training of Medical Emergencies for Physicians Using Virtual Reality). All other authors report no conflict of interest.

Multimedia Appendix 1

Feedback components in the adapted software (video).
[MP4 File (MP4 video File), 122263 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Rubric for open-ended questions (German and English) applied for knowledge tests.
[DOCX File (Microsoft Word File), 28 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Subjective measures questionnaire containing questions on perceived stress and estimated learning success (1 to 5 on a 5-point Likert scale) as well as questions on demographic information.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 3]

Multimedia Appendix 4

P values of the correlation matrix (Figure 3) between baseline-corrected parameters of the physiological stress response and knowledge tests at different time points, as well as the subjective measures of perceived stress and estimated learning success. *P* values below the significance level of .05 are printed in bold. LP: learning phase; RP: repetition phase; SCL: skin conductance level; nsSCR: nonspecific skin conductance response.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 4]

Checklist 1

CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist.

[PDF File (Adobe File), 54390 KB-Checklist 1]

References

1. Bugaj TJ, Nikendei C, Groener JB, et al. Ready to run the wards? - A descriptive follow-up study assessing future doctors' clinical skills. BMC Med Educ. Nov 12, 2018;18(1):257. [doi: 10.1186/s12909-018-1370-4] [Medline: 30419869]

2. Monrouxe LV, Grundy L, Mann M, et al. How prepared are UK medical graduates for practice? A rapid review of the literature 2009-2014. BMJ Open. Jan 13, 2017;7(1):e013656. [doi: 10.1136/bmjopen-2016-013656] [Medline: 28087554]

3. Burridge S, Shanmugalingam T, Nawrozzadeh F, Leedham-Green K, Sharif A. A qualitative analysis of junior doctors' journeys to preparedness in acute care. BMC Med Educ. Jan 13, 2020;20(1):12. [doi: 10.1186/s12909-020-1929-8] [Medline: 31931784]

4. Issleib M, Kromer A, Pinnschmidt HO, Süss-Havemann C, Kubitz JC. Virtual reality as a teaching method for resuscitation training in undergraduate first year medical students: a randomized controlled trial. Scand J Trauma Resusc Emerg Med. Feb 1, 2021;29(1):27. [doi: 10.1186/s13049-021-00836-y] [Medline: 33526042]

5. Mahling M, Wunderlich R, Steiner D, Gorgati E, Festl-Wietek T, Herrmann-Werner A. Virtual reality for emergency medicine training in medical school: prospective, large-cohort implementation study. J Med Internet Res. Mar 3, 2023;25:e43649. [doi: 10.2196/43649] [Medline: 36867440]

6. Lerner D, Mohr S, Schild J, Göring M, Luiz T. An immersive multi-user virtual reality for emergency simulation training: usability study. JMIR Serious Games. Jul 31, 2020;8(3):e18822. [doi: 10.2196/18822] [Medline: 32735548]

7. Wolff G, Bruno RR, Reiter M, Kantzow B, Kelm M, Jung C. Virtual reality device training for extracorporeal membrane oxygenation. Crit Care. Jul 2, 2020;24(1):390. [doi: 10.1186/s13054-020-03095-y] [Medline: 32616025]

8. Mühling T, Schreiner V, Appel M, Leutritz T, König S. Comparing virtual reality-based and traditional physical Objective Structured Clinical Examination (OSCE) stations for clinical competency assessments: randomized controlled trial. J Med Internet Res. Jan 10, 2025;27:e55066. [doi: 10.2196/55066] [Medline: 39793025]

9. Kim HY, Kim EY. Effects of medical education program using virtual reality: a systematic review and meta-analysis. Int J Environ Res Public Health. Feb 22, 2023;20(5):3895. [doi: 10.3390/ijerph20053895] [Medline: 36900904]

10. Mergen M, Graf N, Meyerheim M. Reviewing the current state of virtual reality integration in medical education - a scoping review. BMC Med Educ. Jul 23, 2024;24(1):788. [doi: 10.1186/s12909-024-05777-5] [Medline: 39044186]

11. Einloft J, Meyer HL, Bedenbender S, et al. Immersive medical training: a comprehensive longitudinal study of extended reality in emergency scenarios for large student groups. BMC Med Educ. Sep 9, 2024;24(1):978. [doi: 10.1186/s12909-024-05957-3] [Medline: 39252064]

12. Rickenbacher-Frey S, Adam S, Exadaktylos AK, Müller M, Sauter TC, Birrenbach T. Development and evaluation of a virtual reality training for emergency treatment of shortness of breath based on frameworks for serious games. GMS J Med Educ. 2023;40(2):Doc16. [doi: 10.3205/zma001598] [Medline: 37361243]

13. Mühling T, Späth I, Backhaus J, et al. Virtual reality in medical emergencies training: benefits, perceived stress, and learning success. Multimedia Systems. Aug 2023;29(4):2239-2252. [doi: 10.1007/s00530-023-01102-0]

14. Gabbard T, Romanelli F. The accuracy of health professions students' self-assessments compared to objective measures of competence. Am J Pharm Educ. Apr 2021;85(4):8405. [doi: 10.5688/ajpe8405] [Medline: 34283796]

15. Keicher F, Backhaus J, König S, Mühling T. Virtual reality for assessing emergency medical competencies in junior doctors - a pilot study. Int J Emerg Med. Sep 27, 2024;17(1):125. [doi: 10.1186/s12245-024-00721-2] [Medline: 39333858]

16. Liu JYW, Yin YH, Kor PPK, et al. The effects of immersive virtual reality applications on enhancing the learning outcomes of undergraduate health care students: systematic review with meta-synthesis. J Med Internet Res. Mar 6, 2023;25:e39989. [doi: 10.2196/39989] [Medline: 36877550]

17. Gutiérrez F, Pierce J, Vergara VM, et al. The effect of degree of immersion upon learning performance in virtual reality simulations for medical education. Stud Health Technol Inform. 2007;125(155-160):155-160. [Medline: 17377256]

18. Wetzel CM, Black SA, Hanna GB, et al. The effects of stress and coping on surgical performance during simulations. Ann Surg. Jan 2010;251(1):171-176. [doi: 10.1097/SLA.0b013e3181b3b2be] [Medline: 20032721]

19. McGraw LK, Out D, Hammermeister JJ, Ohlson CJ, Pickering MA, Granger DA. Nature, correlates, and consequences of stress-related biological reactivity and regulation in Army nurses during combat casualty simulation. Psychoneuroendocrinology. Jan 2013;38(1):135-144. [doi: 10.1016/j.psyneuen.2012.05.009] [Medline: 22710003]

20. Mauriz E, Caloca-Amber S, Córdoba-Murga L, Vázquez-Casares AM. Effect of psychophysiological stress and socio-emotional competencies on the clinical performance of nursing students during a simulation practice. Int J Environ Res Public Health. May 19, 2021;18(10):5448. [doi: 10.3390/ijerph18105448] [Medline: 34069709]

21. Gellisch M, Bablok M, Brand-Saberi B, Schäfer T. Neurobiological stress markers in educational research: a systematic review of physiological insights in health science education. Trends Neurosci Educ. Dec 2024;37:100242. [doi: 10.1016/j.tine.2024.100242] [Medline: 39638492]

22. Bhoja R, Guttman OT, Fox AA, Melikman E, Kosemund M, Gingrich KJ. Psychophysiological stress indicators of heart rate variability and electrodermal activity with application in healthcare simulation research. Simul Healthc. Feb 2020;15(1):39-45. [doi: 10.1097/SIH.0000000000000402] [Medline: 32028446]

23. Boucsein W. Electrodermal Activity. Springer; 2012. ISBN: 978-1-4614-1125-3

24. Fadeev KA, Smirnov AS, Zhigalova OP, Bazhina PS, Tumialis AV, Golokhvast KS. Too real to be virtual: autonomic and EEG responses to extreme stress scenarios in virtual reality. Behav Neurol. 2020;2020(5758038):5758038. [doi: 10.1155/2020/5758038] [Medline: 32256856]

25. Halbig A, Latoschik ME. A systematic review of physiological measurements, factors, methods, and applications in virtual reality. Front Virtual Real. 2021;2. [doi: 10.3389/frvir.2021.694567]

26. Ibanez B, James S, Agewall S, et al. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J. Jan 7, 2018;39(2):119-177. [doi: 10.1093/eurheartj/ehx393] [Medline: 28886621]

27. Vogelmeier C, Buhl R, Burghuber O, et al. Leitlinie zur Diagnostik und Therapie von Patienten mit chronisch obstruktiver Bronchitis und Lungenemphysem (COPD) [Article in German]. Pneumologie. Apr 2018;72(4):253-308. [doi: 10.1055/s-0043-125031] [Medline: 29523017]

28. Posada-Quintero HF, Chon KH. Innovations in electrodermal activity data collection and signal processing: a systematic review. Sensors (Basel). Jan 15, 2020;20(2):479. [doi: 10.3390/s20020479] [Medline: 31952141]

29. Kleckner IR, Jones RM, Wilder-Smith O, et al. Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. IEEE Trans Biomed Eng. Jul 2018;65(7):1460-1467. [doi: 10.1109/TBME.2017.2758643] [Medline: 28976309]

30. de Looff P, Duursma R, Noordzij M, et al. Wearables: an R package with accompanying Shiny application for signal analysis of a wearable device targeted at clinicians and researchers. Front Behav Neurosci. 2022;16:856544. [doi: 10.3389/fnbeh.2022.856544] [Medline: 35813597]

31. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res. Dec 31, 2011;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]

32. Wu TC, Ho CTB. A scoping review of metaverse in emergency medicine. Australas Emerg Care. Mar 2023;26(1):75-83. [doi: 10.1016/j.auec.2022.08.002] [Medline: 35953392]

33. Moll-Khosrawi P, Falb A, Pinnschmidt H, Zöllner C, Issleib M. Virtual reality as a teaching method for resuscitation training in undergraduate first year medical students during COVID-19 pandemic: a randomised controlled trial. BMC Med Educ. Jun 22, 2022;22(1):483. [doi: 10.1186/s12909-022-03533-1] [Medline: 35733135]

34. Agasthya N, Penfil S, Slamon N. Virtual reality simulation for pediatric airway intubation readiness education. Cureus. Dec 13, 2020;12(12):e12059. [doi: 10.7759/cureus.12059] [Medline: 33447488]

35. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA. Sep 7, 2011;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]

36. Li L, Yu F, Shi D, et al. Application of virtual reality technology in clinical medicine. Am J Transl Res. 2017;9(9):3867-3880. [Medline: 28979666]

37. Ochsmann EB, Zier U, Drexler H, Schmid K. Well prepared for work? Junior doctors' self-assessment after medical education. BMC Med Educ. Nov 24, 2011;11(99):99. [doi: 10.1186/1472-6920-11-99] [Medline: 22114989]

38. Bußenius L, Harendza S, van den Bussche H, Selch S. Final-year medical students' self-assessment of facets of competence for beginning residents. BMC Med Educ. Feb 7, 2022;22(1):82. [doi: 10.1186/s12909-021-03039-2] [Medline: 35130891]

39. Sitzmann T, Ely K, Brown KG, Bauer KN. Self-assessment of knowledge: a cognitive learning or affective measure? AMLE. Jun 2010;9(2):169-191. [doi: 10.5465/amle.9.2.zqr169]

40. Marín-Morales J, Higuera-Trujillo JL, Guixeres J, Llinares C, Alcañiz M, Valenza G. Heart rate variability analysis for the assessment of immersive emotional arousal using virtual reality: comparing real and virtual scenarios. PLoS ONE. 2021;16(7):e0254098. [doi: 10.1371/journal.pone.0254098] [Medline: 34197553]

41.   Kuhne C, Kecelioglu ED, Maltby S, et al. Direct comparison of virtual reality and 2D delivery on sense of presence, emotional and physiological outcome measures. Front Virtual Real. 2023;4. [doi: 10.3389/frvir.2023.1211001]

42.   Prachyabrued M, Wattanadhirach D, Dudrow RB, Krairojananan N, Fuengfoo P. Toward virtual stress inoculation training of prehospital healthcare personnel: a stress-inducing environment design and investigation of an emotional connection factor. Presented at: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR); Mar 23-27, 2019:671-679; Osaka, Japan. [doi: 10.1109/VR.2019.8797705]

43.   Tao K, Huang Y, Shen Y, Sun L. Automated stress recognition using supervised learning classifiers by interactive virtual reality scenes. IEEE Trans Neural Syst Rehabil Eng. 2022;30(2060-2066):2060-2066. [doi: 10.1109/TNSRE.2022.3192571] [Medline: 35857724]

44.   Klimek A, Mannheim I, Schouten G, Wouters EJM, Peeters MWH. Wearables measuring electrodermal activity to assess perceived stress in care: a scoping review. Acta Neuropsychiatr. Mar 24, 2023;37:e19. [doi: 10.1017/neu.2023.19] [Medline: 36960675]

45.   Phitayakorn R, Minehart RD, Hemingway MW, Pian-Smith MCM, Petrusa E. Relationship between physiologic and psychological measures of autonomic activation in operating room teams during a simulated airway emergency. The American Journal of Surgery. Jan 2015;209(1):86-92. [doi: 10.1016/j.amjsurg.2014.08.036]

46.   Georgiou K, Larentzakis AV, Khamis NN, Alsuhaibani GI, Alaska YA, Giallafos EJ. Can wearable devices accurately measure heart rate variability? A systematic review. Folia Med (Plovdiv). Mar 1, 2018;60(1):7-20. [doi: 10.2478/folmed-2018-0012] [Medline: 29668452]

47.   Ghazali DA, Darmian-Rafei I, Ragot S, Oriot D. Performance under stress conditions during multidisciplinary team immersive pediatric simulations. Pediatr Crit Care Med. Jun 2018;19(6):e270-e278. [doi: 10.1097/PCC.0000000000001473] [Medline: 29432402]

48.   Bensouda B, Mandel R, Mejri A, Lachapelle J, St-Hilaire M, Ali N. Effect of an audience on trainee stress and performance during simulated neonatal intubation: a randomized crossover trial. BMC Med Educ. Oct 3, 2018;18(1):230. [doi: 10.1186/s12909-018-1338-4] [Medline: 30285715]

49.   Stein C. The effect of clinical simulation assessment on stress and anxiety measures in emergency care students. Afr J Emerg Med. Mar 2020;10(1):35-39. [doi: 10.1016/j.afjem.2019.12.001] [Medline: 32161710]

50.   Mills B, Carter O, Rudd C, Claxton L, O'Brien R. An experimental investigation into the extent social evaluation anxiety impairs performance in simulation-based learning environments amongst final-year undergraduate nursing students. Nurse Educ Today. Oct 2016;45(9-15):9-15. [doi: 10.1016/j.nedt.2016.06.006] [Medline: 27429397]

Abbreviations

CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
EDA: electrodermal activity
LP: learning phase
nsSCR: nonspecific skin conductance response
RP: repetition phase
SCL: skin conductance level
STEP-VR: Simulation-based Training of Medical Emergencies for Physicians Using Virtual Reality
VR: virtual reality