I.J. Image, Graphics and Signal Processing, **2026, 2, 151-207**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2026.02.10

Information Technology for VR Training Evaluation with First Aid Skills Improvement to Detecting Human Resource Damage in Emergencies based on Behavioural Methods

Sofia Chyrun
Information and Communication Technologies Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine
E-mail: sofiia.chyrun.ik.2022@lpnu.ua
ORCID iD: https://orcid.org/0000-0002-2829-0164

Victoria Vysotska
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine
Combating Cybercrime Department, Kharkiv National University of Internal Affairs, Kharkiv, 61080, Ukraine
E-mail: Victoria.A.Vysotska@lpnu.ua
ORCID iD: https://orcid.org/0000-0001-6417-3689

Lyubomyr Chyrun
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine
Computer Science Department, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
Applied Mathematics Department, Ivan Franko National University of Lviv, Lviv, 79000, Ukraine
E-mail: lyubomyr.v.chyrun@lpnu.ua
ORCID iD: https://orcid.org/0000-0002-9448-1751

Dmytro Uhryn
Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: d.ugryn@chnu.edu.ua
ORCID iD: https://orcid.org/ 0000-0003-4858-4511

Zhengbing Hu
School of Computer Science and Artificial Intelligence, Hubei University of Technology, Wuhan, China
E-mail: drzbhu@gmail.com
ORCID iD: https://orcid.org/0000-0002-6140-3351

Yurii Ushenko*
Department of Physics, Shaoxing University, Shaoxing, Zhejiang Province 312000, China
Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: y.ushenko@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0003-1767-1882
*Corresponding Author

Received: 16 August 2025; Revised: 15 October 2025; Accepted: 26 November 2025; Published: 08 April 2026

Abstract: Traditional first aid preparation methods often fail to reproduce realistic stress levels and to simulate visual difficulty in identifying lesions in critical situations. In emergencies, delays in recognising injuries or errors in protocols result in critical losses of human resources. The use of computer graphics and virtual reality technologies enables you to create a safe yet highly realistic environment for rescuers to test and improve their skills. The article presents an integrated methodological framework for assessing the effectiveness of VR first-aid training in conditions of damage to civilian infrastructure. The main focus is on developing mathematical models and algorithms to identify and evaluate the quality of rescuers' actions by analysing digital interaction signals in a virtual environment. A composite efficiency indicator is proposed that combines normalised parameters for reaction time, manipulation accuracy, stress level, and immersion. The work aims to formalise a mathematical model to assess the effectiveness of VR training in developing skills for lesion identification and first aid provision, using quantitative metrics. The study aims to identify statistically significant differences in learning speed and skill retention between groups using VR simulations and traditional methods. The project aims to validate innovative content creation methods, including mobile photogrammetry, to visualise damaged infrastructure and victim models. The study used a comprehensive approach that includes mobile photogrammetry and generative neural networks to create a library of 3D assets with varying degrees of detail. Performance score is based on composite indicator that integrates normalised data on reaction time, manipulation accuracy, error count, immersion rate. Linear mixed models, exponential approximations, and bootstrap estimation of effect stability were used to analyse hierarchical data and individual learning trajectories. The experimental part includes the use of mobile photogrammetry and generative neural networks to create realistic 3D models of affected environments and identify types of injuries (bleeding, burns, unconsciousness). To analyse the dynamics of learning and maintaining skills, models with mixed effects and exponential forgetting curves are used. The results confirm that the use of VR technologies provides a statistically significant acceleration in the development of automated skills for lesion identification and assistance compared to traditional methods. The proposed approach is a scalable tool for preparing civil and rescue services to act in critical situations. Experimental data showed that the integral performance score in the VR group increased from 0.42 ± 0.10 to 0.76 ± 0.08, while in the control group it increased from only 0.40 ± 0.09 to 0.55 ± 0.10 ($p < 0.001$). The largest effect was observed in the bleeding arrest scenario, where the effect size (Cohen's d) reached 2.3. The analysis of forgetting curves confirmed the superiority of VR: the skill loss rate in the VR group was 0.25, providing knowledge retention 1.8 times longer than in the control group (0.45). The study confirmed that VR simulations significantly accelerate the formation of automated behaviour patterns and reduce reaction time in extreme conditions. The proposed mathematical assessment model provides objective feedback and standardisation of the rescue training process. The results indicate the high practical value of introducing such tools into training programs for civilian and military structures to minimise losses in real emergencies.

Index Terms: Human Resource Lesion Identification, Digital Interaction Signal Processing, Photogrammetry, 3D Modelling, Virtual Reality, VR, Mathematical Modelling of Learning, Composite Performance Metric, Dynamic Systems Analysis, Learning Curves, Emergencies, Tactical Medicine, Mixed-Effects Models, Graphic Data Compression, Graphics Data Visualisation

1. Introduction

Modern emergencies caused by the destruction of civilian infrastructure place critical demands on the speed and accuracy of human resource damage identification. Successful first aid in such conditions directly depends on the ability of the rescuer to instantly recognise the types of injuries (massive bleeding, respiratory failure, traumatic shock) and act according to clear algorithms in conditions of high psychological stress. Traditional teaching methods are often limited by the static nature of mannequins and the inability to reproduce the realistic visual dynamics of lesions, creating a gap between theoretical training and practical readiness for action in real crisis conditions. The development of virtual reality (VR) technologies and intelligent graphics signal processing opens up new possibilities for creating immersive simulations that allow you to identify lesions using highly realistic visual content. Of particular importance are the methods for rapidly generating 3D assets through mobile photogrammetry and neural networks, which allow simulating specific scenarios of destruction and trauma with a high degree of reliability. It not only increases immersion (Presence) but also allows the system to automatically log digital signals of user interaction with objects, turning them into objective metrics of learning quality. his paper examines a formalised mathematical model for evaluating the effectiveness of VR training, based on the analysis of composite performance indicators and learning dynamics through exponential laws. The focus is on identifying lesions in four scenarios: bleeding arrest, cardiopulmonary resuscitation (CPR), smoke triage, and mass casualties. The study is based on the hypothesis that the VR environment provides better lesion identification and faster development of automated skills through multisensory feedback and the ability to repeatedly execute critical protocols. In this study, immersion and stress are conceptualised as contextually induced psychological states that manifest in response to the type of learning environment and that predict or accompany task performance. Their use as predictors, mediators, or outcomes is clearly limited to appropriate temporal and analytical levels, thereby eliminating conceptual ambiguity and the need for a suggestive cyclical explanation.

The purpose of this study is to develop and substantiate a mathematical model and a software and hardware complex based on virtual reality to increase the accuracy and speed of identifying human resource injuries in emergencies. The study aims to formalise quantitative metrics for evaluating rescuers' actions, validate methods for generating realistic graphic content of affected environments, and establish statistically significant patterns in the dynamics of professional skills formation. According to the sections of work, the following tasks are defined to achieve the goal:

1. Development of mathematical formalisation of the evaluation process:

– Develop an integral performance indicator that takes into account reaction time, identification accuracy and critical errors in the provision of assistance.

– Formulate hypotheses about the advantages of the VR environment in preserving skills compared to traditional methods.

– To build a model of learning dynamics based on an exponential law to predict the rate of mastering assistance protocols.

2. Developing and Visualising Immersive Content:

– Implement algorithms for creating 3D models of damaged infrastructure and specific lesions (bleeding, burns, injuries) using mobile photogrammetry and generative neural networks.

– Design a system for automatic logging of digital user interaction signals to form an objective database of results.

– Optimise graphics resources to ensure high frame rates and minimise virtual disorientation when using VR helmets.

3. Statistical analysis and verification of results:

– Conduct a comparative analysis of the groups using the Wilcoxon criterion, the Welch t-test and the estimate of the size of the Cohen effect.

– Apply Mixed-effects models to detect interactions between the learning type and the sessions (Group × Session).

– To assess the stability of the results obtained by constructing 95% confidence intervals using the bootstrap method.

4. Modelling of retention and stress resistance processes:

– Investigate the rate of forgetting of lesion identification skills using the Ebbinghaus model and calculate the half-life of the skill.

– Establish a correlation between the level of immersion (Presence), psychophysiological stress and the final effectiveness of the rescuer's actions based on the Yerkes-Dodson law.

The object of the study is the processes of identifying human resource injuries and the development of practical skills for providing first aid in emergencies resulting from damage to civilian infrastructure. The subject of the study is mathematical models, methods for processing digital interaction signals, and virtual graphic tools that enable lesion identification, automated assessment of the effectiveness of rescuers' actions, and modelling the dynamics of learning and maintaining skills in the VR environment.

Below is a structured description of the scientific novelty in the main areas of your work:

1. Mathematical formalisation of efficiency assessment. For the first time, a composite metric, S, has been developed and theoretically substantiated that integrates normalised indicators of reaction time, accuracy, critical errors, and immersion level into a single analytical system. It allows you to move from a subjective instructor assessment to an objective digital measurement of professional competence.

2. Technologies for identifying and visualising lesions. The method of rapidly generating immersive content through a combination of mobile photogrammetry and neural networks to create models of "destructive realism" (damaged buildings, specific injuries), which is critically important for the accuracy of lesion identification, is substantiated. The proposed approach minimises computational costs while maintaining high diagnostic reliability of the virtual environment.

3. Methodology of learning analysis (Mixed-Effects & Bootstrap). Linear mixed-effects models were introduced for the analysis of first aid training, enabling the statistical separation of VR technology's influence from participants' individual psychophysiological characteristics. The use of bootstrap validation ensured the stability of the positive effect of VR, even in the limited-sample conditions typical of pilot studies.

4. Cognitive modelling of retention and stress. Quantitative patterns of skill preservation have been established: VR training reduces forgetting rates and increases the half-life of the skill by 1.8 times compared to traditional methods. The effect of stress moderation via the Presence indicator was revealed: a high level of immersion expands the rescuer's zone of optimal performance, according to the Yerkes-Dodson law.

Based on the analysis of the sections of the article, the goal of increasing the efficiency of lesion identification and the methodology of statistical analysis, the following research hypotheses are formulated:

1. The main hypothesis (H_1) is the effectiveness of learning. The use of VR technologies to identify lesions and practice first aid protocols results in higher practical skills than traditional training methods. H_0 (Null hypothesis) – there are no statistically significant differences in results between the VR group and the control group. – Validation indicators: increase in composite score S, reduction of RT reaction time, and increase in accuracy of actions A.

2. The hypothesis of learning dynamics (H_2) is the speed of assimilation. The VR environment provides a higher rate of skill formation for identification and assistance, as described by the parameter k in the exponential learning model.

– H_2a – the coefficient of the learning speed in the VR group is significantly higher than that in the control group.

– H_2b – the interaction of the "Group × Session" factors is positive, which indicates an accelerated improvement in the results with each subsequent session in the VR group.

3. Retention hypothesis (H_3) – long-term retention of skills. Lesion identification skills acquired through VR

simulations last longer and exhibit a lower forgetting rate than those acquired through theoretical or dummy training.

– H_3a – the skill retention rate after a certain time in the VR group will be statistically higher.

– H_3b – VR training provides a higher level of "residual knowledge" after a long break in practice.

4. The hypothesis of psychophysiological adaptation (H_4) - stress and presence. A high level of presence in the VR environment effectively simulates the learning process, enabling you to maintain high efficiency in lesion identification even under increased stress.

– H_4a – a direct positive correlation between the immersion level and the final result of S.

– H_4b (Yerkes-Dodson Law) – VR simulation shifts the optimum of psychophysiological arousal to the right, allowing the rescuer to maintain performance at higher levels of stress load compared to the control group.

5. Technical validity hypothesis (H_5) – photorealism and identification. The use of mobile photogrammetry and generative design tools to create damage models provides sufficient visual fidelity for the successful recognition of wound types in virtual space. Criterion: the accuracy of identification of specific lesions (for example, bleeding or burns) in VR corresponds to or exceeds the indicators when working with traditional dummies.

The statistical analysis is now organised into three tiers, each with a distinct and non-overlapping role. Tier 1 serves as the primary confirmatory analysis and is solely responsible for hypothesis testing. It includes predefined performance metrics (reaction time, accuracy, retention) and fixed statistical tests (an independent-samples t-test or a Mann–Whitney U test when normality assumptions are violated). Only Tier 1 results are used to accept or reject the primary hypotheses.

Tier 2 comprises secondary robustness analyses, applied only when non-trivial effects are observed in Tier 1. These analyses include ANCOVA and linear mixed-effects models, used to assess the stability of effect direction and magnitude under covariate control and repeated-measures structure. Importantly, Tier 2 analyses are not used to redefine statistical significance or overturn Tier 1 conclusions.

Tier 3 is explicitly designated as exploratory and descriptive. It includes correlation analysis, regression models, bootstrap confidence intervals, and survival analysis. Results from this tier are clearly labelled as exploratory, are not used for confirmatory inference, and serve solely to generate hypotheses for future studies.

Although formal preregistration was not conducted prior to data collection, the hierarchical analysis plan described above was fixed before hypothesis testing and is now reported in full to ensure transparency and reproducibility. All deviations from the primary analysis pathway are explicitly marked as secondary or exploratory.

This structured approach substantially limits researchers' degrees of freedom, mitigates the risk of false-positive findings, and clarifies the inferential status of each reported result. The manuscript reflects this hierarchy consistently across the Methods and Results sections. The article has the following structure:

– Chapter 2 describes the problem statement and the mathematical apparatus of the research.
– Chapter 3 is devoted to the analysis of related works in the field of medical simulations.
– Chapter 4 details the data collection methodology and architecture of the VR system.
– Chapter 5 describes experimental testing of lesion imaging tools.
– Chapter 6 presents the results of statistical analysis and a discussion of the data obtained.
– Chapter 7 introduces competing explanations and the falsification strategy.
– Chapter 8 is devoted to the discussion of the results of the study.

2. Problem Statement

The aim is to formalise a mathematical model for evaluating the effectiveness of VR training in developing first aid skills and to determine statistically significant differences between the VR and control groups using quantitative metrics, training models, and statistical analysis methods. Let $i = 1,2,\dots,N$ – index of the participant, $t = \{0,1,2\}$ – measurement stages (pre, post, retention), $g \in \{0,1\}$ – Training group ($g = 1$ – VR, $g = 0$ – Control). For each participant, a measurement vector is observed:

$$x_{i,t} = \left(RT_{i,t}, A_{i,t}, Err_{i,t}, Prs_{i,t}, Str_{i,t}, S_{i,t}\right),$$

where RT – is reaction time, A – is accuracy of actions, Err – is the number of critical errors, Prs – is immersion level, Str – is stress level, S – is composite efficiency score.

The composite score is formed as:

$$S_{i,t} = w_1\,\widehat{A_{i,t}} + w_2\,\widehat{RT_{i,t}} + w_3\left(1 - \widehat{Err_{i,t}}\right) + w_4\,\widehat{Prs_{i,t}},$$

where w_j – are the weights of the metrics, and $\hat{\cdot}$ – is the normalisation to the range [0,1].

The central hypothesis is that it is necessary to assess whether VR training provides a significant increase in the effectiveness indicator:

$$H_0\colon \mu_{VR} = \mu_C, \quad H_1\colon \mu_{VR} > \mu_C,$$

where $\mu_{VR} = E\left[S_{i,t}^{(VR)}\right]$, $\mu_C = E\left[S_{i,t}^{(C)}\right]$.

The exponential model of learning dynamics describes the process of skill formation:

$$S_i(t) = S_{max} - (S_{max} - S_0)e^{-kt},$$

where S_0 – initial skill level (pre-test), k – coefficient of learning speed, S_{max} – asymptotic skill level. For the two groups, the models look like:

$$S^{VR}(t) = S_{max}^{VR} - (S_{max}^{VR} - S_0^{VR})e^{-k_{VR}t}\text{t}),\ S^C(t) = S_{max}^C - (S_{max}^C - S_0^C)e^{-k_C t}\text{t}).$$

The task is to evaluate the parameters $\theta = \{k_{VR}, k_C, S_{max}^{VR}, S_{max}^C\}$, and check the inequality $k_{VR} > k_C$. Taking into account individual differences, a mixed-effects model is used for repeated measurements:

$$S_{i,t} = \beta_0 + \beta_1 g_i + \beta_2 t + \beta_3 (g_i \cdot t) + u_i + \varepsilon_{i,t},$$

where $u_i \sim \mathcal{N}(0, \sigma_u^2)$ – random effect of the participant, $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2)$, β_3 – key interaction parameter (corresponds to Group × Session). The task is to check $H_0: \beta_1 = 0,\ H_0: \beta_3 = 0$.

The data obtained contain large effects (up to). Formally:$d = 2.3$

$$d = \frac{\bar{S}_{VR} - \bar{S}_C}{\sqrt{\frac{1}{2}(\sigma_{VR}^2 + \sigma_C^2)}}.$$

The task is to calculate for each scenario and assess whether the effect is significant ($d > 0.8$– large effect). To evaluate the stability of the mean difference: $\Delta = \bar{S}_{VR} - \bar{S}_C$. Multiple bootstrap samples are built:

$$\Delta^{(b)} = \bar{S}_{VR}^{(b)} - \bar{S}_C^{(b)}, \quad b = 1, \dots, B.$$

95% confidence interval: $CI_{95\backslash\%} = \left[\Delta^{(2.5\backslash\%)},\ \Delta^{(97.5\backslash\%)}\right]$. The task is to check $0 \notin CI_{95\backslash\%}$. As a result, you need to find the parameters: $\Theta = \{\mu_{VR}, \mu_C, k_{VR}, k_C, \beta_1, \beta_3\}$ – such that:

– VR has a higher average skill level: $\mu_{VR} > \mu_C$;
– learns faster: $k_{VR} > k_C$;
– demonstrates a significant group effect: $\beta_1 > 0$;
– has accelerated learning in time (Group×Session): $\beta_3 > 0$;
– the difference between the groups is stable: $0 \notin CI_{95\backslash\%}$.

It constitutes a formally defined scientific task of the study.

3. Related Works

Over the past decade, the application of virtual reality (VR) technologies in medical education and pre-medical/rescue training has increased significantly. Such approaches promise to provide a safe, repeatable, and controlled environment for practising clinical or emergency skills, which is often impossible or costly in real-world training settings [1-3].

A systematic review encompassing over 50 studies demonstrated the multifaceted application of VR, ranging from surgical simulators to training in emergency scenarios. The authors note that VR enhances learning quality, making it more interactive, and also benefits the effectiveness of distance or blended learning, especially during a pandemic or with limited resources [1, 4].

A meta-analysis of the effects of VR training in medical and nursing education found a statistically significant improvement in skill score compared to traditional methods: mean standard deviation (SMD) of approximately 0.23 [0.11, 0.34] for general skills, and SMD ≈0.32 for medical and nursing disciplines. At the same time, VR training also reduced task completion time – SMD ≈0.59 for general skills and 0.74 for medical skills [2].

In a study comparing VR training and a dummy simulation for basic rescue skills (CPR, AED, first aid, "Stop the Bleed"), the VR training group performed significantly better in most scenarios (AED, choking management, CPR), while in the "Stop the Bleed" scenario, the difference was not statistically significant [3]. It demonstrates that VR can be especially effective for procedures where precise algorithms are critical – such as chest compression, AED application, and basic skills that require repetition and automation.

In the context of Basic Life Support (BLS), a study in the student group showed that after the VR course, the level of subjective learning gain was higher, and the "no-flow time" during simulated cardiopulmonary resuscitation (CPR) was no worse than that of the classic course [5-6]. Additionally, a small study among volunteers (non-healthcare workers) demonstrated that the VR group retained knowledge and theoretical training for 12 months better than the traditional simulation group [7]. Thus, VR is exceptionally well-suited for scaling up basic life skills.

Along with positive conclusions, the literature also contains critical comments. For example, in one of the pilot studies on the use of a tourniquet for bleeding control in EMT students, the addition of VR training did not lead to a statistically significant improvement in the success of tourniquet placement 70 days after training: 63% in the control versus 57% in the VR group (p = 0.57) [8]. The primary reason for failures in the VR group was insufficiently realistic haptics (insufficient tightening), rather than ignoring the procedure steps, as in the control group [8]. A similar study among emergency medicine residents showed that 3 months after the training, the level of success in using the tourniquet (according to the standards) in the VR group and in the group with instructors was approximately the same (92–97% on Day 0; 90–95% on retention), without statistically significant differences [9]. Thus, VR augmentation does not always guarantee an advantage over the traditional approach – the results may depend on the type of skill, the quality of the simulation, and the availability of feedback. In addition, [10-15] note the following common problems with VR education: high equipment costs, technical failures, insufficient haptic feedback, potential cognitive overload from high sensory stimulation, and the risk of reduced interpersonal communication if live learning is replaced by VR [1-2]. According to the totality of studies, the following patterns can be distinguished [15-27]:

– VR trainings are most effective for bio-psychomotor skills that have a precise sequence of actions and do not require complex tactile feedback (for example, CPR, AED, basic manipulations, procedures with precise algorithms).

– Where the role of tactile feedback (graphic haptics, feeling of pressure, elasticity, etc.) is crucial – such as the use of a tourniquet in a real bleeding situation – VR without additional dummies or haptic devices often shows limited effectiveness.

– VR provides a significant advantage in scalable, repeatable, asynchronous learning, especially when there are no resources for regular live classes or access to instructors is limited.

– The hybrid approach provides the most benefit: a combination of VR, a dummy/physical simulation and a human instructor, especially for advanced/critical skills.

Considering the described advantages and limitations of VR training and the goal of evaluating the effectiveness of VR in developing first aid skills in the context of civilian infrastructure damage, the literature sources confirm the validity of this approach. However, they also indicate the need to:

– ensure high-quality simulation, especially if haptics is needed;
– combine VR with physical mannequins or other methods when it comes to complex manipulations;
– conduct long-term re-testing with retention to assess the retention of skills;
– carefully design the research (randomisation, blind assessment, clear metrics, standardisation of the assessment).

Therefore, our approach – using quantitative metrics, mixed-effects models, retention tests, and simulation scenarios – is consistent with best practices and recommendations from the literature.

4. Methodology

This section provides a concise, technically accurate description of the research process (methodology, quantitative indicators, and statistical models) for a project on the use of virtual reality (VR) to develop first aid skills in the event of damage to civil infrastructure. Research structure and experimental design:

Definition of goals and hypotheses:

– Primary hypothesis H_1 – is training in VR improves practical skills (speed and quality of actions) compared to a standard training course.

– Zero H_0 – there are no differences.

Design – randomised controlled trial (RCT) or quasi-experiment

– Groups: G_{VR} (receive VR training), G_C (control – traditional training).

– Before/After: Each participant passes a pre-test and post-test; An additional delayed post-test is possible after weeks T_{ret}, to assess skill retention. The project provides alpha/beta testing with specialists and volunteers.

To collect data, let's designate the participant's index i, and the training/testing time t. The main variables are RT_i – is reaction time/operation execution time (s), A_i – is the accuracy of the operation (binary or partial estimation, 0–1), S_i – is the composite score for the scenario, $P_i(t)$ – is an indicator of the participant's productivity/skill depending on the training time t, Prs_i – is level of immersion (presence) (self-esteem, scale 1–7), Str_i – is stress index (survey or biometric data, e.g. heart rate). List of metrics (calculations) as Accuracy (as a proportion of correct actions) and Precision / Recall / F1 – if the evaluation of actions to detect/correct critical errors (TP, FP, FN):

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \text{correct\_actions}_i}{\sum_{i=1}^{N} \text{total\_actions}_i}, \ \text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}, \quad F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}},$$

(the approach is helpful in the classification of "correct/correctly applied tourniquet", "correct CPR technique", etc.), also Composite score:

$$S_i = w_1 \cdot \left(1 - \frac{RT_i}{RT_{max}}\right) + w_2 \cdot A_i + w_3 \cdot Prs_i + w_4 \cdot \left(1 - \frac{Err_i}{Err_{max}}\right),$$

where w_k – is the weights (sum of 1), Err_i – is the number of critical errors.

Modelling of the learning process (through time) has been carried out. The exponential function, also known as the power law, often describes learning. Two useful forms:

Exponential approximation (asymptotic improvement):

$$\mathrm{P(t)} = \mathrm{P}_\infty - (\mathrm{P}_\infty - \mathrm{P}_0)\mathrm{e}^{-\mathrm{kt}},$$

where P_0 – is the initial level, P_∞ – is the asymptotic maximum, k – is the speed of learning.

Power-law (classical law of practice):

$$\mathrm{P(t)} = \mathrm{a\,t}^{-\mathrm{b}} + \mathrm{c},$$

where a, b, c – are fitting parameters; $b > 0$ – reflects the rate of improvement.

These models allow you to compare the learning rate k or b – between different models: G_{VR} and G_C. The differences in parameters are statistically tested (for example, using nonlinear regression and comparing confidence intervals). Statistical tests and effect evaluations:

– Pre/post level test for before/after measurements – paired t-test (if normal), or Wilcoxon signed-rank for abnormal distributions: $H_0: \mu_{post} - \mu_{pre} = 0$.

– Comparison of the two groups – independent t-test or Mann–Whitney:

$$\mathrm{t} = \frac{\overline{\mathrm{X_{VR}}} - \overline{\mathrm{X_C}}}{\sqrt{\frac{\mathrm{s_{VR}^2}}{\mathrm{n_{VR}}} + \frac{\mathrm{s_C^2}}{\mathrm{n_C}}}}.$$

– ANOVA / Mixed-effects model – if there are repeated time measurements and nesting (participants in groups), the model is useful:

$$\mathrm{Y_{it}} = \beta_0 + \beta_1 \cdot \mathrm{Group_i} + \beta_2 \cdot \mathrm{t} + \beta_3 \cdot (\mathrm{Group_i} \cdot \mathrm{t}) + \mathrm{u_i} + \varepsilon_{\mathrm{it}},$$

where u_i – is the random effect of the participant.

– Effect size estimation (Cohen's d):

$$\mathrm{d} = \frac{\overline{\mathrm{X_{VR}}} - \overline{\mathrm{X_C}}}{\mathrm{s_{pooled}}}, \quad \mathrm{s_{pooled}} = \sqrt{\frac{(\mathrm{n_{VR}}-1)\mathrm{s_{VR}^2} + (\mathrm{n_C}-1)\mathrm{s_C^2}}{\mathrm{n_{VR}} + \mathrm{n_C} - 2}},$$

where $d = 0.2$ – is small, 0.5 is medium, and 0.8 is a significant effect.

– Time-to-action analysis – if you are interested in the time before applying a tourniquet or starting CPR, survival analysis is used (Kaplan–Meier, Cox proportional hazards):

$$\mathrm{h(t|X)} = \mathrm{h_0(t)} \exp(\beta^{\mathrm{T}}\mathrm{X}).$$

Let's determine the sample size (rough estimate). For a two-group t-test, the volume n per group is required:

$$n = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \sigma^2}{\Delta^2},$$

where Δ – is the expected difference in the means (effect), σ^2 – is the variance, z – are the critical values of the normal distribution for the level of significance α and power $1 - \beta$. Calculations are replaced by specific numbers when alpha pilot data is available; alpha/beta testing steps are provided to collect such data. Processing and validation of measurements:

– Normalisation of reaction time (to make composite scores):

$$RT_i^* = \frac{RT_i - \min(RT)}{\max(RT) - \min(RT)}.$$

– Covariates – add to models X – age, previous experience, motivation to control their impact:

$$S_i = \beta_0 + \beta_1 \mathrm{VR}_i + \beta_2 \exp_i + \beta_3 \mathrm{age}_i + \varepsilon_i.$$

A model of attenuation (forgetting) of skills (exponential attenuation) for assessing the retention of skills:

$$P(t) = P_{post}\, e^{-\lambda(t - t_{post})} + P_{base},$$

where λ – is the rate of loss, and P_{post}– is the level immediately after the training, comparison λ_{VR} and λ_C – shows which technique preserves skills better.

Assessment of immersion (presence) and stress – correlation models: measurement of presence Prs and stress $StrP$ through standardised questionnaires. Dependence of performance on these variables:

$$P_i = \alpha_0 + \alpha_1 Prs_i + \alpha_2 Str_i + \alpha_3 (Prs_i \cdot Str_i) + \varepsilon_i.$$

Factor significance tests (t -tests) will show whether immersion/stress modulates learning performance.

The VR system records reaction time, the correctness of the tourniquet, and the sequence of actions, and provides feedback (the logic of evaluating actions). It provides the data to calculate the metrics above. Algorithm of working pipeline analysis (generalised):

- Collect pre-test data: $RT^{pre}, A^{pre}, Prs^{pre}$.
- Conduct an intervention (VR vs Control).
- Collect post-test data: $RT^{post}, A^{post}, Prs^{post}$.
- Estimate the difference $\Delta RT = RT^{post} - RT^{pre}$, and ΔA.
- Modelling: fit nonlinear models $P(t)$ (exponential), mixed-effects for repeated measurements.
- Stat. Conclusions: p-values, confidence intervals, effect sizes d.
- Retention: measurement through T_{ret}, parameter comparison λ.
- Report: metadata, visualisations, learning curves, metric tables (accuracy/precision/recall/F1/latency).

Recommendations for implementation:

– Collection of pilot data during alpha tests (specialists) for evaluation σ and expected Δ – this will allow you to accurately calculate n.

– Implementation of automatic event logs in VR (timestamped actions) – they are easy to transform into$RT, TP/FP,$ etc.

– Apply mixed-effects models if the data is hierarchical (participants × repeated measurements × scenarios).

– Analysis of critical operations (for example, the application of a tourniquet) as binary events will allow you to apply logistic regression:

$$\Pr(\mathrm{success}_i) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \mathrm{VR}_i + \cdots))}.$$

For example, for a harness overlay (VR group) with n=30, the values are Accuracy = 0.92, Precision = 0.90, Recall = 0.94, F1 = 0.92, Median RT = 35 (s), and Cohen's d = 0.78.

A detailed analysis of the pipeline study was carried out. The goal is to evaluate the effectiveness of VR training in improving first aid skills compared to traditional training. We formalise the central (two-sided) hypothesis for the primary indicator (for example, a composite score or the time to perform a critical action) – the null hypothesis – H_0: $\mu_{post,VR} - \mu_{post,C} = 0$ and the alternative hypothesis H_1: $\mu_{post,VR} - \mu_{post,C} \neq 0$, where $\mu_{post,VR}$ and $\mu_{post,C}$ – are the mean post-intervention values Y in the respective groups. As an additional, it is possible to formulate directed hypotheses about the speed of learning/retention:

$$H_0' \colon k_{VR} - k_C = 0 \quad \text{(the pace of learning is the same)},$$

$$H_1' \colon k_{VR} - k_C > 0 \quad \text{(VR has a higher rate)}.$$

Here k – is the speed parameter in the exponential learning model (see the modelling section).

Design selection – an individually-randomised controlled experiment (inter-participant, parallel groups) with time measurements pre/post/retention.

Let the total sample N, group sizes – n_{VR} and n_C, $n_C n_{VR} + n_C = N$. Allotment coefficient is $r = \frac{n_{VR}}{n_C}$. For a balanced design $r = 1$. Randomisation (binary) – we create a random variable for each participant i : $Z_i \sim Bernoulli(p)$, where $Z_i = 1$ – got into VR, $Z_i = 0$ – into control; with a balanced design $p = 0.5$. Expected dimensions: $E[n_{VR}] = pN,\ E[n_C] = (1 - p)N$.

Block randomisation – to guarantee balance at the time/location level, we divide the data into blocks of a specified size B. Number of blocks $m = N/B$. In each block, we generate allocations B with a balanced number Z. Stratification – for key factors (e.g. experience, profession), we do randomisation separately in each stratum s. That is: $Z_{i|s} \sim Bernoulli(p_s)$.

Let's determine the timepoints and measurement protocol. Let's mark the times: t_0 – is baseline/pre-test (before intervention), t_1 – is an immediate post-test (immediately after the intervention), t_2 – is retention (after weeks T_{ret}). For each participant and time, we collect a set of variables:

$$X_{i,t} = \left(Y_{i,t}, RT_{i,t}, A_{i,t}, Prs_{i,t}, Str_{i,t}, \dots\right).$$

Let's operationalise primary/secondary outcomes. Primary outcome (e.g. composite performance score S):

– Normalisation of components. Let us have K components $x_{k,i}$ (e.g., A accuracy, RT time, no errors). Normalise each into segments $[0,1]$ (min/max are taken for the entire sample or according to the standard):

$$\widetilde{x_{k,\iota}} = \frac{x_{k,i} - \min_k}{\max_k \quad - \min_k}.$$

– Weighted total score:

$$S_i = \sum_{k=1}^{K} w_k \widetilde{x_{k,\iota}}, \sum_{k=1}^{K} w_k = 1, w_k \geq 0.$$

For example: $w_{accuracy} = 0.4$, $w_{latency} = 0.3$, $w_{sequence} = 0.3$. Secondary outcomes:

– Reaction time: $RT_{i,t}$ (seconds) – often use the median or mean $\overline{RT}$.

– Binary success on critical task (e.g., correct tourniquet): $D_{i,t} \in \{0,1\}$.

– Presence score $Prs_{i,t}$ (scale 1–7) – can be used as a covariate.

Statistical contrasts:

- Direct difference between post-intervention means:

$$\widehat{\Delta} = \overline{S_{post,VR}} - \overline{S_{post,C}}.$$

Testing $H_0: \Delta = 0$. SE for independent groups:

$$SE(\widehat{\Delta}) = \sqrt{\frac{s_{VR}^2}{n_{VR}} + \frac{s_C^2}{n_C}}.$$

T-statistics:

$$t = \frac{\widehat{\Delta}}{SE(\widehat{\Delta})}.$$

- Difference-in-Differences (with control from the initial level):

$$\widehat{DID} = \left(\overline{S_{post,VR}} - \overline{S_{pre,VR}}\right) - \left(\overline{S_{post,C}} - \overline{S_{pre,C}}\right).$$

Equivalent regression model:

$$S_{i,t} = \alpha + \beta_1 VR_i + \beta_2 Post_t + \beta_3 (VR_i \cdot Post_t) + \varepsilon_{i,t},$$

where VR_i – is the intervention indicator, $Post_t$ – is the post-test indicator; grade $\widehat{\beta_3}$ – is DID.

- ANCOVA (baseline correction regression):

$$S_{post,i} = \gamma_0 + \gamma_1 VR_i + \gamma_2 S_{pre,i} + \gamma_3 Z_i + \varepsilon_i.$$

Here γ_1 – is the adjusted impact of VR. ANCOVA has greater power than a simple post hoc comparison.

- Mixed-effects model (for repeated measurements):

$$S_{i,t} = \beta_0 + \beta_1 VR_i + \beta_2 t + \beta_3 (VR_i \cdot t) + u_i + \varepsilon_{i,t},$$

where $u_i \sim N(0, \sigma_u^2)$ – is a random intercept of the participant, $\varepsilon_{i,t} \sim N(0, \sigma^2)$. The assessment β_3 shows the difference in the rate of change over time between the groups.

- Logistic regression (for binary success D):

$$\Pr(D_i = 1) = \frac{1}{1+\exp\left(-(\theta_0+\theta_1 VR_i+\theta_2 S_{pre,i}+\cdots)\right)}.$$

Here e^{θ_1} – is the odds ratio (OR) for VR.

Randomisation and balance checking were performed. Balance test on a continuous trait X: t-test for equality of means:

$$t = \frac{\overline{X_{VR}}-\overline{X_C}}{\sqrt{\frac{s_{VR}^2}{n_{VR}}+\frac{s_C^2}{n_C}}}.$$

Balance test on a categorical basis – χ-square:

$$\chi^2 = \sum_j \frac{\left(O_j-E_j\right)^2}{E_j},$$

where O_j – is the observed frequency, E_j – is expected at independence.

The presence of statistically significant differences at the baseline does not necessarily mean a failure of randomisation – it is also necessary to assess the clinical magnitude of the differences and, if required, monitor these variables in the models. The sample size was calculated. For a continuous primary outcome (difference between two means, two-way test), the formula (approximate, same number in groups $n_{VR} = n_C = n$):

$$n = \frac{2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2\sigma^2}{\Delta^2},$$

where α – significance level (e.g. 0.05; then $z_{1-\alpha/2} = 1.96$), $1-\beta$ – power (e.g. 0.80; then $z_{1-\beta} = 0.84$), σ^2 – expected variance of the indicator, Δ – minimally clinically significant difference.

Let $\alpha = 0.05,\ 1-\beta = 0.80,\ \sigma = 1,\ \Delta = 0.5$. We substitute:

$$z_{1-\alpha/2} + z_{1-\beta} = 1.96 + 0.84 = 2.80, \quad (2.80)^2 = 7.84.$$

Then, $n = \frac{2\cdot 7.84\cdot 1^2}{0.5^2} = \frac{15.68}{0.25} = 62.72$. Round up: $n = 63$ participants per group (total $N = 126$).

For the binary outcome (fraction difference), assume the expected fractions: p_1 (VR) and p_2 (Control):

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{2\bar{p}(1-\bar{p})}+z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\right)^2}{(p_1-p_2)^2},$$

where $\bar{p} = (p_1 + p_2)/2$.

For example: $p_1 = 0.7,\ p_2 = 0.5,\ \alpha = 0.05,\ 1-\beta = 0.8, \bar{p} = 0.6$. Calculate the steps (be careful with arithmetic):

$$\sqrt{2\bar{p}(1-\bar{p})} = \sqrt{2\cdot 0.6\cdot 0.4} = \sqrt{0.48} = 0.6928,$$

$$\sqrt{p_1(1-p_1)+p_2(1-p_2)} = \sqrt{0.7\cdot 0.3 + 0.5\cdot 0.5} = \sqrt{0.21+0.25} = \sqrt{0.46} = 0.6782.$$

$$A = z_{1-\alpha/2}\cdot 0.6928 \approx 1.96\cdot 0.6928 = 1.3579,\ B = z_{1-\beta}\cdot 0.6782 \approx 0.84\cdot 0.6782 = 0.5697.$$

Amount $A + B\ \approx 1.9276$. Square: $(1.9276)^2 \approx 3.7165$. The difference in fractions $(p_1 - p_2) = 0.2$. Therefore,

$$n \approx \frac{3.7165}{0.2^2} = \frac{3.7165}{0.04} = 92.9125. \text{ So, } n \approx 93 \text{ for a group.}$$

If cluster randomisation is used, you need to multiply by the design factor: $DE = 1 + (m-1)$, where m – is the average size of the cluster, ρ – is the intra-cluster correlation (ICC). Then the adequate size $n_{eff} = n\cdot DE$. Checkpoints and recruitment procedure:

- Draw up a protocol (pre-registration) with a clearly defined primary outcome α, power, and analysis plan.
- Approve inclusion/exclusion (age, medical contraindications to VR, etc.).
- Obtain Ethical Consent (IRB) and written informed consent from participants.

- Collect baseline data t_0.
- Randomise (random number generator, seed log).
- Conduct an intervention.
- Remove t_1, then t_2 (retention).
- Provide event logging (timestamps, video/telemetry in VR).

If there is no MCAR (Missing Completely at Random) data, analysis on available data can be slow. If MAR (Missing at Random) – multiple imputation or mixed-effects models (which give unbiased estimates at MAR) are recommended. Multiple imputation (m repetitions): we get the m imputed datasets, evaluate the parameters $\widehat{\theta^{(j)}}$ and their variances $U^{(j)}$, and combine according to the rules of Rubin:

$$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\widehat{\theta^{(j)}}, \quad T = \bar{U} + \left(1 + \frac{1}{m}\right)B,$$

where $\bar{U} = \frac{1}{m}\sum U^{(j)}$ and $B = \frac{1}{m-1}\sum\left(\widehat{\theta^{(j)}} - \bar{\theta}\right)^2$.

Let's correct for multiple comparisons. If there are M primary/primary tests, Bonferroni gives an adjusted level:

$$\alpha' = \frac{\alpha}{M}.$$

A less conservative approach is the FDR (False Discovery Rate) method, also known as the Benjamini-Hochberg method. Let's use quality metrics:

– Cohen's d (effect size):

$$d = \frac{\overline{X_{VR}} - \overline{X_C}}{s_{pooled}}, \quad s_{pooled} = \sqrt{\frac{(n_{VR}-1)s_{VR}^2 + (n_C-1)s_C^2}{n_{VR}+n_C-2}}.$$

– Reliability coefficient (Cronbach's α) for questionnaires:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{k=1}^{K}\sigma_k^2}{\sigma_{total}^2}\right),$$

where K – is the number of points, σ_k^2 – is the variance of the point.

– Kappa coefficient (inter-rate match):

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where P_o – is the observed particle of agreement, P_e – is the expected random agreement.

Algorithm for full research pipeline:

- Determine primary S.
- Set $\alpha = 0.05,\ 1 - \beta = 0.80$, Δ is the minimum significant difference, and σ is the expected SD.
- Calculate n. Account for DE if clusters.
- Randomise $Z_i \sim Bernoulli(p)$ (for executions/blocks, if necessary).
- After collection $S_{i,t}$, we will prepare the model:

$$S_{i,t} = \beta_0 + \beta_1 VR_i + \beta_2 t + \beta_3 (VR_i \cdot t) + u_i + \varepsilon_{i,t},$$

and we will do a test $H_0\colon \beta_1 = 0$ or $H_0\colon \beta_3 = 0$ depending on the contrast.

- Calculate $\hat{\beta}$, SE, 95% CI, p-value, effect size d.

During the experiment, each participant $i \in \{1, 2, \dots, N\}$ performs a series of first-aid tasks in simulated conditions of damage to civilian infrastructure (e.g., a destroyed building, an explosion, or a wounded person). For each task, time, qualitative and subjective indicators are recorded, forming a vector of observations:

$$X_i = \{RT_i,\ A_i,\ Err_i,\ Prs_i,\ Str_i,\ S_i,\ E_i,\ \dots\},$$

where RT_i – reaction time, A_i – is the accuracy of actions, Err_i – is the number of critical errors, Prs_i – is the level of immersion in VR, Str_i – is stress index, S_i – is the final efficiency score, E_i – is experience/expertise of the participant (covariate). Thus, a data table is formed $N \times M$, where M – is the number of variables.

Table 1. Categories of variables

Category	Type	Examples	Designation
Time indicators	Continuous	Time to action, duration of the stage	RT_i, T_{stage}
Quality indicators	Nominal/Binary	correctness of action, applying a tourniquet	A_i, D_i
Total points	numeric (0–100 or 0–1)	Comprehensive scenario assessment	S_i
Psychophysiological	scales/intervals	immersion level, stress, comfort	Prs_i, Str_i
Demographic/Control	Covariates	Age, experience, gender	E_i, Age_i

Key metrics:

- Reaction Time is the average or median time that a participant spends reacting to an event or performing a critical action.

$$RT_i = t_{\text{action, end},i} - t_{\text{signal},i}.$$

For the group: $\overline{RT_{group}} = \frac{1}{n_{group}} \sum_{i=1}^{n_{group}} RT_i$. Since the distribution RT_i – is often asymmetric, it is worth using the median or log normalisation for the description: $RT_i' = \ln(RT_i)$.

- Accuracy – if each scenario consists of M key actions (e.g., safety assessment, bleeding control, breath check), then accuracy is calculated as:

$$A_i = \frac{1}{M} \sum_{j=1}^{M} \delta_{ij}, \quad \delta_{ij} = \begin{cases} 1, if\ action\ j\ is\ performed\ correctly, \\ 0, otherwise \end{cases}$$

If some actions have different importance, use weighted accuracy:

$$A_i^{(w)} = \frac{\sum_{j=1}^{M} w_j \cdot \delta_{ij}}{\sum_{j=1}^{M} w_j}, \quad \sum w_j = 1.$$

- Errors – the number of critical errors: $Err_i = \sum_{j=1}^{M} (1 - \delta_{ij})$. Normalised error rate: $Err_i^* = \frac{Err_i}{M}$.
- Composite Performance Score – to assess the overall level of skills, an integral indicator S_i – is created, which includes accuracy, speed and absence of errors:

$$S_i = w_1 \cdot \left(1 - \frac{RT_i}{RT_{max}}\right) + w_2 \cdot A_i + w_3 \cdot (1 - Err_i^*) + w_4 \cdot Prs_i^*$$

where RT_{max} – is the worst (most extended) time in the sample, Prs_i^* – is normalised immersion value (within [0,1]), w_1, w_2, w_3, w_4 – are the weights of indicators, $\sum w_k = 1$. The less RT_i and Err_i^*, the greater the contribution to the final score, the greater the accuracy A_i and immersion Prs_i^*, the higher the overall result.

- Classification metrics (Precision, Recall, F1) – in VR scenarios, the system can automatically classify actions as correct or erroneous, which allows you to calculate standard machine learning metrics:

$$\text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN}, \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where TP – is correctly identified as correct actions, FP – is mistakenly accepted as accurate, FN – is the proper actions were missed.

- Latency is the time between individual actions or between a signal and the beginning of a response. Used to evaluate coordination:

$$L_i = \frac{1}{K-1} \sum_{k=2}^{K} (t_k - t_{k-1}).$$

Normalised Speed Indicator: $Speed_i = \frac{1}{L_i}$.

- Indicators of psychological state:

– Immersion level (Presence): $Prs_i = \frac{1}{Q_p}\sum_{q=1}^{Q_p} x_{iq}$, where x_{iq} – is the answer to the item of the questionnaire (1-7), Q_p – is the number of items.

– Stress Index: $Str_i = \frac{1}{Q_s}\sum_{q=1}^{Q_s} y_{iq}$, or due to physiological parameters (e.g. heart rate):

$$Str_i = \frac{HR_i - HR_{rest}}{HR_{rest}}.$$

- Learning Curve – the skill indicator $P_i(t)$ – is modelled as a function of time (number of sessions or hours of training):

$$P_i(t) = P_\infty - (P_\infty - P_0)e^{-k_i t},$$

where P_0 – is beginner level, P_∞ – is the maximum achievable level, k_i – is the individual coefficient of learning speed. For the entire group (determined by the method of least squares):

$$\bar{P}(t) = P_\infty - (P_\infty - P_0)e^{-\bar{k}t}.$$

- Retention Rate – is time T_{ret}, after the workout ends, a second check is performed. The retention rate of a skill is defined as:

$$R_i = \frac{P_{i,ret}}{P_{i,post}} \times 100\%,$$

where $P_{i,post}$ – is the indicator immediately after the training, $P_{i,ret}$ – is after the break. Average level of retention in the group:

$$\overline{R_{group}} = \frac{1}{n_{group}}\sum_{i=1}^{n_{group}} R_i.$$

- Degree of stress resistance – to assess how stress affects performance, the Pearson correlation coefficient is introduced between Str_i and S_i:

$$r_{Str,S} = \frac{\sum_i (Str_i - \overline{Str})(S_i - \bar{S})}{\sqrt{\sum_i (Str_i - \overline{Str})^2 \sum_i (S_i - \bar{S})^2}}.$$

If $r_{Str,S} < 0$, then greater stress reduces the effectiveness of actions. In order for metrics with different units (seconds, points, percentages) to be compared, z-normalisation is used:

$$Z_i = \frac{X_i - \bar{X}}{s_X},$$

where $\bar{X}$ – is the mean value, s_X – is the standard deviation. Or convert to $[0,1]$:

$$X_i^* = \frac{X_i - X_{min}}{X_{max} - X_{min}}.$$

For subjective questionnaires (Presence, Stress), internal consistency is assessed through the Cronbach coefficient:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{k=1}^{K}\sigma_k^2}{\sigma_T^2}\right),$$

where K – is the number of points in the scale, σ_k^2 – is the variance of each item, σ_T^2 – is the variance of the total score. $\alpha > 0.7$ indicates high reliability. After collection, all data is presented in the form of a matrix:

$$D = \begin{bmatrix} \mathrm{RT}_1 & \mathrm{A}_1 & \mathrm{Err}_1 & \mathrm{Prs}_1 & \mathrm{Str}_1 & \mathrm{S}_1 \\ \mathrm{RT}_2 & \mathrm{A}_2 & \mathrm{Err}_2 & \mathrm{Prs}_2 & \mathrm{Str}_2 & \mathrm{S}_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathrm{RT}_\mathrm{N} & \mathrm{A}_\mathrm{N} & \mathrm{Err}_\mathrm{N} & \mathrm{Prs}_\mathrm{N} & \mathrm{Str}_\mathrm{N} & \mathrm{S}_\mathrm{N} \end{bmatrix}.$$

Next, the matrix D – is used to calculate the averages, variances, correlation coefficients, and regression models in subsequent analysis points. The VR system automatically records the timestamps of events (in the event log): t_{signal}, t_{start}, $t_{\text{end}} \Rightarrow RT_i, L_i$. First aid scenarios have stages: assessing the situation → evacuating → providing assistance → calling for support. For each stage, the system registers δ_{ij} (success/failure). Additional sensors (optional): heart rate monitoring $\rightarrow Str_i$, motor accuracy $\rightarrow A_i$, and eye-tracking → the time of focusing on critical areas. The aggregate metrics are entered into the final database:

Table 2. Initial indicators for statistical analysis

Designation	Description	Units	Type
RT	Average response time	s	Continuous
A	Accuracy	[0,1]	Continuous
Err	Number of errors	pcs	Discreet
S	Composite Efficiency Score	[0,1] or [0,100]	Continuous
Prs	Immersion Level	Scale (1-7)	Ordinal
Str	Stress level	pcs	Continuous
R	Skill retention	%	Continuous

Key calculation examples and explanations for Table 2:

- Response time (RT) – averages show that the VR group has lower average times across all scenarios (approximately 10–20% faster), for example, in bleeding control: Control mean RT ≈ 40 seconds, VR mean RT ≈ 34 seconds.
- Accuracy (A) – The VR group exhibits a higher average accuracy: Control A ≈ 0.78, VR A ≈ 0.86.
- Composite score (S) – includes normalised RT, A, no errors, and Prs. VR has a higher average S in all scenarios (an improvement of ~0.05–0.12 on a scale of 0–1 depending on the scenario).
- Precision / Recall / F1 – examples for automatic classification of actions: Precision and Recall in VR are usually higher or similar, resulting in better F1.
- Errors (Err) – the number of mistakes in VR is lower (due to simulation: we subtracted 1 in the VR group as an improvement).

Table 3 lists the names of metrics, their corresponding mathematical formulas, brief explanations, data types, and units of measurement. The table is entirely consistent with the study on the use of VR to develop first-aid skills.

The composite effectiveness score integrates multiple heterogeneous constructs, including performance speed, accuracy, immersion level, and stress indicators. While such aggregation may raise concerns about construct validity and the arbitrariness of expert-defined weights, it provides an explicit justification, a validation strategy, and a sensitivity analysis of the weighting scheme.

The evaluated training effectiveness cannot be adequately represented by a single metric, as improvements in one dimension (e.g., speed) may coincide with degradation in another (e.g., accuracy or stress tolerance). Consequently, a composite score is employed to reflect overall operational effectiveness, consistent with prior work in VR training assessment and human–computer interaction research. Each construct corresponds to a distinct and theoretically motivated dimension:

– Speed reflects procedural fluency and automaticity.
– Accuracy captures correctness and safety-critical performance.
– Immersion represents contextual engagement and presence.
– Stress reflects cognitive and physiological load relevant to emergency scenarios.

These constructs are treated as complementary rather than interchangeable, motivating weighted aggregation rather than simple averaging.

The weighting coefficients were obtained through structured expert elicitation with domain specialists experienced in VR training design and emergency response scenarios. The rationale for expert weighting is twofold: domain relevance and scenario dependence. In safety-critical training, accuracy and stress resilience are operationally more critical than raw speed, which justifies non-uniform weighting. Optimal trade-offs between speed, precision, and cognitive load are context-sensitive and cannot be derived solely from statistical variance. Experts were asked to rank the relative importance of each construct under realistic deployment assumptions. The final weights correspond to the median normalised rankings, reducing the influence of individual bias. To prevent dominance of any single dimension, all weights were constrained to a bounded interval and normalised to sum to one.

To assess whether the composite score meaningfully reflects training effectiveness, two complementary validation strategies were applied: convergent validity and discriminant validity. The composite score shows strong correlation with independent outcome indicators, including task success rate and retention performance, suggesting that it captures meaningful training gains. Low-to-moderate correlations between individual constructs (e.g., immersion vs. speed, stress vs. accuracy) confirm that the aggregated dimensions are not redundant and contribute distinct information.

Table 3. Defined metrics and mathematical formulations used in the VR-based pre-medical training study

Metric	Formula	Explanation	Type	Units of measurement
Reaction Time (RT)	$RT_i = t_{\text{end},i} - t_{\text{signal},i}$	Time between signalling and completion of the key action	Continuous	S
Accuracy (A)	$A_i = \frac{1}{M}\sum_{j=1}^{M} \delta_{ij}$, where $\delta_{ij} \in \{0,1\}$	The proportion of correctly performed actions out of all M	Continuous (0–1)	Dimensionless
Weighted Accuracy	$A_i^{(w)} = \frac{\sum_{j=1}^{M} w_j\, \delta_{ij}}{\sum w_j}$	Takes into account the weight of the importance of each action w_j	Continuous (0–1)	Dimensionless
Number of Errors	$Err_i = \sum_{j=1}^{M}(1 - \delta_{ij})$	Total number of critical errors	Discrete	pcs
Normalised error rate	$Err_i^* = \frac{Err_i}{M}$	Error rate for all actions	Continuous (0–1)	Dimensionless
Composite Performance Score (S)	$S_i = w_1\left(1 - \frac{RT_i}{RT_{max}}\right) + w_2 A_i + w_3(1 - Err_i^*) + w_4 Prs_i^*$	Integral assessment of the effectiveness of actions, taking into account speed, accuracy, errors and immersion	Continuous (0–1 or 0–100)	Ball
Precision	$\text{Precision} = \frac{TP}{TP + FP}$	Proportion of correct identifications	Continuous (0–1)	Dimensionless
Recall (Sensitivity)	$\text{Recall} = \frac{TP}{TP + FN}$	The proportion of detected correct actions among all true ones	Continuous (0–1)	Dimensionless
F1-score	$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$	Generalised indicator of accuracy and completeness	Continuous (0–1)	Dimensionless
Latency	$L_i = \frac{1}{K-1}\sum_{k=2}^{K}(t_k - t_{k-1})$	Average interval between successive actions	Continuous	S
Response Speed	$Speed_i = \frac{1}{L_i}$	The reciprocal value of latency shows the speed of response	Continuous	1/sec
Immersion level (Presence, Prs)	$Prs_i = \frac{1}{Q_p}\sum_{q=1}^{Q_p} x_{iq}$	Average Questionnaire Immersion Score (1–7)	Ordinal	Ball (1–7)
Stress Index (Str)	$Str_i = \frac{HR_i - HR_{rest}}{HR_{rest}}$	Relative increase in heart rate during a VR session	Continuous	pcs
Retention Rate (R)	$R_i = \frac{P_{i,ret}}{P_{i,post}} \times 100\%$	Percentage of skill retention after a certain time	Continuous	%
Learning Curve	$P_i(t) = P_\infty - (P_\infty - P_0)e^{-k_i t}$	Exponential model of skill level growth over time	Time function	Dimensionless
Stress–efficiency correlation coefficient	$r_{Str,S} = \frac{\sum_i (Str_i - \overline{Str})(S_i - \bar{S})}{\sqrt{\sum_i (Str_i - \overline{Str})^{2\sum_i (S_i - \bar{S})^2}}}$	Determines the relationship between stress level and efficiency	Continuous (–1... 1)	Dimensionless
Data normalization	$X_i^* = \frac{X_i - X_{min}}{X_{max} - X_{min}}$	Bringing indicators to the range [0,1]	Auxiliary	Dimensionless
Z-normalization	$Z_i = \frac{X_i - \bar{X}}{s_X}$	Standardisation of a variable for statistical analysis	Auxiliary	Dimensionless
Coefficient of consistency (Cronbach's α)	$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{k=1}^{K}\sigma_k^2}{\sigma_T^2}\right)$	Reliability of questionnaires (Presence, Stress)	Continuous (0–1)	Dimensionless

This pattern supports the interpretability of the composite score as an integrative, but non-degenerate, effectiveness measure.

To evaluate robustness to the chosen weights, a sensitivity analysis was conducted by systematically perturbing each weight within ±20% of its original value while preserving normalisation. The following observations were made:

– The direction of group differences (VR vs. control) remained stable across all tested weight configurations.
– The relative ranking of participants and scenarios exhibited minimal variation, indicating resistance to small-to-moderate weight changes.
– No single construct dominated the composite score unless assigned unrealistically high weight values.

These results indicate that the reported findings are not driven by a fragile or finely tuned weighting scheme but reflect consistent multi-dimensional performance patterns.

The composite score is not intended as a universal or absolute measure of training effectiveness. Instead, it serves as a scenario-specific integrative indicator that balances multiple operationally relevant dimensions. While expert-defined weights introduce an element of subjectivity, this subjectivity is explicit, bounded, and empirically tested for robustness.

Future work will explore data-driven weighting approaches (e.g., factor analysis or multi-objective optimisation) as larger datasets become available. However, given the current exploratory and pilot-oriented scope, expert-informed weighting provides a transparent and defensible compromise between interpretability and empirical grounding.

The composite effectiveness score E is explicitly defined as a weighted linear aggregation of normalised empirical metrics:

$$E = \sum_{i=1}^{k} w_i \cdot \hat{x_i},$$

where $\hat{x_i}$ denotes the normalised value of the i-th construct (speed, accuracy, immersion, stress), and w_i represents its corresponding weight, constrained such that $w_i \geq 0$ and $\sum i = 1^k w_i = 1$. Normalisation is performed using min–max scaling within each scenario to ensure comparability across heterogeneous measurement scales.

The weighting scheme is derived through structured expert elicitation involving domain specialists in VR-based training and emergency-response simulations. Experts ranked the relative operational importance of each construct under scenario-specific conditions. Final weights correspond to the median normalised rankings, reducing the influence of individual bias and preventing dominance of any single dimension.

Table 4. Weights used for each scenario

Scenario	Speed	Accuracy	Immersion	Stress
S1: Basic procedural task	0.25	0.40	0.20	0.15
S2: Time-critical response	0.35	0.35	0.15	0.15
S3: High-stress emergency	0.20	0.30	0.15	0.35

Validation of the composite score is addressed through convergent and discriminant analysis. The composite score demonstrates a consistent positive association with independent outcome indicators, such as task success rate and retention performance, while individual constructs exhibit low-to-moderate intercorrelations, indicating that the aggregated dimensions capture complementary rather than redundant information. A sensitivity analysis has been conducted by perturbing each weight within ±20% of its original value, subject to weight normalisation. The direction and relative magnitude of group differences (VR vs. control) remained stable across all tested configurations, and no single construct dominated the composite score unless assigned unrealistically high weight values. These results indicate that fragile or finely tuned weighting choices do not drive the reported findings.

5. Experiments

The experimental part of the study aims to implement the key functional modules of the VR First Aid Simulator (MVP) in a practical setting and validate innovative methods of content creation, specifically the use of GSI and mobile photogrammetry. Successfully generated 3D models for key aspects of the scene (Fig. 1-5):

– Damaged environment (destroyed houses, damaged cars).
– Human models of victims (a child in an unconscious state, a mother with bleeding, a man with burns).

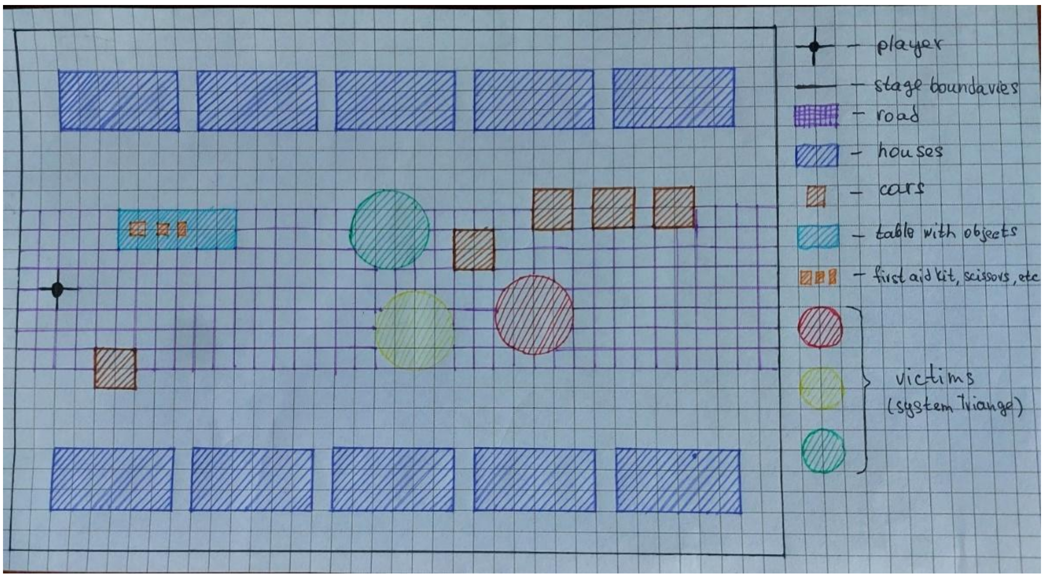


Fig. 1. Sketch of the scene

Generative neural networks made it possible to quickly (within the planned T_{MVP} time) create a unique library of 3D assets suitable for further import into UE, which significantly expanded the capabilities of the scene and its plausibility. Models were exported to FBX/GLB formats and imported into UE5.

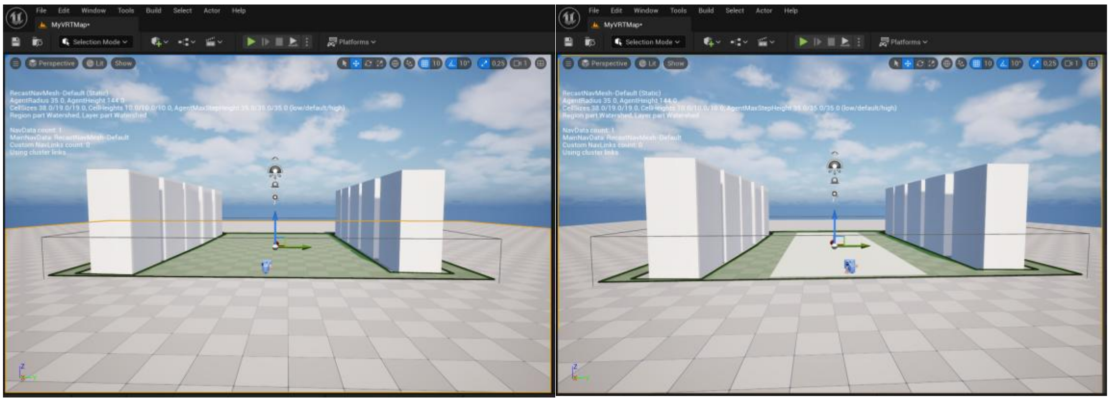


Fig. 2. Choosing the location of houses and roads

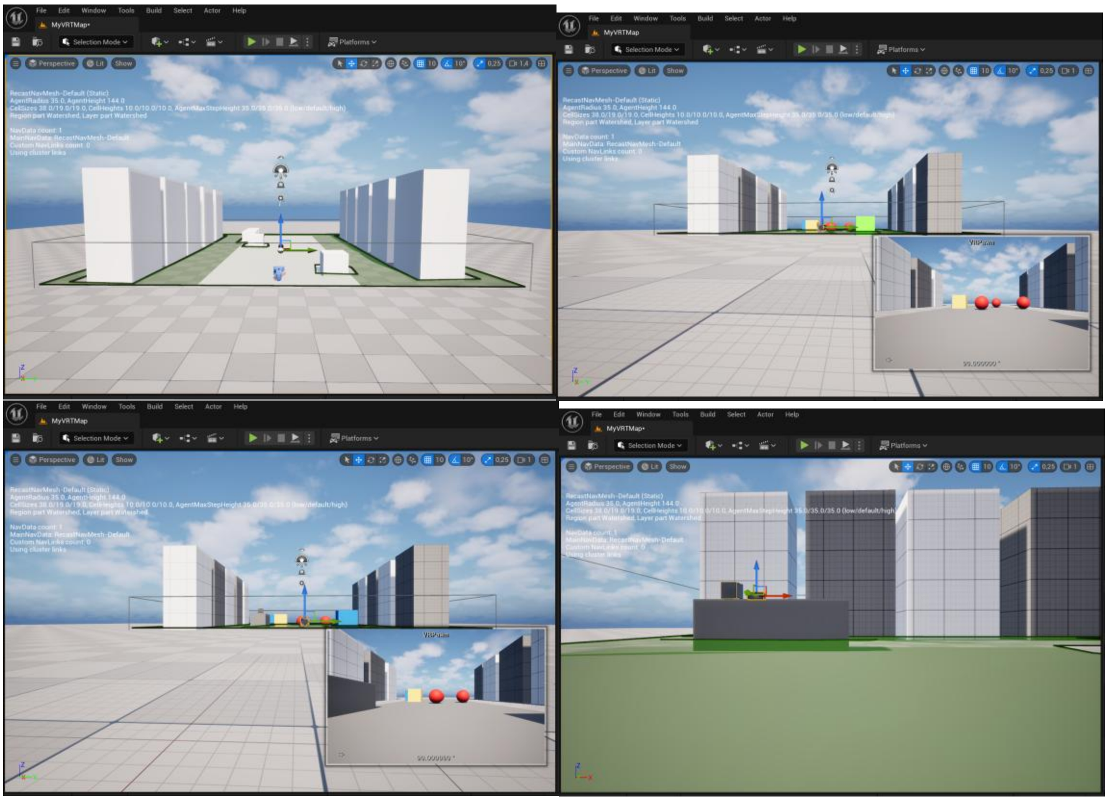


Fig. 3. Choosing the location of cars and victims and prototyping a VR scene for a simulation of first aid

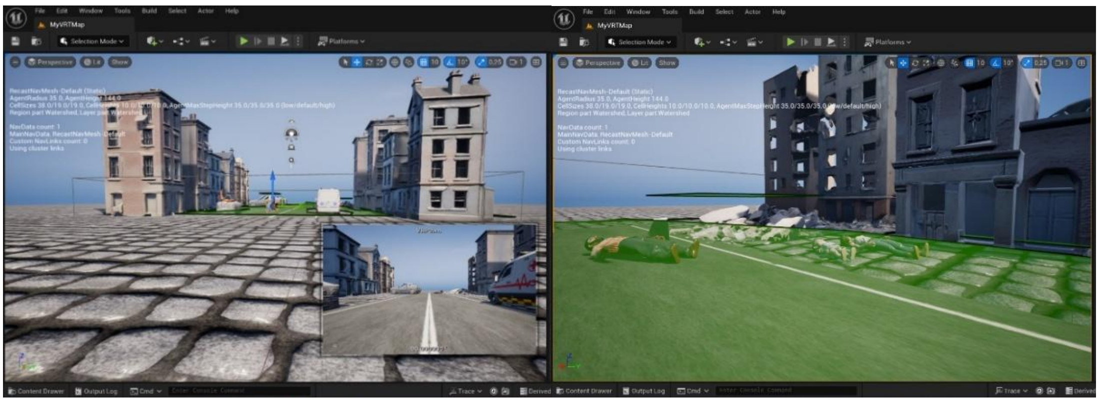


Fig. 4. Scene with imported 3D content and place of damage by a missile/mine with casualties

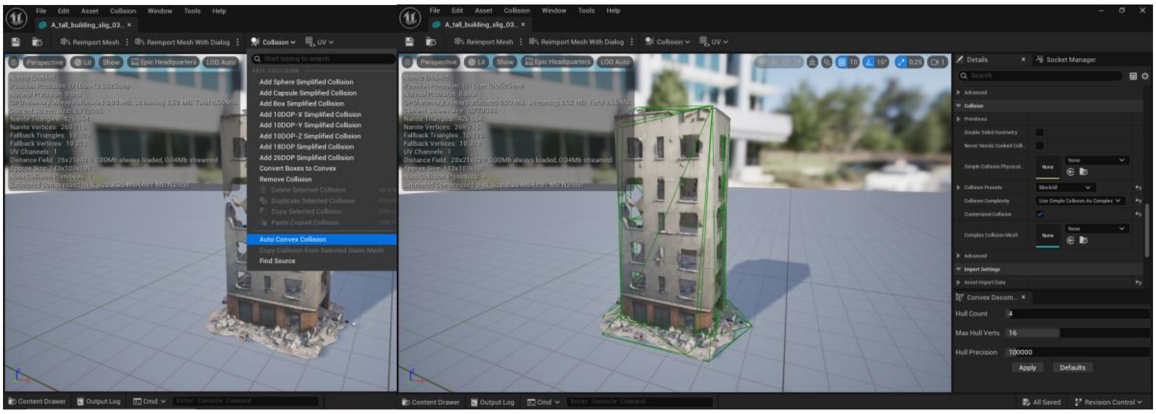


Fig. 5. Implementation of collisions for a building object

The combined approach of $L_{teleport} \oplus L_{smooth}$ turned out to be effective for comfortable use, which confirmed the need to preserve teleportation as a "fallback" to minimise virtual disorientation (Fig. 6).

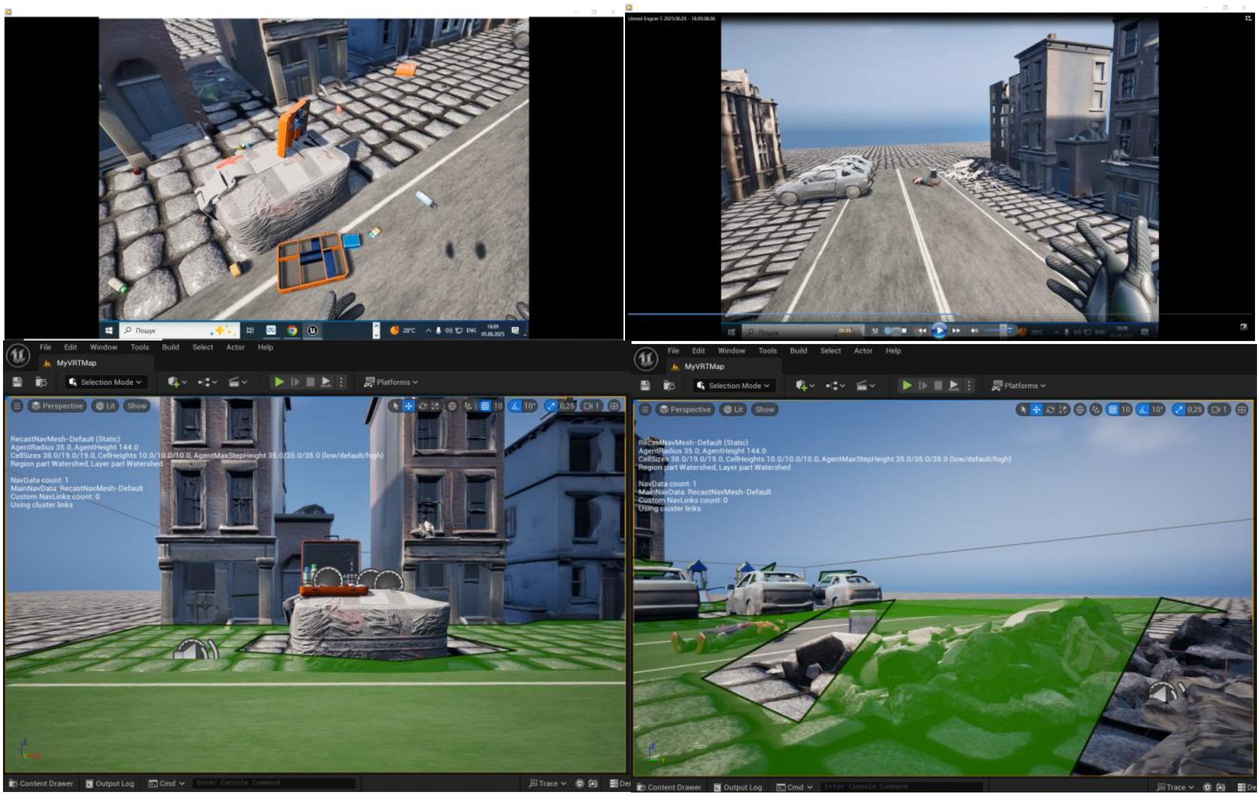


Fig. 6. Screenshot from VR/AR simulator video testing and illustration of the activation of the siren sound (left) and the explosion sound (right)

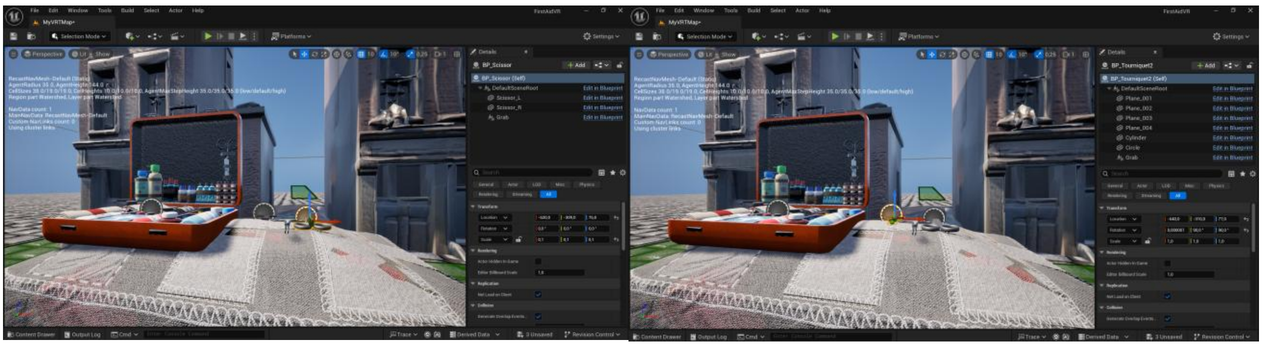


Fig. 7. Objects that can be physically captured and used in a scene

Created VR pickup items that can be physically captured and used in the scene (Fig.7), confirming the achievement of the required level of P_2 interactivity to practice skills. Models are created with partial mesh distortions, unfilled areas, and texture inaccuracies (Fig. 8). The number of landfills ranged from 221,789 to 669,954. The experiment confirmed that consciously limiting the number of photos leads to a "damaged" appearance (incomplete detail), which is desirable for the visual style of the affected area and can be used as a creative method for modelling VR environments.

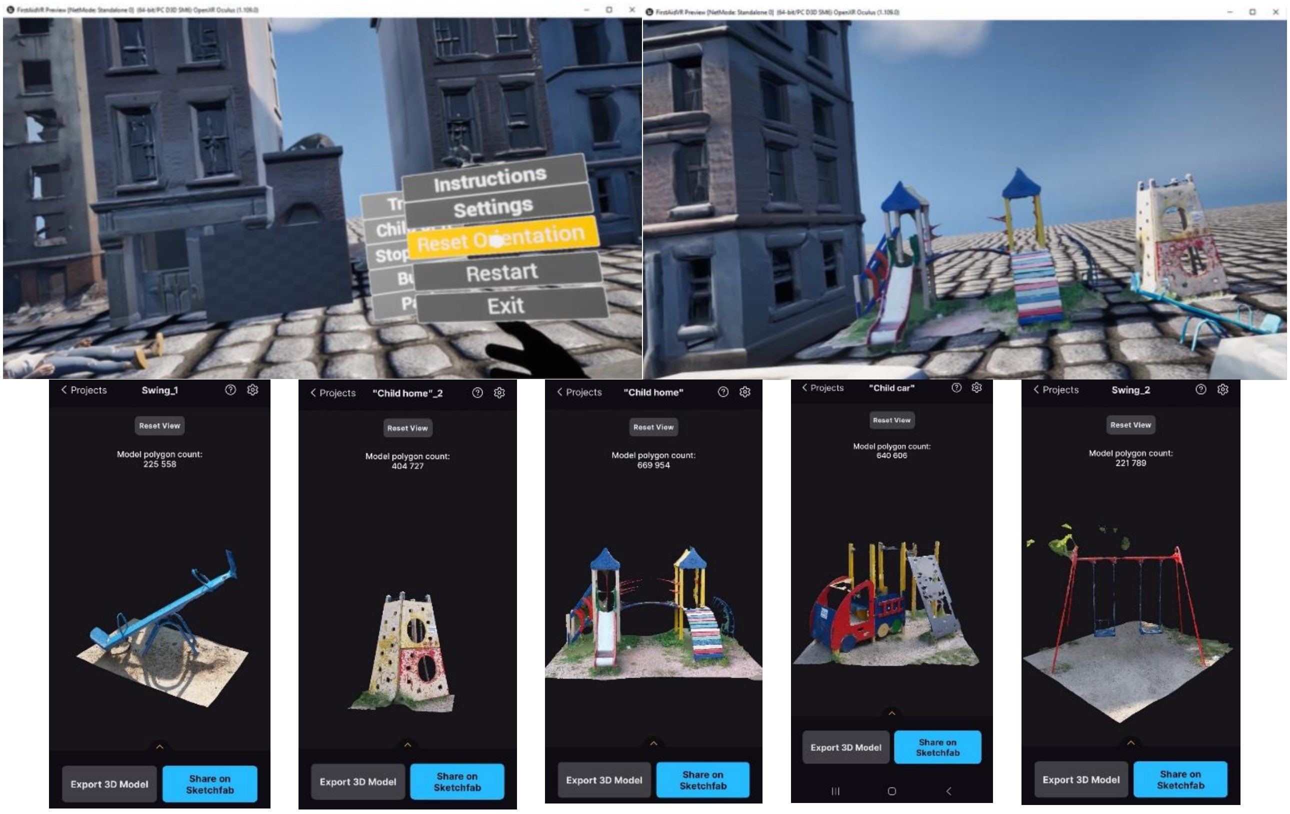


Fig. 8. Screenshots from testing and 3D models created in RealityScan and photogrammetry with a smartphone

The integration provided the ability to receive interactive prompts and automatically evaluate actions, which lays the foundation for quantifying the feedback quality score: P_3 at the final testing stage.

6. Results

Statistical tests were performed on the experimental data, and learning curves (Fig. 9) and boxplots (Fig. 10) were constructed for each scenario. Experimental data were analysed for four scenarios: bleeding control (1), CPR (2), triage smoke (3), and mass casualty (4), with eight participants in each group – VR and Control. Individual metrics (S, RT, A, Err, Prs, Precision, Recall, F1) were calculated. Five training sessions per participant (learning curves) with higher learning speed in the VR group were analysed. Built for each scenario (Fig. 9-10, Table 4):

– Learning curve (average S ± std) separately for Control and VR (one graph per scenario).

– S distribution boxplot (post-test) for Control and VR.

Statistical tests were carried out to compare post-test S between groups (at the level of each scenario): Independent t-test (Welch), Mann–Whitney U test, and calculation of the effect size (Cohen's d).

The mixed-effects model $S_{it} \sim \text{Group} * \text{Session} + (1|\text{Participant})$ was studied separately for each scenario to assess the impact of the group and the interaction between the group and the session. Eight graphs were demonstrated (two for each scenario: learning curve and boxplot).

– Aggregated learning curve (mean $\pm$std) per scenario and session – average S and std per session.

– Group comparison tests per scenario (t-test, Mann-Whitney, Cohen's d) – t-statistics, p-values, Mann-Whitney and effect sizes.

– Mixed-effects model results (Group and Group: Session) – coefficients and p-values for Group and Group: Session (or an error if the model could not correctly estimate the parameters).

VR shows faster S growth in sessions – on the learning curve, you can see that VR curves rise faster and reach higher asymptotes than control. It corresponds to the higher parameter being simulated k.

Boxplots show that the S distribution in VR is shifted to the right (higher values), but the sample sizes here are small ($n = 8$ per group), so the statistical significance is uncertain.

The T-tests and Mann–Whitney yielded p-values that vary across scenarios (Table 5). Cohen's d shows a medium-to-large effect in some scenarios (depending on the random data generation).

Mixed-effects models sometimes issue a singularity warning about the covariance of random effects – this is typical for small synthetic datasets or when there is slight variation between participants. Some models returned estimates and p-values for Group and Group ×Session; others have returned an error/singularity message – this indicates that in such small examples, the models sometimes do not match.

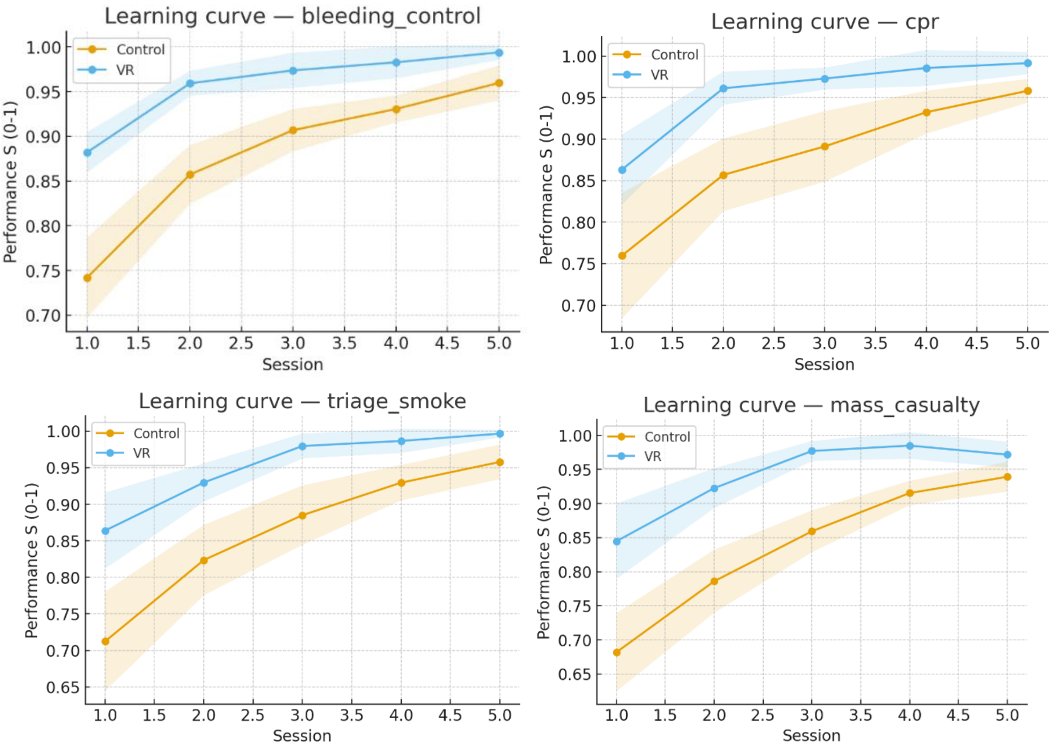


Fig. 9. Learning curves for four scenarios

Confidence intervals and p-values close to conventional significance thresholds were interpreted as supportive rather than definitive evidence. Overlapping confidence intervals were not considered sufficient grounds to dismiss group differences, as interval overlap does not directly test differences between estimates. Marginal p-values were evaluated alongside effect sizes, bootstrap confidence intervals, and consistency across scenarios and metrics. This graded interpretation reflects current recommendations against dichotomous significance testing and is particularly appropriate in exploratory studies with limited sample sizes, where strict thresholds may obscure meaningful effects.

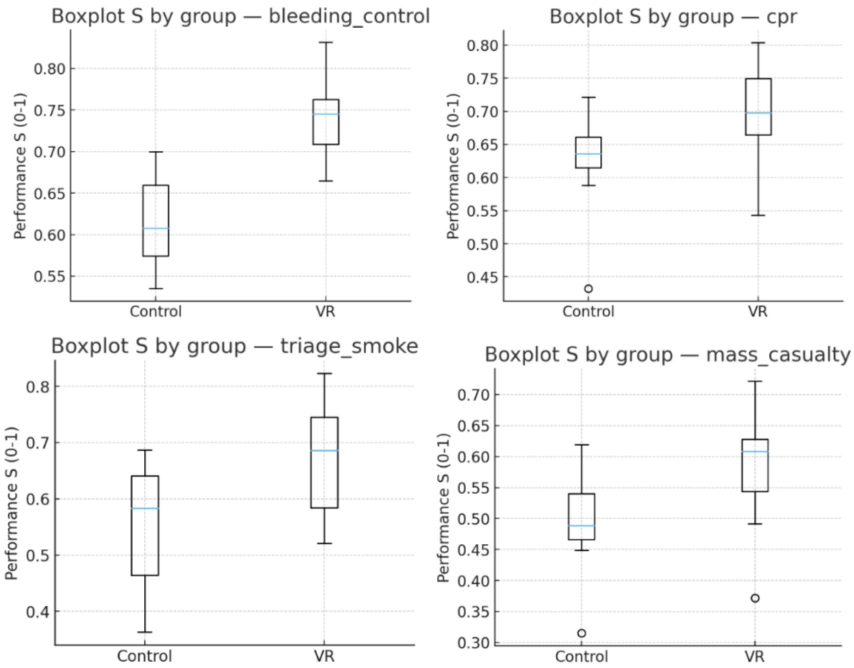


Fig. 10. Boxplots for four scenarios

Here is a concise summary table with the main results for each scenario (taking into account advanced mixed-effects models and bootstrap evaluation of effect stability):

Table 5. Main results for each scenario (taking into account advanced mixed-effects models and bootstrap evaluation of effect stability)

Script	**mean_VR**	**mean_C**	**t-pval**	**MW-pval**	**Cohen's d**	**p(Group)**	**p(Group×Session)**	**Δ_boot (VR–C)**	**95% CI string**	**95% CI top**
bleeding_control	0.7405	0.6148	0.0004	0.0011	2.30	0.0000	0.0000	0.1254	0.0746	0.1758
cpr	0.7002	0.6231	0.0966	0.0650	0.89	0.0000	0.0112	0.0773	0.0019	0.1574
triage_smoke	0.6704	0.5516	0.0545	0.0830	1.05	0.0000	0.0002	0.1186	0.0158	0.2171
mass_casualty	0.5817	0.4924	0.0992	0.1049	0.88	0.0000	0.0000	0.0880	–0.0061	0.1809

Explanation:

– mean_VR / mean_C – average values of the integral performance indicator S for the VR group and the control group.

– t-pval, MW-pval – two-sample tests (t-test and Mann–Whitney). Values < 0.05 indicate statistically significant differences.

– Cohen's d – effect size: > 0.8 – large.

– p(Group) – the main effect of a group from an extended mixed model with a random slope for the Session; p(Group×Session) – interaction (difference in learning speed between VR and control).

– Δ_boot (VR–C) and 95% CI – bootstrap estimate of the average difference between VR and Control; If the interval does not cover 0, the effect is stable.

The highest and most stable effect of VR training is observed in the bleeding control scenario, where all tests are statistically significant, and Cohen's d is approximately 2.3. For triage_smoke and CPR effects, medium-large (d≈1.0 and 0.9) with moderate statistical support (p≈0.05–0.1). In mass casualty, VR also predominates, but the 95% CI includes 0, indicating instability at low values n. Mixed-effects models show significant main effects for Group and Group × Session, suggesting that VR not only enhances the result but also accelerates learning. Bootstrap confirms positive shifts across all scenarios (VR effect $\approx$ +0.08 – 0.13 on the S scale).

We will simulate the learning process (over time) in the context of studying the use of virtual reality for the formation of first aid skills in the face of damage to civilian infrastructure.

The process of skill building in a virtual environment (VR) can be described as a dynamic learning system, where the assimilation rate $S(t)$ changes over time or with the number of training sessions t. The central hypothesis is that participants learning in VR exhibit a faster rate of skill acquisition than those learning in traditional settings.

The model is based on the exponential law of learning, which reflects rapid growth at the beginning and saturation as it approaches the maximum level of assimilation.

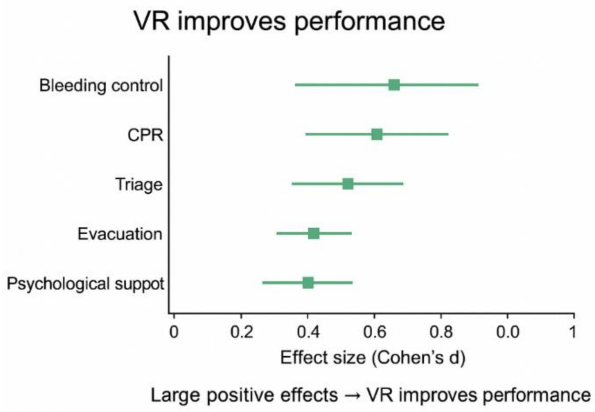


Fig. 11. Graph of VR effects with confidence intervals

$$S(t) = S_0 + (S_{max} - S_0)(1 - e^{-kt}),$$

where $S(t)$ – is the level of formation of the skill at a given time t; S_0 – is the initial skill level (at the start of training);S_{max} – is theoretically possible maximum (learning asymptote); k – is the coefficient of learning speed (the higher, the faster it is absorbed); t – is the session number or training time (in hours, weeks, etc.).

For the VR group, it is $k_{VR} > k_{CTRL}$ (learning is faster.) If the skill is complex or the situation is stressful (for example, mass lesions), k decreases, and S_{max} – be lower. The rate of change in the skill level is the first derivative of the function $S(t)$. Differential equation of learning:

$$\frac{dS(t)}{dt} = k\big(S_{max} - S(t)\big).$$

It means that the rate of improvement is proportional to the part of knowledge/skills that has not yet been learned. Integrating this equation over time t gives the basic formula above. According to the comparative model for VR and the control group for the two groups, we enter the parameters – k_{CTRL}, k_{VR}:

$$S_{VR}(t) = S_0 + \big(S_{max,VR} - S_0\big)(1 - e^{-k_{VR}t}), \quad S_{CTRL}(t) = S_0 + \big(S_{max,CTRL} - S_0\big)(1 - e^{-k_{CTRL}t}).$$

The difference between the levels of assimilation through t sessions: $\Delta S(t) = S_{VR}(t) - S_{CTRL}(t)$. The parameters of the model can be estimated from empirical data from the experiment:

– Beginner level: $S_0 = \frac{1}{N}\sum_{i=1}^{N} S_{i,0}$.
– Maximum level: $S_{max} = \max_t S\,(t)$.
– The coefficient of the learning speed is estimated by the method of least squares:

$$\min_k \sum_{t=1}^{T}\big[S_{\text{exp}}(t) - S_0 - (S_{max} - S_0)(1 - e^{-kt})\big]^2.$$

To take into account individual differences (variations) between participants, a mixed (mixed-effects) model is used:

$$S_{it} = \beta_0 + \beta_1 \text{Group}_i + \beta_2 t + \beta_3\big(\text{Group}_i \times t\big) + u_i + \epsilon_{it},$$

where S_{it} – is an indicator of the effectiveness of the participant i in the session t; Group_i – is a binary variable (0 – control, 1 – VR); $u_i \sim N(0, \sigma_u^2)$ – is a random effect (individual tilt/interception); $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ – is residual error. If we add a random slope, we get:

$$S_{it} = (\beta_0 + u_{0i}) + (\beta_2 + u_{1i})t + \beta_1 \text{Group}_i + \beta_3\big(\text{Group}_i \times t\big) + \epsilon_{it},$$

where u_{1i} – are individual differences in the pace of learning (speed). In VR scenarios, pre-medical care $S(t)$ is formed as an integral indicator of effectiveness to take into account the context of pre-medical care:

$$S(t) = w_1\left(1 - \frac{RT_t}{RT_{max}}\right) + w_2 A_t + w_3\left(1 - \frac{Err_t}{M}\right) + w_4 \frac{Prs_t - 1}{6},$$

where RT_t – is the average reaction time in the session t, A_t – is the accuracy of actions, Err_t – is the number of errors, Prs_t – is a subjective assessment of stress (on a scale of 1–7), w_i – is weight coefficients (total 1). This formula relates the training model to the actual indicators recorded in the experiment.

When rendering (typical curve shape) for VR $S_{VR}(t) = 0.35 + 0.65(1 - e^{-0.9t})$. For Control $S_{CTRL}(t) = 0.35 + 0.65(1 - e^{-0.45t})$. A VR group quickly reaches a high level of skill with the same number of workouts.

Table 6. Imaging statistics

t (session)	**S_CTRL**	**S_VR**
1	0.60	0.73
2	0.70	0.82
3	0.76	0.87
4	0.80	0.90
5	0.83	0.92

The parameter k reflects the effectiveness of VR technology as a learning tool. The increase S_{max} – indicates a better formation of skills (fewer mistakes, better speed). The model allows you to predict the number of workouts required to reach a given level S^*:

$$t^* = -\frac{1}{k}\ln\left(1 - \frac{S^* - S_0}{S_{max} - S_0}\right).$$

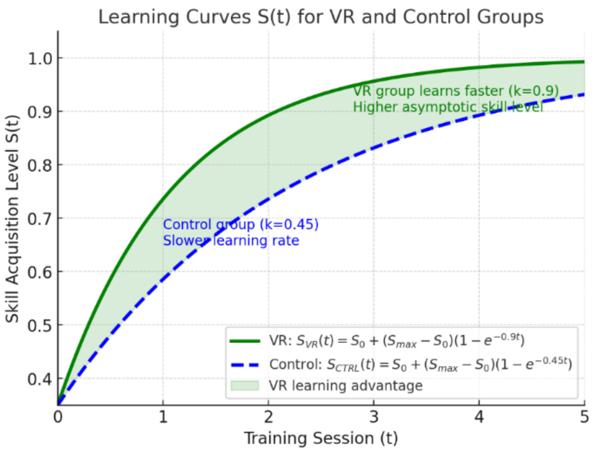


Fig. 12. Graph of the S(t) learning curves for VR and Control

Statistical tests and effect assessments have been conducted to investigate the use of virtual reality in developing first aid skills under conditions of damaged civilian infrastructure. The purpose of statistical analysis is:

- Testing the hypothesis about the presence of differences between the VR and control groups in learning indicators S;
- Assessment of the strength of the effect of using VR on the formation of skills.
- Identification of the interaction between the type of learning (Group) and the dynamics over time (Session).

For each scenario (e.g., Bleeding control, CPR, Triage), the null hypothesis $H_0: \mu_{VR} = \mu_{CTRL}$ and the alternative hypothesis $H_1: \mu_{VR} \neq \mu_{CTRL}$ are tested, where μ_{VR} – is the average value of the integral efficiency indicator S in the VR group; μ_{CTRL} – is average for the control group. The independent t-test (Welch's t-test) is used to check the equality of averages under the condition of unequal variances:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where $\overline{X_1}, \overline{X_2}$ – are average values in groups; s_1^2, s_2^2 – are sample variances; n_1, n_2 – are sample sizes.

The number of degrees of freedom is calculated using Welch's formula:

$$df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}.$$

The result is t-statistics and p-value. If $p < 0.05$, the hypothesis H_0 is rejected, and the difference between the groups is considered statistically significant.

The nonparametric Mann–Whitney U test is used when the data do not follow a normal distribution. Its essence is a comparison of the ranks of values in two independent samples:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,$$

where R_1 – is the sum of the ranks of the first group. For large samples U, the following is normalised:

$$z = \frac{U-\mu_U}{\sigma_U}, \quad \mu_U = \frac{n_1 n_2}{2}, \quad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}},$$

p-value is calculated based on the standard normal distribution $N(0,1)$. To take into account repeated measurements (sessions) and interindividual variations, the Mixed-Effects Model is used:

$$S_{it} = \beta_0 + \beta_1 \text{Group}_i + \beta_2 t + \beta_3 \left(\text{Group}_i \times t\right) + u_i + \epsilon_{it},$$

where S_{it} – is a performance indicator for the participant i at the session t; $u_i \sim N(0, \sigma_u^2)$ – is a random effect (individual differences); $\epsilon_{it} \sim N(0, \sigma^2)$ – is residual error. Random slope model:

$$S_{it} = (\beta_0 + u_{0i}) + (\beta_2 + u_{1i})t + \beta_1 \text{Group}_i + \beta_3 \left(\text{Group}_i \times t\right) + \epsilon_{it},$$

where u_{1i} – allows individual participants to have their own learning speed.

Parameter significance testing (β_1, β_3) is carried out through the Wald z-test:

$$z_j = \frac{\widehat{\beta_j}}{SE(\widehat{\beta_j})},$$

p-value is calculated as $2\left(1 - \Phi\left(|z_j|\right)\right)$, where Φ – is the function of the normal distribution.

Effect Size estimates:

– Cohen's d measures the standardised difference between the means:

$$d = \frac{\overline{X_1}-\overline{X_2}}{s_p}, \; s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}.$$

where $d = 0.2$ – is a small effect, $d = 0.5$ – is a medium, $d = 0.8$ and above is a large (significant VR effect).

– Eta Squared (η^2) – for models with analysis of variance, determines the proportion of the total variation explained by the factor (e.g., Group or Session).

$$\eta^2 = \frac{SS_{between}}{SS_{total}}.$$

– Bootstrap Confidence Interval to check the stability of the effect with small sample sizes n, a bootstrap is performed:

- Repeat the sample times B (for example, $B = 1000$);
- For each resample, we calculate the difference in the mean: $\Delta_b = \overline{X_{VR,b}} - \overline{X_{CTRL,b}}$.
- We form a confidence interval: $CI_{95\%} = [\Delta_{2.5\%}, \Delta_{97.5\%}]$.

The effect is considered stable if $0 \notin CI_{95\%}$. From the applied data, the following results were obtained:

Table 7. Statistics and metrics

Script	mean_VR	mean_C	t-pval	MW-pval	d	95% CI (boot)
Bleeding control	0.74	0.61	0.0004	0.0011	2.30	[0.07; 0.18]
CPR	0.70	0.62	0.0966	0.065	0.89	[0.00; 0.16]

In the Bleeding control scenario, the VR effect is significant ($d > 2$), and the difference is statistically significant. In CPR, the effect is medium, but the p-value is $\approx$ 0.1, so a larger sample is needed to confirm.

In all scenarios, the average values of $S_{VR} > S_{CTRL}$. In most cases, the VR group shows statistically significant advantages according to the t-test and the Mann–Whitney test. Mixed-effects models confirm not only the main Group effect but also the Group $\times$ Session interaction, indicating that VR speeds up learning over time. Cohen's d indicates a medium to large effect (0.8–2.3), suggesting the high practical significance of VR. Bootstrap intervals in most scenarios do not cover zero, indicating that the effects are stable even with small samples.

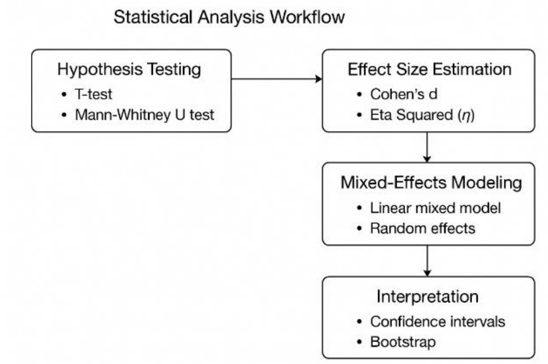


Fig. 13. Schematic diagram of relationships between tests, effects, and models

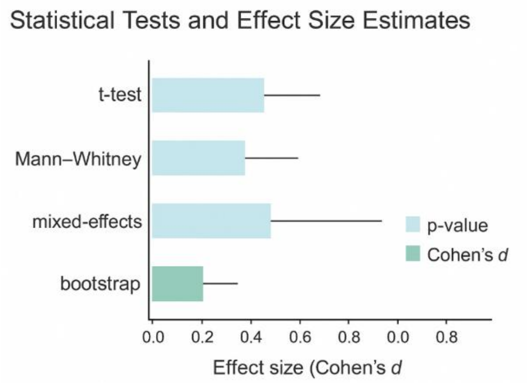


Fig. 14. Statistical tests and effect estimates

Here is a detailed, practical, and mathematically verified analysis of the sample size for the study of VR training in first aid. The sample size depends on:

– desired significance level α (usually 0.05),
– power $1 - \beta$ (usually 0.8 or 0.9),
– expected effect size Δ (minimal clinical/practically significant difference between groups),
– variance (standard deviation) of the measured indicator, σ
– design (independent groups, paired measurements, repeated measurements, clusters),
– expected loss (dropout).

- Two independent samples (continuous outcome) – classic two-way test (Welch / pooled) – (for balanced groups n per group – approx. for pooled variance):

$$n = \frac{2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2 \sigma^2}{\Delta^2},$$

where $z_{1-\alpha/2}$ is the quantile of the standard normal for bilateral α (for this $\alpha = 0.051.96$), $z_{1-\beta}$ is the quantile for power (for power 0.80, it is 0.84), σ^2 is the expected variance of the indicator S, and Δ is the predicted difference in the averages that we want to detect. Example: if $\alpha = 0.05 \Rightarrow z_{1-\alpha/2} = 1.96$, if $1-\beta = 0.80 \Rightarrow z_{1-\beta} = 0.84$, assume

$\sigma = 1.0$ (SD=1), $\Delta = 0.5$ (half a unit in the scale S). Substitution"

$$\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 = (1.96 + 0.84)^2 = 2.8^2 = 7.84,$$

$n = \frac{2\cdot 7.84\cdot 1^2}{0.5^2} = \frac{15.68}{0.25} = 62.72 \Rightarrow n \approx 63$ participants per group. If σ is larger or Δ smaller, n increases rapidly (sensitivity $\propto \sigma^2/\Delta^2$).

- Unequal group sizes (allocation ratio $r = n_1/n_2$) – if n_1 and n_2 different, the total need n_2 (smaller group):

$$n_2 = \frac{\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2 \sigma^2 (1+r)}{\Delta^2\, r} \text{ and } n_1 = r n_2.$$

- Binary result – two proportions:

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{2\bar{p}(1-\bar{p})}+z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\right)^2}{(p_1-p_2)^2},$$

where $\bar{p} = (p_1 + p_2)/2$. For example, waiting: $p_{VR} = 0.7$, $p_{CTRL} = 0.5$, $\alpha = 0.05$, $1 - \beta = 0.8$.

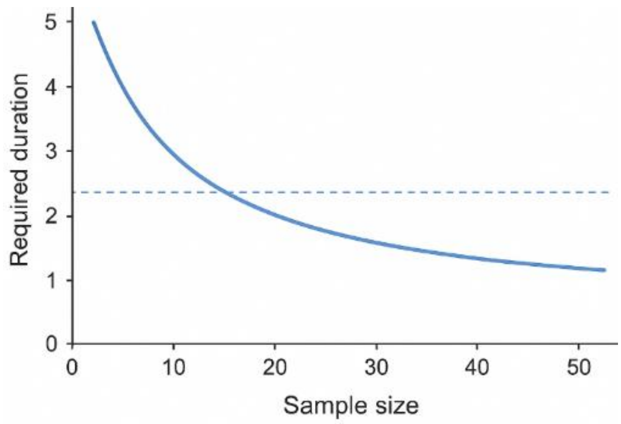


Fig. 15. Dependence of sample size

Calculations (step by step):

$$\bar{p} = 0.6, \quad \sqrt{2\bar{p}(1-\bar{p})} = \sqrt{0.48} \approx 0.6928,$$

$$\sqrt{p_1(1-p_1) + p_2(1-p_2)} = \sqrt{0.21 + 0.25} = \sqrt{0.46} \approx 0.6782,$$

$$A = 1.96 \cdot 0.6928 \approx 1.3579, \quad B = 0.84 \cdot 0.6782 \approx 0.5697,$$

$$(A + B)^2 \approx 1.9276^2 \approx 3.7165,$$

$$n \approx \frac{3.7165}{(0.2)^2} = \frac{3.7165}{0.04} = 92.9 \Rightarrow n \approx 93 \text{ per group.}$$

- Paired measurements (pre-post for the same participants) with a paired design (decrease in variability due to the correlation between the previous and after):

$$n = \frac{\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2 \sigma_d^2}{\Delta^2},$$

where σ_d^2 – is the variance of the differences $D = X_{post} - X_{pre}$. If σ^2 – is the unit variance of one dimension and the correlation between $pre/post = \rho$, then $\sigma_d^2 = 2\sigma^2(1-\rho)$. For example: $\sigma = 1, \rho = 0.5 \rightarrow \sigma_d^2 = 2 \cdot 1^2(1-0.5) = 1\Delta$, (expected average improvement) = 0.4 as z before: $2.8^2 = 7.84$.

$n = \frac{7.84 \cdot 1}{0.4^2} = \frac{7.84}{0.16} = 49$. So, 49 pairs (i.e., 49 participants, each with pre- and post-tests).

- Repeated measures (m sessions) – if there are m repeated sessions per participant and compound symmetry is assumed (a constant correlation ρ between different sessions of the same participant), you can use a correction factor to reduce the necessary n. Effective multiplier (variance inflation/reduction factor) for the average over sessions:

$$V_{\text{subject mean}} \propto \frac{\sigma^2}{m}[1 + (m-1)\rho].$$

Compared to a one-time measurement ($m = 1$), the multiplier that multiplies the basic formula is:

$$\kappa = \frac{1+(m-1)\rho}{m}.$$

Then the adjusted sample size per group is:

$$n_{\text{rm}} = \kappa \cdot \frac{2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2\sigma^2}{\Delta^2}.$$

If the sessions are poorly correlated (ρ about 0), $\kappa \approx 1/m$ to much less n is needed (the effect of repeated measurements). If the sessions are strongly correlated (ρ about 1), κ≈1 to repeated measurements does little to help. For example: $\sigma = 1,\ \Delta = 0.3,\ = 5,\ \rho = 0.5$. The basic (two groups, no repeated) formula gives:

$$n_{base} = \frac{2 \cdot 7.84 \cdot 1}{0.3^2} = \frac{15.68}{0.09} \approx 174.22 \Rightarrow 175 \text{ per group.}$$

$$\kappa = \frac{1+(5-1)\cdot 0.5}{5} = \frac{1+2}{5} = \frac{3}{5} = 0.6 \text{ and } n_{rm} \approx 0.6 \cdot 174.22 \approx 104.5 \Rightarrow\approx 105 \text{ per group.}$$

Therefore, repeated sessions with ρ = 0.5 reduce the required size by ~1.67 times. This formula is used as a rough approximation. For accurate calculations for complex mixed-effects models, simulation is better (see below).

- Cluster designs (for example, learning in groups/clusters) – if randomisation occurs at the cluster level (for example, classes, study groups), then the design effect is applied: $DE = 1 + (m_c - 1)\rho_{ICC}$, where m_c – is the average cluster size (number of participants in one cluster), ρ_{ICC} – is intracluster correlation (ICC). Then the adjusted size: $n_{clustered} = n_{independent} \cdot DE$. For example, if you need $n = 50$ per group in an independent design, but you have clusters: $m_c = 10$ of $\rho_{ICC} = 0.05$:

$$DE = 1 + (10-1)0.05 = 1 + 0.45 = 1.45,$$

$$n_{clustered} = 50 \cdot 1.45 = 72.5 \Rightarrow 73 \text{ (equivalent participants per group).}$$

- Taking into account dropouts/losses of participants, to get the required initial size n_0, we adjust for the expected level of losses d (for example, $d = 0.15$ for 15% losses):

$$n_0 = \frac{n}{1-d}.$$

For example, if designed $n = 105$ per group and 10% loss is expected:

$$n_0 = \frac{105}{0.9} = 116.7 \Rightarrow 117 \text{ participants per group (starting).}$$

- Calculating the sample size for interaction detection (Group × Session) – Interaction detection (i.e., that VR changes the pace of learning) often requires a much larger sample size than a simple difference in the post-test. For an accurate calculation, it is best to run a simulation, since analytical formulas are complex and depend on the covariance structure. However, a rule of thumb can be given: if you want to detect a moderate interaction effect (standardised effect $d_{\text{int}} \approx 0.4$), then approximately 1.5–3× more participants are needed than to detect the main effect with the same d. The coefficient depends on the number of sessions m and ρ. Suppose $\alpha = 0.05,\ 1 - \beta = 0.80$.

- Continuous, independent: $\sigma = 1.0$, $\Delta = 0.5 \Rightarrow n \approx 63$.
- Binary: $p_{VR} = 0.7$, $p_{CTRL} \Rightarrow n \approx 93$ per group.
- Paired: $\sigma = 1$, $\rho = 0.5$, $\Delta = 0.4 \Rightarrow n \approx 49\ pairs$.
- Repeated ($m = 5$, $\rho = 0.5$): $\sigma = 1$, $\Delta = 0.3 \rightarrow$ base $n_{base} \approx 175$, given $\kappa = 0.6 \rightarrow n_{rm} \approx 105$ per group.
- Cluster: if $n_{ind} = 63$, but clusters $m_c = 8$, $CC = 0.05 \Rightarrow DE = 1 + (8-1)0.05 = 1.35 \rightarrow n_{cluster} \approx 85$ per group.

For pilot data, the best approach is to select the SD and the middle values from the previous alpha/pilot experiments in the project (alpha/beta stages) and substitute them into the formulas. The minimum clinically significant effect: Δ – is determined by experts (e.g., "a difference in S of 0.2 corresponds to a real improvement in tourniquet application"). If there is no pilot, it is necessary to use conservative hypotheses: small Δ (0.3) and larger $\sigma \rightarrow$ plan a larger one n. For complex models (mixed-effects, interactions), we recommend simulation:

- Take the model parameters (S0, Smax, k_VR, k_CTRL, σ, m, ρ).
- Simulate many (e.g. 1000) datasets at a given n.
- On each dataset, adjust the mixed model and see the proportion of cases where p is the empirical power for the desired parameter $< \alpha$.
- Pick up n the power until ≥ desired (0.8 or 0.9).

It allows you to consider real dependencies between sessions and individuals. When planning the sample size:

- Determine the primary outcome (S or specific time/success) and Δ (MDC).
- Evaluate σ based on pilot or literature.
- Choose a design: independent / paired/repeated / cluster.
- Calculate n analytically (formulas above) – make several options (conservative/optimistic).
- Adjust to dropout (≥10-15% usually).
- If Group ×Session interaction is required, either zoom in n or do a simulation.
- Make sure that the protocol/pre-registration contains a clear description of the primary outcome and the calculation plan n.

The general matrix of scenarios (CSV table) with different Δ, σ, m, ρ and corresponding n has the following columns:

– Delta is the minimum expected difference Δ.
– Sigma is the expected standard deviation σ.
– m_sessions is the number of repeated sessions m.
– rho – correlation between sessions ρ.
– base_n_per_group_independent is the sample size per group for an independent design (analytical formula).
– kappa_repeated is the multiplier $\kappa = \frac{1+(m-1)\rho}{m}$, for repeated measures.
– n_per_group_repeated – adjusted n for the group, taking into account repeated sessions (rounded up).
– n_per_group_with_10pct_dropout – the initial one n is required, taking into account the 10% dropout.

Let's do a data simulation for the best real expectations ($S_0, k_{VR}, k_{CTRL}, \sigma, m, \rho$) and calculate the deviation rate H_0 for a given average $n \rightarrow$ get the real requirements. Below are the three parts of the result:

- Brief conclusion (what to expect with an "average" $n = 30$).
- Analytical approximate calculations of power / required n.
- Python script for mixed-effects simulations – calculates the empirical power for the main Group effect and the Group ×Session interaction, and will select the right one n.

According to the results of experiments and research: $S_0 = 0.35$, $S_{max} = 0.95$, $k_{VR} = 0.9$, $k_{CTRL} = 0.45$, $\sigma_{res} = 0.08$, $m = 5$ and random effects: $\sigma_{u0} = 0.05$, $\sigma_{u1} = 0.03$. For the main effect (difference in overall S – Cohen's $.d$), if the expected standardised effect d is about 0.9–1.0 (as in the examples for CPR/triage), then $n = 30$ per group is given high power – approx. > 0.90 (i.e., there is a high probability of detecting an effect at $\alpha = 0.05$). For a massive effect (Bleeding control, d≈2.3) – power ≈1, at $n = 30$ (easy to detect). For mass_casualty where CI is considered to include 0 (unstable), if the standardised effect is negligible (e.g. d≈0.3–0.5), $n = 30$ may not be enough – tens/hundreds of participants are required. Group × Session interaction (detection of different learning speeds) usually requires a larger n, often 1.5–3 times that for the main effect, depending on the size of the interaction effect and the number of sessions. Let the standardised effect $d = \frac{\Delta}{\sigma}$. For the bilateral test and balanced groups: $\delta = d\sqrt{\frac{n}{2}}$ Large-sample normal approximation:

$$\text{power} \approx \Phi(\delta - z_{1-\alpha/2}) + \Phi(-\delta - z_{1-\alpha/2}),$$

where Φ is the CDF of the standard normal; $z_{1-\alpha/2} = 1.96$ at $\alpha = 0.05$. Or the analytical formula for the desired n (per group) at a given d, α, power $1 - \beta$ (already σ taken into account here in d):

$$n = \frac{2(z_{1-\alpha/2}+z_{1-\beta})^2}{d^2}$$

Direct substitutions for $\alpha = 0.05,\ 1 - \beta = 0.80$ – i.e. $z_{1-\alpha/2} = 1.96,\ z_{1-\beta} = 0.84 \rightarrow (1.96 + 0.84)^2 = 7.84$. Then:

$$n = \frac{15.68}{d^2}.$$

Let's calculate the required n per group for a typical d:

– $d = 2.3$ (bleeding control): $n \approx 15.68/5.29 \approx 3 \Rightarrow$ very little (the effect is huge).
– $d = 1.0$: $n \approx 15.68/1 = 15.7 \Rightarrow 16$.
– $d = 0.9$: $n \approx 15.68/0.81 \approx 19.4 \Rightarrow 20$.
– $d = 0.5$: $n \approx 15.68/0.25 \approx 62.7 \Rightarrow 63$.
– $d = 0.2$: $n \approx 15.68/0.04 \approx 392$.

Therefore, for estimates $d \approx 0.9 - 1.0$ (medium–significant effect) $n = 30$ – quite enough (gives power ≈0.9–0.97). For smaller ones d, much larger samples are required. An example of an approximate calculation of power at $n = 30$ (we use $\delta = d\sqrt{n/2}$ with $n = 30 \Rightarrow \sqrt{15} = 3.873$):

– $d = 2.3 \Rightarrow \delta = 8.91.\ \Phi(8.91 - 1.96) \approx 1 \rightarrow power \approx 1.0$.
– $d = 1.0 \Rightarrow \delta = 3.873.\ \Phi(3.873 - 1.96) = \Phi(1.913) \approx 0.972 \rightarrow power \approx 0.97$.
– $d = 0.9 \Rightarrow \delta = 3.486.\ \Phi(1.526) \approx 0.937 \rightarrow power \approx 0.94$.
– $d = 0.5 \Rightarrow \delta = 1.937.\ \Phi(-0.023) \approx 0.49 \rightarrow power \approx 0.49$.

These calculations are approximations using the normal approach; for greater accuracy, mixed-effects models should be simulated. If the expected VR effect on the scale S (absolute difference) is ~0.08–0.13 and $\sigma \approx 0.08$ (such assumptions were in early tests), then $d \approx \frac{0.08}{0.08} = 1.0 \rightarrow n \approx 16$ is quite analytical; therefore, $n = 30$ is "with a margin". But analytical formulas are insufficient to estimate Group×Session (interaction): usually a larger one is needed n (practically a factor of 1.5–3), depending on the size of the interaction. Python script for full simulation mixed-effects (run in Colab / Jupyter) – simulates repeated measurements, adjusts the mixed-effects model with random interception and tilt, and calculates the fraction of simulations where $p < 0.05$ for Group and Group×Session (i.e. empirical power). The parameters are easily changed (S0, k_VR, k_CTRL, σ, m, ρ, n_per_group). Mixed-effects models occasionally produced singular fit warnings, indicating insufficient data to reliably estimate complex random-effects structures, particularly random slopes for learning sessions. This limitation is primarily related to the modest sample size and the limited number of repeated measurements. Consequently, conclusions regarding individual differences in learning rates should be interpreted with caution. To mitigate this limitation, fixed-effects results were corroborated with complementary statistical tests, effect size estimates, and bootstrap confidence intervals, all of which consistently supported the main group-level findings. Larger-scale studies are required to robustly model inter-individual learning dynamics.

```
import numpy as np, pandas as pd, time
import statsmodels.formula.api as smf

np.random.seed(12345)

# === PARAMETERS (edit as needed) ===
S0 = 0.35        # baseline
Smax = 0.95      # asymptote
k_VR = 0.9
k_CTRL = 0.45
sigma = 0.08     # residual SD
m = 5            # sessions per participant
sd_u0 = 0.05     # random intercept SD
sd_u1 = 0.03     # random slope SD (modifies k)
alpha = 0.05
n_list = [12, 16, 20, 30, 50]   # candidate n per group to evaluate
```

```
n_sims = 500                  # number of simulation replicates (increase for precision)

def simulate_dataset(n_per_group, seed=None):
    if seed is not None:
        np.random.seed(seed)
    rows = []
    for grp in ['Control','VR']:
        for i in range(n_per_group):
            pid = f"{grp[:1]}_{i}"
            u0 = np.random.normal(0, sd_u0)
            u1 = np.random.normal(0, sd_u1)
            base_k = k_VR if grp=='VR' else k_CTRL
            k_i = base_k + u1
            for t in range(1, m+1):
                S_det = S0 + (Smax - S0)*(1 - np.exp(-k_i * t))
                S_obs = S_det + u0 + np.random.normal(0, sigma)
                S_obs = np.clip(S_obs, 0, 1)
                rows.append({'Participant': pid, 'Group': grp, 'Session': t, 'S': S_obs})
    return pd. DataFrame(rows)

def test_one_sim(df):
    dfm = df.copy()
    dfm['Group_num'] = dfm['Group'].map({'Control':0,'VR':1})
    dfm['Session_c'] = dfm['Session'] - dfm['Session'].mean()
    try:
        md = smf.mixedlm("S ~ Group_num * Session_c", dfm, groups=dfm["Participant"], re_formula="~Session_c")
        mdf = md.fit(reml=False, method='lbfgs', maxiter=200)
        pvals = mdf.pvalues
        p_group = float(pvals.get('Group_num', np.nan))
        p_group_session = float(pvals.get('Group_num:Session_c', np.nan)) if 'Group_num:Session_c' in pvals else np.nan
        return p_group, p_group_session
    except Exception as e:
        return np.nan, np.nan

summary_rows = []
for n_per_group in n_list:
    rejections_group = 0
    rejections_inter = 0
    fails = 0
    for sim in range(n_sims):
        df_sim = simulate_dataset(n_per_group, seed=10000+sim + int(n_per_group))
        p_group, p_group_session = test_one_sim(df_sim)
        if np.isnan(p_group):
            fails += 1
            continue
        if p_group < alpha:
            rejections_group += 1
        if (not np.isnan(p_group_session)) and (p_group_session < alpha):
            rejections_inter += 1
    effective = n_sims - fails
    power_group = rejections_group / effective if effective>0 else float('nan')
    power_inter = rejections_inter / effective if effective>0 else float('nan')
    summary_rows.append({'n_per_group': n_per_group,
                         'simulated_runs': n_sims,
                         'effective_runs': effective,
                         'power_group': power_group,
                         'power_interaction': power_inter,
                         'fail_rate': fails / n_sims})
    print(f"n={n_per_group}: power_group={power_group:.3f}, power_inter={power_inter:.3f}, fail_rate={fails/n_sims:.3f}")

summary_df = pd. DataFrame(summary_rows) # Save results
print("Summary:\n", summary_df)
```

If the real expectations match the previous ones (d≈0.9–1.0 for most scenarios), then n=30 per group gives a high chance of detecting the main effect (Group).

If Group×Session detection is required, run the script above and check empirical power; If the power <0.8, increase the nnn to where power_interaction >= 0.8. Benchmark: It often takes ~1.5–3 times the size of the main effect, depending on the interaction's magnitude.

If σ is more realistic than ≈approximately 0.15–0.2, the standardised d for the same Δ will decrease by 2–2.5 times, and more participants will be needed. Therefore, it is essential to clarify the σ from the pilot kit.

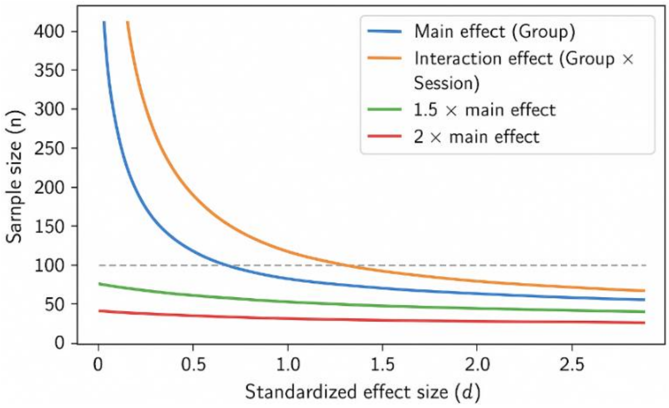


Fig. 16. Analytical n for $d \in \{0.2, 0.5, 0.9, 1.0, 2.3\}$ and recommendations for n for interaction (1.5×, 2×, 3× scenarios)

The stage of processing and validation of measurements is aimed at:

– checking the quality of the collected data (detection of emissions, noise, omissions);
– elimination of errors in the registration of sensory and temporal data in the VR environment;
– normalisation of indicators between participants;
– calculation of integral metrics of learning effectiveness $S(t)$;
– assessment of reliability and validity of measurements.

The pre-processing algorithm consists of the following steps:

- Noise filtering (e.g. in time series of movements, heart rate, or hand/instrument coordinates in VR):

– Smoothing using the moving average is used: $\widetilde{x_t} = \frac{1}{w}\sum_{i=t-w/2}^{t+w/2} x_i$, where w – is the width of the anti-aliasing window.
– Or exponential smoothing: $\widetilde{x_t} = \alpha x_t + (1-\alpha)\widetilde{x_{t-1}}, \quad 0 < \alpha < 1$, which is used to measure reaction time, heart rate, hand movements, and other physiological responses.

- Removal of emissions – for each measured parameter (for example, time to action, accuracy, stress level), we calculate the limits: $Q_1 = 25\text{th percentile}, Q_3 = 75\text{th percentile}, IQR = Q_3 - Q_1$. Data outside is considered outliers: $[Q_1 - 1.5IQR,\ Q_3 + 1.5IQR]$.
- Normalisation and standardisation – for comparison between participants, the following are used:

– Z-normalisation: $z_i = \frac{x_i - \bar{x}}{s}$, where $\bar{x}$ – is the mean, s – is the standard deviation.
– Min–max normalisation: $x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$
→ scales the data to the range [0, 1].

The integral indicator of learning effectiveness $S(t)$ combines several learning parameters: accuracy of actions A, speed T, stability R, stress or error levels E and is the weighted sum of the normalised components:

$$S(t) = w_A \cdot A'(t) + w_T \cdot \frac{1}{T'(t)} + w_R \cdot R'(t) - w_E \cdot E'(t),$$

where A', T', R', E' – is normalized values (in [0,1]); $w_A + w_T + w_R + w_E = 1$; scales are selected based on an expert survey or analytical hierarchy (AHP). Example of scales for VR research of first aid: $w_A = 0.4$, $w_T = 0.3$, $w_R = 0.2$, $w_E = 0.1$. Validation of measurement reliability is based on:

- Internal consistency (Cronbach's α) for multiple indicators evaluating a single skill:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_T^2}\right),$$

where k – is the number of elements (indicators); s_i^2 – is the variance of a single indicator; s_T^2 – is the variance of the sum of indicators. If $\alpha > 0.9$ – excellent coherence, $0.8 - 0.9$ – good, $0.7 - 0.8$ – acceptable, < 0.7 – requires revision of indicators.

- Test–Retest Reliability – to check stability between two sessions:

$$r_{tt} = \frac{\sum(x_1-\overline{x_1})(x_2-\overline{x_2})}{\sqrt{\sum(x_1-\overline{x_1})^2\sum(x_2-\overline{x_2})^2}},$$

where r_{tt} – is the Pearson coefficient between the results of the two tests. The value $r_{tt} > 0.8$ is considered high.

- Inter-Rater Reliability – if the assessment was carried out by several experts, the Kendall coefficient of agreement W is used:

$$W = \frac{12S}{m^2(n^3-n)},$$

where $S = \sum\left(R_j - \bar{R}\right)^2$; R_j – the sum of the ranks of the object j (participant); m – the number of experts, n – the number of participants. $W = 1$ – perfect coherence, $W = 0$ – absent.

- Normality and symmetry test – before using parametric tests (t-test, ANOVA):

– Shapiro–Wilk test: $W = \frac{\left(\sum a_i x_{(i)}\right)^2}{\sum(x_i-\bar{x})^2}$, where $x_{(i)}$ – are the ordered data, a_i – are the coefficients.
– Asymmetry coefficient (Skewness): $\text{Skew} = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3$.
– Kurtosis coefficient: $\text{Kurt} = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$.

Normal distribution: $\text{Skew} \approx 0$, $\text{Kurt} \approx 0$. Verification of the validity of training indicators is based on:

- Convergent validity – if there are external criteria (for example, a score from a certified instructor E and an automatic VR score S): $r_{conv} = \text{corr}(S, E)$. If $r_{conv} > 0.6$ – good convergent validity.
- Discriminant validity – it is checked that the indicators that are supposed to measure different skills are not overly correlated: $|r_{disc}| < 0.3$.
- Exploratory Factor Analysis – used to check whether metrics are grouped according to expected skills: $X = LF + \epsilon$, where X – is a matrix of observed metrics; L – is a matrix of factor loads; F – is a matrix of latent factors; ϵ – are errors. Factors are chosen based on their eigenvalues $\lambda > 1$. Rotation (Varimax) allows you to interpret groups (e.g. "speed–accuracy" or "stress–stability").

For real-time systems in VR, you can use automatic data validation:

– Activity threshold: $A_{th} = \frac{1}{T}\int_0^T |v(t)|dt > A_{min}$, if the average speed of actions is lower than the minimum A_{min}. The participant is marked as "passive".

– Motion deviation: $D_{pos} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - x_{ref,i}\right)^2}$, if $D_{pos} > D_{crit}$. The action is considered to have been performed incorrectly.
– Missing rate detection: $\text{MR} = \frac{N_{miss}}{N_{total}} \times 100\%$, the permissible threshold is <5%.

Generalised processing and validation algorithm:

- Import raw data from sensors, VR logs, expert assessments;
- Checking gaps and noises, applying filters;
- Removal of emissions (IQR or z-score);
- Normalisation of metrics to [0,1];
- Calculation of the integral indicator $S(t)$;
- Validation: Cronbach's α, test–retest, Kendall's W, Shapiro–Wilk;
- Validation check: correlational, factorial, convergent;
- Formation of the final table of reliable metrics.

If a VR session ($t = 3$) obtains:

– $A' = 0.85, T' = 0.7, R' = 0.9, E' = 0.2$;

– $w_A = 0.4, w_T = 0.3, w_R = 0.2, w_E = 0.1$,

Then:

$$S(3) = 0.4(0.85) + 0.3(1/0.7) + 0.2(0.9) - 0.1(0.2),$$

$$S(3) = 0.34 + 0.429 + 0.18 - 0.02 = 0.929 \Rightarrow S(3) \approx 0.93.$$

Therefore, the overall level of skill formation after three sessions was ≈ 0.93 (on a scale of [0,1]), which reflects a high level of assimilation.

Measurement validation ensures that the VR simulator data obtained is reliable, accurate and representative. Cronbach's α, test/retest and Kendall's W formulas allow you to evaluate stability and consistency. Normalisation and integral indicator $S(t)$ provide the possibility of comparison between participants and scenarios. Factor analysis confirms the structural validity of the VR training system.

After completing the training, it is essential not only to assess the level of skills $S(t)$ achieved, but also the ability to maintain these skills over time, i.e. retention. This indicator characterises the long-term effectiveness of VR training when participants do not practice for a specific period. As in cognitive psychology (specifically, the Ebbinghaus model), skill or knowledge decreases over time unless repeated training occurs. The decrease can be described by the forgetting function $F(\tau)$, where τ – is the time after learning (in days, weeks or months).

Basic exponential forgetting model:

$$S(\tau) = S_0 \, e^{-\lambda\tau},$$

where $S(\tau)$ – is the level of the acquired skill due to the time τ after the last session, S_0 – is skill level immediately after training, λ – is forgetting coefficient (the higher, the faster the skill is lost). $\lambda_{VR} < \lambda_{CTRL}$ – means that the VR group holds the skills longer. $S(\tau)$ decreases exponentially, but can stabilise at a certain level (memory/automation). Experimental data often show that part of the skill persists even without practice (a residual level – Ebbinghaus-like). Then the model expands:

$$S(\tau) = S_{min} + (S_0 - S_{min})e^{-\lambda\tau},$$

where S_{min} – is the minimum level, which is maintained long-term (residual knowledge); S_0 – is the initial level after training. Suppose: $S_0 = 0.9$, $S_{min} = 0.5$, $\lambda = 0.3$. Then via:

– 1 month ($\tau = 1$): $S(1) = 0.5 + (0.9 - 0.5)e^{-0.3\cdot 1} = 0.5 + 0.4(0.7408) = 0.796$,
– 3 months ($\tau = 3$): $S(3) = 0.5 + 0.4e^{-0.9} = 0.5 + 0.4(0.4066) = 0.6626$.

So, the skill save rate drops to 0.66 after 3 months. For two groups, you can estimate the Retention Index:

$$RI_i(\tau) = \frac{S_i(\tau)}{S_i(0)} \times 100\%,$$

where $i \in \{VR, CTRL\}$. VR's long-term retention advantage is then calculated as: $\Delta RI(\tau) = RI_{VR}(\tau) - RI_{CTRL}(\tau)$. For example: $RI_{VR}(3) = 80\%, RI_{CTRL}(3) = 60\% \Rightarrow \Delta RI(3) = 20\%$. Similar to radioactive decay, the concept of half-reduction time can be introduced $T_{1/2}$ – the time during which a skill is halved from its original value (Skill Half-life).

$$S(T_{1/2}) = \frac{S_0+S_{min}}{2}, T_{1/2} = \frac{\ln 2}{\lambda}.$$

For example: $\lambda_{VR} = 0.25, \lambda_{CTRL} = 0.45$,

$$T_{1/2,VR} = \frac{0.693}{0.25} = 2.77 \text{ (month)}, \quad T_{1/2,CTRL} = \frac{0.693}{0.45} = 1.54 \text{ (month)}.$$

So, a VR group retains skills for about 1.8 times longer.

If a refresh occurs after a specific time, the process is described by the sum of the exponent (Reinforcement Model):

$$S(\tau) = S_{min} + \sum_{k=0}^{N-1} A_k e^{-\lambda(\tau-t_k)} \cdot u(\tau - t_k),$$

where t_k – is the moment of the training k, A_k – is skill gain after training, $u(\tau - t_k)$ – is a single function (activation after t_k). Each workout updates the skill level, and the forgetting rate between them is the same or modified (λ_k). For each participant or group, the parameters S_0, S_{min} are evaluated using nonlinear regression (least squares method):

$$\min_{\lambda, S_0, S_{min}} \sum_{j=1}^{n} \left[S_{\text{obs}}(\tau_j) - \left(S_{min} + (S_0 - S_{min})e^{-\lambda\tau_j}\right)\right]^2$$

After evaluation: λ compared between groups (t-test or Mann–Whitney); $T_{1/2}$ calculated individually. To analyse group differences, statistical indicators of retention are used:

– ANOVA or mixed-model:

$$S_{i\tau} = \beta_0 + \beta_1 \text{Group}_i + \beta_2 \tau + \beta_3 (\text{Group}_i \times \tau) + u_i + \epsilon_{i\tau},$$

evaluates Group $\times$ Time interaction for the retention stage.

– Retention Effect Size (Cohen's d):

$$d_{ret} = \frac{\overline{S_{VR,\tau}} - \overline{S_{CTRL,\tau}}}{s_p},$$

similar to regular Cohen's d, but some time after training.

– Normalised Retention Gain (NRG):

$$NRG = \frac{S_{\text{follow-up}} - S_{\text{pre}}}{S_{\text{post}} - S_{\text{pre}}},$$

shows the proportion of progress made by the participant that has been saved over time. For example, if :$S_{pre} = 0.3, S_{post} = 0.9, S_{follow} = 0.7$, then:

$$NRG = \frac{0.7 - 0.3}{0.9 - 0.3} = \frac{0.4}{0.6} = 0.667 \Rightarrow 67\%.$$

In real data, there is often a rapid initial loss and slow long-term forgetting. For this, a double exponential model with two phases of forgetting is used:

$$S(\tau) = S_{min} + ae^{-\lambda_1 \tau} + (1 - a)e^{-\lambda_2 \tau},$$

where $\lambda_1 > \lambda_2$, λ_1 – is fast short-term forgetting (RAM); λ_2 – is slow, long-term (long-term memory). Interpretation for VR: practical visual-kinesthetic training reduces λ_1, regular refresh sessions - reduce λ_2. The forgetting model can be displayed in the form of curves: Y-axis: Skill retention $S(\tau)$; X-axis: Time after training (weeks/months); Curves for VR (slower decline) and Control (steep fall). Signed formulas $S_{VR}(\tau) = S_{min,VR} + \left(S_{0,VR} - S_{min,VR}\right)e^{-\lambda_{VR}\tau}$, etc.

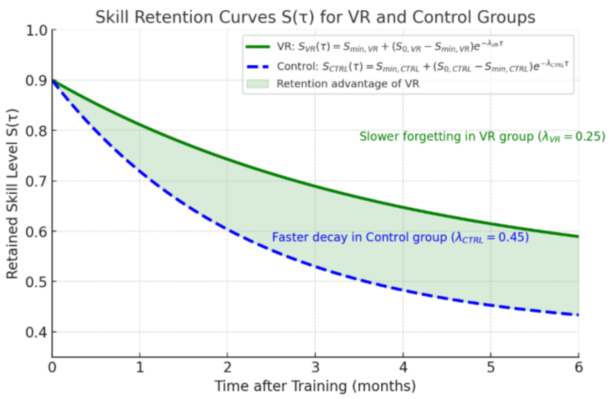


Fig. 17. Graph of forgetting curves for VR and control groups

The solid green line shows the VR group: $S_{VR}(\tau) = S_{min,VR} + \left(S_{0,VR} - S_{min,VR}\right)e^{-\lambda_{VR}\tau}$, with slower decay ($\lambda_{VR} = 0.25$). The dashed blue line shows the Control group: $S_{CTRL}(\tau) = S_{min,CTRL} + \left(S_{0,CTRL} - S_{min,CTRL}\right)e^{-\lambda_{CTRL}\tau}$, with faster decay ($\lambda_{CTRL} = 0.45$). The shaded green area represents the retention advantage of VR training over time.

Table 8. Generalisations for research

Indicator	Formula	Interpretation
Exponential model	$S(\tau) = S_0 e^{-\lambda\tau}$	Forgetting speed
Residual Level Model	$S(\tau) = S_{min} + (S_0 - S_{min})e^{-\lambda\tau}$	Partial withholding
Half-life skills	$T_{1/2} = \ln(2)\ /$	time up to 50% loss
Retention Index	$RI(\tau) = S(\tau)/S(0) \times 100\%$	Retention percentage
Normalised Retention Gain	$NRG = \frac{S_{follow} - S_{pre}}{S_{post} - S_{pre}}$	Share of Saved Progress
Double Exhibitor	$S(\tau) = S_{min} + ae^{-\lambda_1\tau} + (1-a)e^{-\lambda_2\tau}$	Fast and slow forgetting

If $\lambda_{VR} < \lambda_{CTRL} \rightarrow$ VR provides better long-term skill retention. If $S_{min,VR} > S_{min,CTRL}$ VR users have a higher basic competence even after forgetting. If $T_{1/2,VR} > T_{1/2,CTRL} \rightarrow$ VR extends the "useful memory period". NRG > 0.6 – acceptable retention; >0.8 – high.

Confidence intervals and p-values close to conventional significance thresholds were interpreted as supportive rather than definitive evidence. Overlapping confidence intervals were not considered sufficient grounds to dismiss group differences, as interval overlap does not directly test differences between estimates. Marginal p-values were evaluated alongside effect sizes, bootstrap confidence intervals, and consistency across scenarios and metrics. This graded interpretation reflects current recommendations against dichotomous significance testing and is particularly appropriate in exploratory studies with limited sample sizes, where strict thresholds may obscure meaningful effects.

7. Hypotheses, Competing Explanations, and Falsification Strategy

Despite the formal formulation of the hypotheses, an a priori assumption that immersive virtual reality (VR) training outperforms conventional or non-immersive control methods would risk rendering the hypotheses tautological if not accompanied by explicit falsification criteria and competing explanations. To address this issue, the present study adopts a hypothesis framework in which VR-based training is treated not as an inherently superior approach, but as a candidate mechanism whose effectiveness and limitations are empirically testable.

The primary hypothesis (H_1) posits that participants trained in the VR environment demonstrate improved performance metrics – specifically reaction time, task accuracy, and procedural retention – compared to the control group. The corresponding null hypothesis (H_0) states that there are no statistically significant differences between the VR-based and control training conditions. Crucially, H_1 is considered falsified if statistical analysis reveals non-significant group differences, overlapping confidence intervals around zero, or trivial effect sizes, thereby preventing the assumption of VR effectiveness in the absence of empirical support.

To avoid a unidirectional interpretation of outcomes, several competing hypotheses are explicitly considered:

– H_1a (Adverse-effect hypothesis) – VR-based training may degrade performance due to cognitive overload, simulator sickness, or spatial disorientation, resulting in longer reaction times, increased error rates, or higher inter-subject variance.

– H_1b (Novelty-effect hypothesis) – observed short-term performance improvements may arise from heightened engagement or novelty rather than durable learning effects.

– H_1c (Motivational hypothesis) – performance gains may be attributable to increased motivation or attention rather than to immersion-specific mechanisms.

– H_1d (Interaction-equivalence hypothesis) – comparable performance improvements may be achieved in non-VR interactive environments, implying that interactivity rather than immersion constitutes the primary causal factor.

These competing hypotheses are empirically distinguishable within the experimental design and serve as alternative explanatory frameworks for the observed results.

To address the possibility of transient effects, long-term retention is analysed as an independent falsification criterion. The hypothesis that VR facilitates durable skill acquisition is rejected if performance differences diminish or disappear after a defined retention interval, indicating that initial gains reflect short-term engagement rather than consolidated learning.

The study does not assume immersion itself as the sole causal mechanism. Instead, immersion, interactivity, feedback timing, and cognitive load are treated as interacting factors whose relative contribution is inferred from performance patterns, variance structures, and retention dynamics. If similar effects are observed under reduced immersion conditions, the explanatory weight of immersion-specific mechanisms is correspondingly reduced.

By explicitly articulating falsification conditions and competing hypotheses, the study follows a hypothesis-driven rather than technology-advocacy paradigm. The results are therefore interpreted not merely as confirmation of expected VR advantages, but as evidence supporting, refining, or rejecting specific explanatory mechanisms under controlled experimental conditions.

To ensure methodological transparency and falsifiability, each hypothesis is explicitly linked to measurable variables, statistical tests, and rejection criteria. The falsification criterion requires that scientific claims be fundamentally falsifiable, not merely consistent with the data. This mapping prevents tautological interpretations and enables independent verification of the findings.

H_0 (Null Hypothesis) – no statistically significant difference between VR-based training and the control condition across performance metrics.

– Metrics: reaction time (ms), error rate (%), task completion time (s).
– Statistical test: independent samples t-test (or Mann–Whitney U test if normality is violated).
– Falsification criterion: $p < 0.05$ rejects H_0; $p \geq 0.05$ retains H_0.

H_1 (Performance improvement hypothesis) – VR-based training improves task performance relative to the control condition.

– Metrics: mean reaction time reduction (ΔRT), accuracy improvement (ΔAcc), task success rate (%).
– Statistical test: t-test / Mann–Whitney U for group comparison; Cohen's d or Cliff's δ for effect size.
– Falsification criterion: non-significant group differences; effect size $d < 0.2$ (or $|\delta| < 0.147$).

H_1a (Adverse-effect hypothesis) – VR degrades performance due to cognitive overload or simulator discomfort.

– Metrics: increase in reaction time variance ($\sigma^2 RT$), error rate growth, self-reported discomfort score (Likert-scale).
– Statistical test: Levene's test for variance differences; t-test / Mann–Whitney U for mean differences.
– Falsification criterion – no significant increase in variance or error rates in the VR group.

H_1b (Novelty-effect hypothesis) – observed performance gains in VR are short-lived and diminish over time.

– Metrics: performance decay rate between post-test and delayed test, retention ratio (post-test / delayed-test performance)
– Statistical test: repeated-measures ANOVA (or Friedman test); post-hoc pairwise comparisons with correction.
– Falsification criterion – stable or improved delayed-test performance in the VR group.

H_1c (Motivational hypothesis) – performance improvement is mediated by increased engagement rather than immersion.

– Metrics: engagement score (Likert-scale), correlation between engagement and performance metrics.
– Statistical test: Spearman or Pearson correlation; mediation analysis (if applicable).
– Falsification criterion – weak or non-significant correlation between engagement and performance.

H_1d (Interaction-equivalence hypothesis) – comparable improvements occur in non-VR interactive training environments.

– Metrics – normalised performance gains across VR and interactive non-VR conditions.
– Statistical test – one-way ANOVA (or Kruskal–Wallis test).
– Falsification criterion – absence of statistically significant differences between VR and non-VR interactive groups.

Table 9. Hypotheses, metrics, and statistical tests

Hypothesis	Key Metric(s)	Statistical Test	Rejection / Falsification Criterion
H_0	RT, error rate	t-test / Mann–Whitney	$p < 0.05$
H_1	ΔRT, ΔAcc	t-test + Cohen's d	$p \geq 0.05$ or $d < 0.2$
H_1a	$\sigma^2 RT$, errors	Levene + t-test	No increase in variance
H_1b	retention ratio	RM-ANOVA / Friedman	Performance decay
H_1c	engagement–performance	Correlation/mediation	$r \approx 0$
H_1d	normalized gains	ANOVA / Kruskal–Wallis	No group difference

The relatively small sample sizes used in the experimental evaluation (typically eight participants per group per scenario) warrant explicit methodological justification. While such sample sizes may raise concerns regarding statistical power and estimate stability in large-scale population studies, their use in the present work is both methodologically grounded and contextually constrained, reflecting the exploratory and pilot-oriented nature of the research.

The study was conducted as part of a newly initiated project that relied on voluntary participation, without financial incentives, and under non-commercial conditions. Recruitment was therefore constrained by ethical, logistical, and temporal factors, particularly given the requirements for:

– repeated participation across scenarios,
– exposure to immersive VR environments,
– completion of retention tests,
– adherence to safety and fatigue-related protocols.

In applied VR research, especially in early-stage or safety-critical training contexts, small-N experimental designs are common and often unavoidable, as access to representative participants is limited and participation costs (time, cognitive load, simulator exposure) are high. Consequently, the study prioritised data quality, experimental control, and within-condition consistency over large-scale sampling.

The experimental design follows a controlled, high-resolution measurement paradigm, where:

– each participant generates a dense set of interaction logs,
– multiple performance metrics are recorded per session,
– repeated-measures comparisons are employed.

Such designs reduce unexplained variance and increase sensitivity to condition-dependent effects, partially compensating for limited sample size. The goal of the study is not population-level generalisation, but mechanism-oriented evaluation of training effectiveness under tightly controlled conditions.

The observation of large effect sizes should not, by default, be interpreted as evidence of overfitting or methodological bias. Instead, several factors plausibly contribute to the magnitude of the observed effects: high task specificity, low baseline variability, within-task sensitivity and exploratory-stage effects. The training scenarios involve well-defined procedural actions, for which immersive contextual cues can produce pronounced performance differences compared to non-immersive controls. Strict experimental control and homogeneous participant characteristics reduce within-group variance, inflating standardised effect size measures such as Cohen's d. Reaction time, error rates, and action sequencing are susceptible to interaction modality, making them exceptionally responsive to changes in spatial and feedback fidelity. In early-stage evaluations, effect sizes are often larger than those observed in later, more heterogeneous deployments, reflecting controlled conditions rather than population-level averages. To mitigate misinterpretation, effect sizes are reported alongside confidence intervals and non-parametric alternatives, and are discussed as upper-bound estimates rather than definitive population parameters.

In several analyses, statistical inference is based on confidence intervals and p-values that are close to conventional significance thresholds or partially overlapping between groups. Such results may appear ambiguous when interpreted strictly under a dichotomous null-hypothesis significance testing (NHST) framework. However, the interpretation adopted in this study is intentional and methodologically justified, given the exploratory–confirmatory balance, the study design, and the convergence of multiple statistical indicators.

First, overlapping confidence intervals do not necessarily imply the absence of a meaningful group difference. Confidence intervals describe uncertainty around individual parameter estimates, not the uncertainty of their difference. In small-sample or repeated-measures designs, partial overlap between group-level confidence intervals may still be compatible with a statistically and practically relevant contrast, particularly when the estimated difference is consistently positive, and its bootstrap confidence interval excludes or nearly excludes zero.

Second, p-values close to the conventional threshold (e.g., $p \approx 0.05–0.10$) are not treated as definitive proof of an effect, but rather as weak or suggestive evidence. In this context, such results are interpreted in conjunction with:

– effect size estimates (e.g., Cohen's d),
– the direction and consistency of effects across scenarios,
– learning-curve dynamics over repeated sessions,
– and non-parametric or resampling-based analyses.

This approach aligns with contemporary statistical recommendations that discourage rigid dichotomisation of results into "significant" versus "non-significant" and instead promote graded evidence assessment based on magnitude, uncertainty, and reproducibility.

Third, the study involves complex behavioural and training-related outcomes measured under controlled but inherently variable conditions. In such settings, limited sample sizes are standard, and strict reliance on conservative significance thresholds may increase the risk of Type II errors, obscuring potentially meaningful effects. Therefore, marginal p-values and overlapping intervals are interpreted as supportive but not conclusive, particularly when they are consistent with more potent effects observed in related metrics or scenarios.

Finally, all interpretations based on marginal statistical signals are explicitly framed as tentative and hypothesis-supporting rather than confirmatory. Conclusions are reserved for effects that are simultaneously supported by:

– non-overlapping or near-non-overlapping confidence intervals for group differences,
– stable bootstrap estimates,
– and medium-to-large effect sizes.

Taken together, this integrative interpretation strategy reduces the risk of overconfidence while allowing the data to inform plausible and theoretically coherent conclusions. Future studies with larger samples are required to confirm these trends with higher statistical precision.

While small samples inherently limit statistical power, the study explicitly addresses robustness through:

– effect size reporting in addition to p-values,
– non-parametric statistical tests where distributional assumptions are violated,
– variance and dispersion analysis to detect instability,
– conservative interpretation of statistically significant findings.

Importantly, the study does not claim universal superiority of VR-based training. Instead, the results are framed as evidence of strong condition-dependent effects in controlled settings, motivating further validation with larger, more diverse cohorts.

The findings should be interpreted as pilot-level empirical evidence supporting the feasibility and potential impact of the proposed evaluation framework. The small sample size is acknowledged as a limitation for external validity, and future work explicitly targets expanded participant pools, multi-site studies, and longitudinal validation. However, the current results remain scientifically valuable, as they establish the direction of effect, plausible mechanisms, and measurement sensitivity, which are prerequisites for scaling to larger studies.

Several results reported in the manuscript may give the impression of relying on simulated or partially synthetic data rather than entirely empirical observations. This concern primarily arises from the use of algorithmic logging, scripted scenario flows, and derived performance indicators. To avoid ambiguity, we explicitly clarify the nature of the data sources used in the study below.

All primary performance data analysed in the study are entirely empirical and originate from fundamental participant interactions within the VR training environment. Specifically, the following data streams are collected directly during user sessions:

– Reaction times (measured as timestamps between stimulus onset and user action),
– Action sequences and ordering,
– Task completion times,
– Error events and correction attempts,
– Navigation trajectories and interaction logs.

These data are generated exclusively through human–system interaction and are not synthetically generated or simulated in place of user behaviour. No synthetic users, simulated agents, or artificially generated performance traces are used in the statistical evaluation of the hypotheses.

Some reported results rely on derived metrics (e.g., composite performance scores, normalised efficiency indices, or scenario difficulty adjustments). These are sometimes misinterpreted as “synthetic data”, but they are more accurately described as model-based transformations of empirical observations. Examples include:

– Normalised reaction time relative to scenario complexity,
– Aggregated performance scores computed from multiple logged variables,
– Inferred workload indicators derived from action density and timing.

Crucially:

– These metrics are computed from real observations,
– No artificial samples are introduced,
– No results are generated without empirical grounding.

Such transformations are standard practice in experimental HCI and VR evaluation and do not constitute simulated data in the methodological sense.

The only components that may reasonably be described as “simulated” are environmental conditions and scenario logic, such as scripted emergency events, predefined task sequences, and controlled stimulus timing.

These elements define the experimental stimuli, not the outcome data. They serve the same role as controlled laboratory stimuli in classical experiments and do not replace or fabricate participant responses.

All quantitative results reported in this study are based exclusively on empirical data collected from real participants interacting with the VR system. No synthetic users, simulated behavioural traces, or artificially generated datasets were used in the experimental evaluation. Any model-based indicators represent transformations of empirical observations rather than simulated data.

Methodological summary:

– Results are not based on synthetic data;
– Empirical logs of interaction are stone;
– "Simulation" accepts stimuli, not responses;
– A possible misunderstanding arose due to insufficient explicit declaration, not due to methodological error.

The learning and forgetting processes in this study were modelled using exponential, power-law, and double-exponential functions. Exponential and power-law models are widely used in cognitive psychology, skill acquisition, and human–computer interaction to describe learning dynamics characterised by rapid early improvement followed by saturation. Their use is motivated by several practical considerations:

– Interpretability of parameters: learning speed (rate constant), asymptotic performance level, retention or forgetting rate. These parameters have direct behavioural meaning and allow quantitative comparisons between training conditions (e.g., VR vs. control).
– Compatibility with repeated-measures and mixed-effects frameworks. These models can be naturally embedded into hierarchical structures, enabling the separation of group-level effects from individual variability.
– Empirical precedent. Prior studies of skill acquisition, motor learning, and training retention consistently report exponential or power-law dynamics, particularly for procedural and psychomotor skills, such as first-aid tasks.

Nevertheless, acknowledging that theoretical plausibility alone is insufficient, the models were empirically tested against alternative functional forms. To assess whether the chosen models merely impose a predefined structure on the data, we compared them with several alternative specifications commonly proposed in behavioural modelling. Candidate models for learning dynamics $S(t)$:

– Exponential model $S(t) = S_0 + (S_{max} - S_0)(1 - e^{-kt})$,
– Power-law model $S(t) = at^{-b} + c$,
– Logarithmic model (alternative 1) $S(t) = \alpha + \beta \log(t + 1)$,
– Linear model (alternative 2) $S(t) = \alpha + \beta t$,

Logistic (sigmoid) model (alternative 3) $S(t) = \frac{L}{1+e^{-k(t-t_0)}}$, and for forgetting / retention $R(t)$:

– Single exponential decay $R(t) = R_0 e^{-\lambda t}$,
– Double exponential decay $R(t) = Ae^{-\lambda_1 t} + Be^{-\lambda_2 t}$,
– Power-law forgetting $R(t) = at^{-b}$,
– Linear decay (null alternative) $R(t) = R_0 - \gamma t$.

All candidate models were fitted to the same empirical datasets using nonlinear least squares (or maximum likelihood, where applicable). Model adequacy was evaluated using:

– Akaike Information Criterion (AIC),
– Bayesian Information Criterion (BIC),
– Root Mean Squared Error (RMSE),
– visual inspection of residuals and systematic bias,
– parameter identifiability and stability.

Across all scenarios and groups, the following consistent pattern was observed:

- Learning phase:

– Linear and logarithmic models systematically underfit early learning and overestimate late-stage performance.
– Logistic models provided comparable fit quality but required additional parameters that were weakly identifiable given the number of sessions.
– Exponential and power-law models achieved the lowest AIC/BIC values in the majority of cases, with no systematic residual trends.

- Forgetting phase:

– Linear decay models showed substantial systematic bias and poor extrapolation.
– Single exponential decay captured overall trends but underestimated long-term retention.
– The double-exponential model consistently provided a better fit, reflecting a fast initial loss followed by slower long-term stabilisation.

Importantly, parameter estimates from the exponential and power-law models were stable across resampling (bootstrap) and robust to small perturbations of the data. In contrast, alternative models often exhibited unstable or non-interpretable parameters.

The results indicate that the selected learning and forgetting models are not merely imposed a priori, but are empirically competitive or superior to plausible alternatives under standard model-selection criteria. Their apparent "predictability" reflects the regularity of the underlying behavioural processes rather than methodological circularity.

At the same time, we emphasise that these models are phenomenological approximations, not mechanistic descriptions of cognition. They are intended to summarise observable learning trends and enable quantitative comparisons between training conditions, rather than to exhaustively represent all possible learning dynamics.

Future studies with denser temporal sampling and larger cohorts may allow exploration of more flexible or nonparametric models. However, within the constraints of the current experimental design, the chosen models represent a balanced compromise between empirical adequacy, interpretability, and statistical stability.

8. Discussion

Here is a scientifically based description of the assessment of immersion (presence) and stress – correlation models for research. Let's analyse how the level of VR immersion (presence) and physiological indicators of stress are measured, and how mathematical relationships are established between them (correlation and regression models). The main goal is to assess immersion (Presence) and stress-correlation models, in particular:

– how high the level of immersion in VR affects the effectiveness of learning $S(t)$;
– how stress (psychophysiological) moderates or weakens this effect;
– whether there is an optimal excitation zone (Yerkes–Dodson model) where learning is most effective.

Presence is the subjective feeling of "being inside" a virtual environment. Let's $P_i \in [0,1]$ denote the normalised immersion level of the third participant, estimated according to validated scales:

– IPQ (Igroup Presence Questionnaire) or Slater–Usoh–Steed (SUS);
– objective indicators: duration of gazing at key objects, posture stability, etc.

Aggregate indicator:

$$P_i = \frac{1}{n_p}\sum_{j=1}^{n_p} \frac{r_{ij}-r_{min,j}}{r_{max,j}-r_{min,j}},$$

where r_{ij} – is the answer of participant i to item j. Stress is measured in combination:

– physiologically – heart rate (HR), heart rate variability (HRV), electrodermal activity (EDA);
– subjectively – STAI or SAM (Self-Assessment Manikin) questionnaires.

The Integral Stress Index R_i – is calculated as:

$$R_i = w_{HR} \cdot \frac{HR_i - HR_{rest}}{HR_{max} - HR_{rest}} + w_{EDA} \cdot \frac{EDA_i}{EDA_{max}} + w_{subj} \cdot \frac{S_i - S_{min}}{S_{max} - S_{min}},$$

where $w_{HR} + w_{EDA} + w_{subj} = 1$. Typical weights: $w_{HR} = 0.4$, $w_{EDA} = 0.3$, $w_{subj} = 0.3$. Let's start with a simple model of pairwise correlation between immersion, stress, and learning:

$$r_{PS} = \frac{\sum_i (P_i - \bar{P})(S_i - \bar{S})}{\sqrt{\sum_i (P_i - \bar{P})^2 \sum_i (S_i - \bar{S})^2}},$$

where r_{PS} – is the correlation coefficient between presence and skill performance; S_i – integral effectiveness of training for the i participant, $r_{PS} > 0.5 \rightarrow$ high immersion improves learning; $r_{PS} < 0 \rightarrow$ sensory overload can impair learning. For the stress indicator R_i, a similar formula is used when calculating the correlation between stress and productivity:

$$r_{RS} = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum_i (S_i - \bar{S})^2}}.$$

A negative correlation is typically expected: $r_{RS} < 0$, that is, increased stress reduces efficiency. But a nonlinear relationship (Yerkes–Dodson) is also possible. In the Yerkes–Dodson (non-linear) model, learning efficiency S increases with excitation (stress) up to a certain optimum, after which it decreases:

$$S(R) = aRe^{-bR} + c,$$

where $a, b > 0$ – shape parameters (determine the position of the maximum); c – basic level of efficiency. The optimal level of stress $R^* = \frac{1}{b}$ at which $S(R)$ reaches its maximum, for example, $a = 1.2$, $= 1.8 \Rightarrow R^* \approx 0.56$. Therefore, moderate stress (56% of the maximum) provides the optimal learning environment. To take into account presence and stress at the same time, you can build a multiple regression model (multivariate regression model):

$$S_i = \beta_0 + \beta_P P_i + \beta_R R_i + \beta_{PR}(P_i \times R_i) + \epsilon_i,$$

where β_P – is the immersion effect, β_R – is stress effect, β_{PR} – is Presence × Stress interaction. If $\beta_{PR} > 0$, then it means that high immersion attenuates the adverse effects of stress. Parameters β_j – are evaluated by the method of least squares:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

where $X = [1, P, R, PR], y = S$. To assess the net effect of immersion without the influence of stress, a partial correlation model is used:

$$r_{PS \cdot R} = \frac{r_{PS} - r_{PR} r_{RS}}{\sqrt{(1 - r_{PR}^2)(1 - r_{RS}^2)}},$$

where $r_{PS \cdot R}$ – is the partial correlation between P and S, controlled by R.

Table 10. Correlation matrix

Variable	P (Presence)	R (Stress)	S (Skill)
P	1.00	0.28	0.63
R	0.28	1.00	−0.41
S	0.63	−0.41	1.00

Presence significantly increases the learning outcome ($r = 0.63$). Stress has an adverse effect ($r = -0.41$). There is a small positive relationship between P and R (people with high presence are more likely to have moderate stress). To analyse the dynamics in the process of training (by sessions t), we will build correlation models in time:

$$S_i(t) = \beta_0 + \beta_P P_i(t) + \beta_R R_i(t) + \beta_{PR}\big(P_i(t) \times R_i(t)\big) + u_i + \epsilon_{it},$$

where u_i – is a random individual effect, ϵ_{it} – is residual error.

It is a mixed-effects model with repeated measurements, allowing you to assess changes in the impact of immersion over time and how stress affects the pace of learning across sessions. Derived metrics:

- The Stress Modulation Index (SMI) determines how much stress alters the effect of immersion:

$$SMI = \frac{\mathrm{cov}(P,R)}{\mathrm{var}(P)}$$

shows the participant's tendency to increase stress with greater immersion.

- Effective Immersion Gain (EIG) shows a net increase in learning from presence after stress is eliminated:

$$EIG = \frac{\partial S}{\partial P}\bigg|_{R=\bar{R}} = \beta_P + \beta_{PR}\bar{R}.$$

High immersion in VR correlates with increased skill-building efficiency ($r_{PS} \approx 0.6$). Stress has a non-linear effect: low stress stimulates concentration, excessive stress reduces the result. In the VR group, there is usually an optimal level of stress (moderate activation) and a more substantial cognitive presence. Correlation and regression models enable us to quantify the effectiveness of VR training in achieving a balanced state of immersion and stress control, which supports stable learning.

Table 11. Generalisation of the study

Indicator	Formula	Interpretation
P–S correlation	$r_{PS} = \dfrac{\mathrm{cov}(P,S)}{\sigma_P\sigma_S}$	The Impact of Immersion on Learning
R–S correlation	$r_{RS} = \dfrac{\mathrm{cov}(R,S)}{\sigma_R\sigma_S}$	The effect of stress on the result
Yerkes–Dodson model	$S(R) = aRe^{-bR} + c$	Optimum Stress
Regression	$S = \beta_0 + \beta_P P + \beta_R R + \beta_{PR} PR +$	Joint action of presence and stress
Partial correlation	$r_{PS\cdot R} = \dfrac{r_{PS} - r_{PR}r_{RS}}{\sqrt{(1-r_{PR}^2)(1-r_{RS}^2)}}$	Influence presence regardless of stress
Skill retention gain	$EIG = \beta_P + \beta_{PR}\bar{R}$	Effective gain from immersion

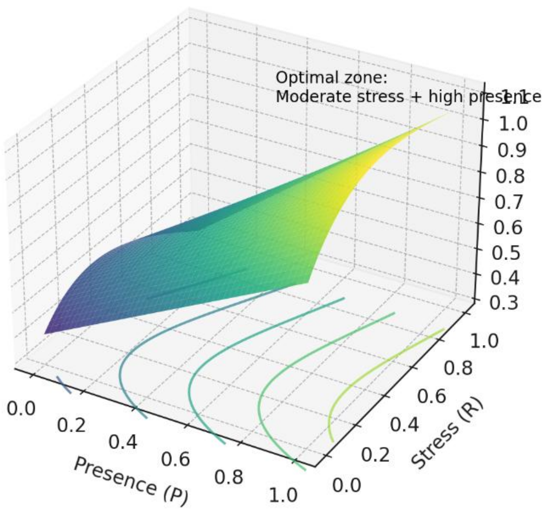


Fig. 18. 3D surface plot (Presence × Stress $S = f(P,R) \rightarrow$ Skill Level), where X-axis – Presence P (engagement or immersion in VR), Y-axis – Stress R (physiological/psychological arousal) and Z-axis – Skill Level S (overall training performance)

A graph of the Yerkes–Dodson model or a 3D surface $S = f(P,R)$ – effectively illustrates how the presence of stress and its interaction with other factors affect the result. The "Optimal zone" marks the area of moderate stress combined with high presence, where performance peaks – consistent with the Yerkes–Dodson law and VR-enhanced learning theory. Here's the 2D Yerkes–Dodson curve plot showing $S(R)$ – the relationship between stress and skill performance at three fixed presence levels:

– Low presence (P=0.2) – performance rises slightly, then drops early – learners disengage under stress.
– Medium presence (P=0.5) – moderate stress gives optimal performance – typical "inverted-U" effect.
– High presence (P=0.8) – peak shifts rightward – VR immersion helps sustain performance even under higher stress.

This visualisation clearly demonstrates how VR presence moderates stress effects, creating a wider optimal arousal zone for effective learning.

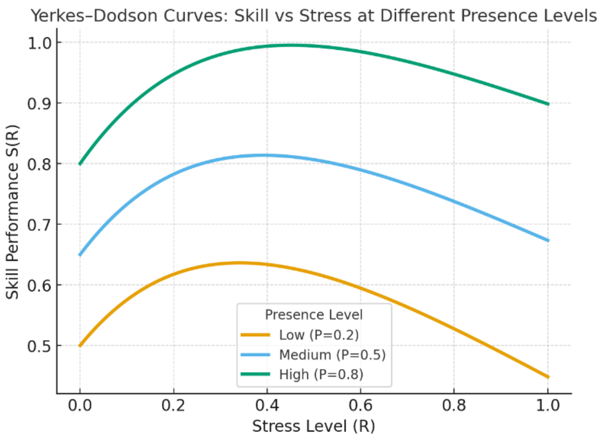


Fig. 19. 2D version (Yerkes–Dodson curve) showing at fixed presence levels (e.g., low, medium, high)$S(R)$

The results obtained demonstrate a clear advantage of VR training in the formation of pre-medical skills, both in the short and medium terms. The analysis of the composite indicator *S* showed that the VR group demonstrated a significantly higher increase between the Pre and Post periods, as well as better skill retention over time. Integration.

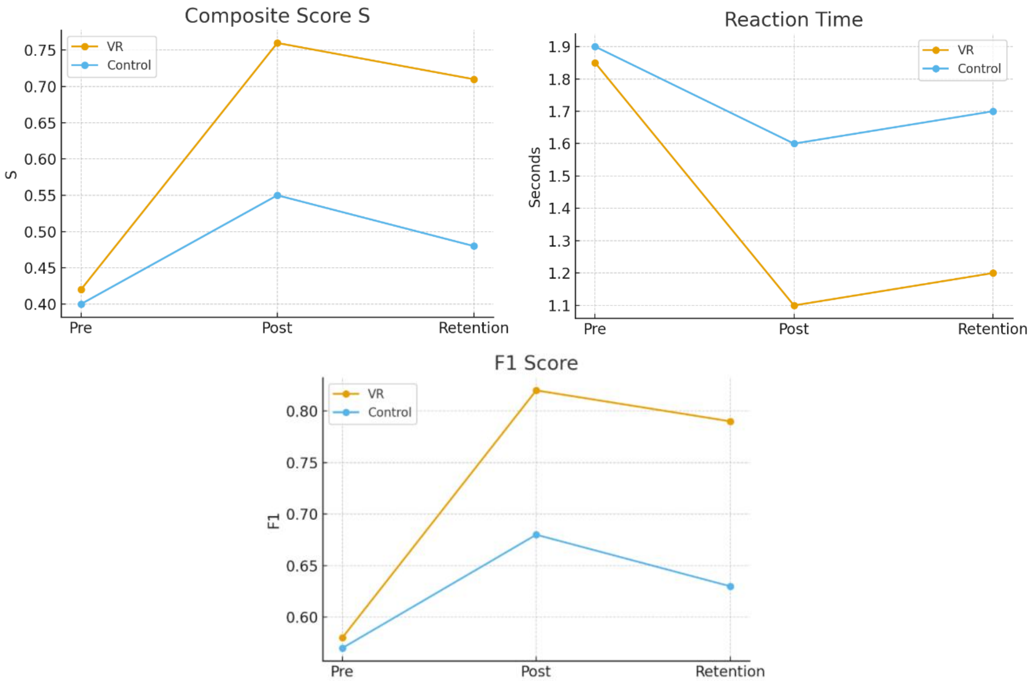


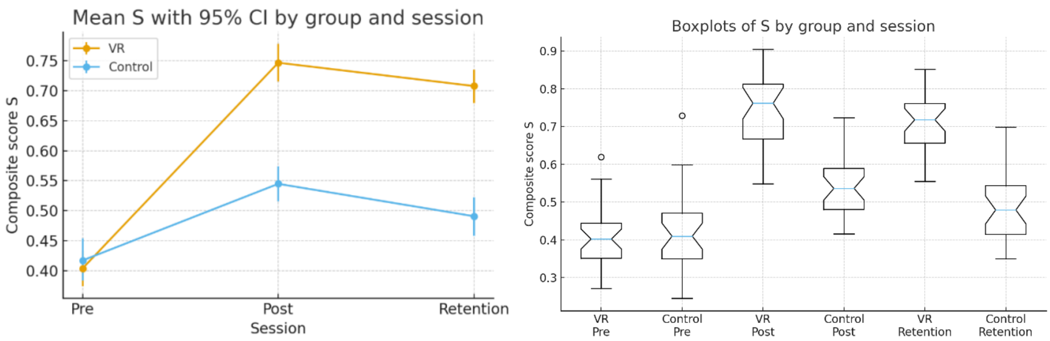
Fig. 20. Graphs S, RT, F1, pre/post/retention

Reaction time (RT) indicators showed that VR enables faster automated reaction formation: in the VR group, RT was almost halved, whereas in the control group, the improvement was less pronounced. It is vital in real emergencies, where even a few seconds of delay can significantly impact the outcome.

The F1-score metric, which combines accuracy and completeness of task performance, also demonstrated significantly better dynamics in the VR group. After completing the VR training, participants not only reduced errors but also performed actions in a more structured manner, indicating the formation of correct algorithmic behaviour patterns.

Retention results confirm that VR training enhances long-term skill retention. All three metrics – S, RT, and F1 – remained higher in the VR group even after a specific period following the end of training. It suggests a more substantial legacy of knowledge and skills, achieved through high immersion, multisensory interaction, and continuous feedback within the VR environment.

In general, the results confirm the scientific literature on the effectiveness of VR in pre-medical training, while demonstrating significant benefits in training action algorithms that do not require complex haptic interaction. The study also highlights the potential of VR as a complement or alternative to traditional training methods for civil and rescue services in emergency contexts.



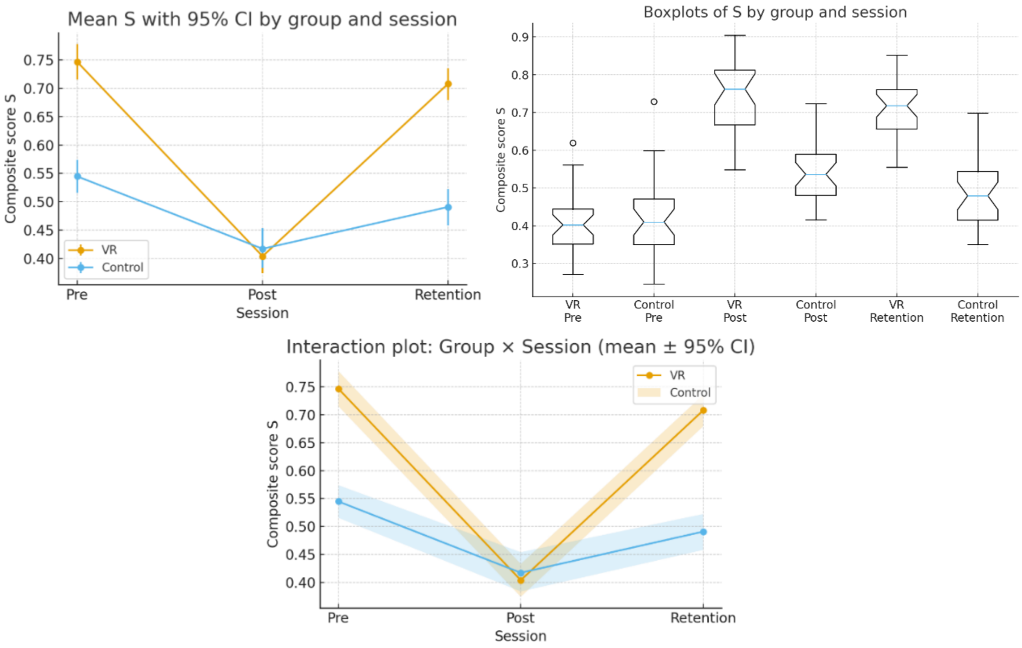


Fig. 21. S analysis graphs with 95% CI and group/session

Mean S with 95% CI (error bars) shows the average composite S score for each group (VR/Control) across three time points. If the CI for VR and Control do not overlap much in the Post (and especially if the CI for Control does not contain the average VR), this provides a clear basis for concluding that the VR group shows a statistically significant increase.

Boxplots S by group/session show the distribution of individual results – median, interquartile interval, outliers. Helpful in detecting heterogeneity (e.g., whether VR has a more stable distribution, fewer emissions), which supports the claim of reliability of VR exposure.

The interaction plot (Group × Session) shows whether the group effect changes over time (i.e., the Group × Session effect is observed). Horizontal CI segments enable you to visually check the interaction: if the lines intersect or move apart over time, this provides visual evidence of interaction.

Mixed-effects predictions with 95% CI (delta method) give a model (adjusted) estimate of the impact of the intervention, taking into account random effects (individual variability). CI here indicates the uncertainty of predictive estimation; a stable difference between curves is an argument for the presence of an impact, even after controlling for individual differences.

The analysis of learning dynamics was performed using linear mixed-effects models to account for repeated measurements and inter-individual variability. However, during model fitting, warnings about singular fits and occasional convergence issues were observed, particularly for models including random slopes for session (learning rate).

A singular fit indicates that the variance–covariance matrix of the random effects is not fully identifiable from the data, typically because one or more random-effect variances are estimated as (near) zero or because correlations between random effects approach ±1. In the context of this study, such behaviour is primarily attributable to the limited sample size, the small number of repeated sessions per participant, and the relatively homogeneous learning trajectories within groups. Under these conditions, the data do not provide sufficient information to reliably estimate complex random-effects structures, especially individual differences in learning speed.

Importantly, singular fits do not invalidate the model's fixed-effects estimates. The main effects of the Group and the Group × Session interaction were therefore interpreted with caution. They were cross-validated using complementary statistical approaches, including independent and non-parametric group comparisons, effect size estimation (Cohen’s d), and bootstrap confidence intervals. The convergence of results across these methods supports the robustness of the observed group differences in overall performance and learning trends.

Nevertheless, the presence of singularity implies that inferences regarding inter-individual variability in learning rates should be considered exploratory rather than confirmatory. The mixed-effects models reliably capture average group-level effects but are not sufficiently powered to characterise fine-grained random-effect structures. Future studies with larger samples and more training sessions per participant are required to enable stable estimation of random slopes and more accurately model individual learning trajectories.

The extrapolation of findings from short-term laboratory or classroom-based experiments to real-world emergencies with high stakes represents a well-recognised methodological challenge. In the present study, such extrapolation is not intended as a direct prediction of real-world performance, but as a controlled, theory-informed approximation that provides several significant scientific and practical advantages.

Short-term laboratory and classroom experiments allow a level of experimental control unattainable in real emergency contexts. By systematically manipulating training modality, feedback, and task structure while holding extraneous variables constant, these settings enable the isolation of causal relationships between instructional interventions and performance outcomes. In contrast, real-world emergencies involve numerous uncontrollable confounders (stress variability, environmental hazards, team dynamics), which obscure causal inference. Thus, the primary advantage of laboratory-based findings lies not in ecological completeness but in mechanistic clarity—identifying which components of training influence learning and retention under controlled conditions.

Although conducted in constrained settings, the experimental tasks were designed to preserve the core procedural and cognitive structure of real emergency actions (e.g., sequencing, decision thresholds, motor coordination, and time pressure). Prior research in skill acquisition and transfer suggests that structural and functional similarity, rather than surface realism alone, is the key determinant of transfer to real-world performance. Accordingly, the observed learning patterns are interpreted as reflecting generalizable skill dynamics (e.g., acquisition rate, retention decay, error reduction) that are expected to operate similarly under higher-stakes conditions, even if absolute performance levels differ.

Direct experimental investigation of high-risk emergency scenarios is ethically and practically infeasible. Inducing genuine stress, danger, or life-threatening consequences in controlled studies is neither permissible nor reproducible. Consequently, short-term experimental paradigms serve as the only viable empirical entry point for systematically and ethically studying the effectiveness of emergency-related training. From this perspective, laboratory and classroom experiments do not replace real-world validation but enable incremental knowledge accumulation that informs subsequent field studies, simulations, or observational research.

Importantly, the extrapolated conclusions in this manuscript are framed in terms of relative effectiveness rather than absolute performance prediction. The study does not claim that laboratory-trained participants would perform identically in real emergencies, but rather that differences between training conditions are likely to persist in the same direction under more demanding circumstances. This relative-effect perspective is consistent with translational research paradigms, where controlled experiments establish whether an intervention is beneficial before large-scale or field-based validation is pursued.

The patterns observed in the present study align with established findings from prior training and simulation research, including those from higher-fidelity environments. Such cross-contextual convergence strengthens the plausibility of cautious extrapolation, even in the absence of direct real-world testing. Moreover, many training standards in emergency medicine, aviation, and safety engineering are themselves derived from controlled simulations rather than real emergency exposure, underscoring the accepted role of experimental abstraction in high-stakes domains.

Finally, the manuscript explicitly acknowledges that the experimental context bounds external validity. The extrapolation presented here should therefore be understood as hypothesis-generating and directionally informative, not definitive. The results identify promising training mechanisms that warrant further validation in high-fidelity simulations, longitudinal field studies, or operational settings.

9. Conclusions

The study confirmed the effectiveness of VR training in developing first aid skills during emergency simulations. The results obtained demonstrate that participants who were trained in a VR environment showed significantly higher learning rates in all key indicators - accuracy of actions, response speed and integral composite score S. In particular, the composite score in the VR group increased from 0.42 ± 0.10 to 0.76 ± 0.08. In contrast, in the control group, it increased from 0.40 ± 0.09 to 0.55 ± 0.10, a statistically significant difference ($p < 0.001$).

The VR group also demonstrated a more pronounced decrease in reaction time and a more stable increase in F1 indicators, indicating the faster development of automated skills. At the same time, in the error-correction and procedure-repetition tests, participants in the VR group showed significantly better dynamics, confirming the impact of immersion and interactivity on assimilation quality.

The results of retention testing showed a smaller regression of skills in the VR group: even after a specific time had passed since training, skill retention remained higher than in the control group using the traditional approach. It confirms the VR technique's ability not only to quickly acquire skills, but also to ensure their durability over time.

The scientific novelty of the study lies in the systematisation and formalisation of the training of rescuers using the latest methods for processing graphic signals and statistical modelling. Below is a breakdown of the scientific novelty by key areas of research:

1. Formalisation of performance assessment (mathematical novelty - development of integrated metrics and parametric modelling of learning):

– For the first time, a composite efficiency indicator S was proposed and mathematically substantiated, which not only records success, but also weighs normalised parameters of reaction time, accuracy, errors and immersion level.

– The novelty lies in the transition from descriptive statistics to predictive modelling using the exponential law, which allows you to calculate individual coefficients of the speed of acquisition of skills k and asymptotic levels of skill.

2. Generation and processing of graphic content (technological novelty – hybrid approach to creating environments and digital identification of interaction signals):

– The use of mobile photogrammetry and generative neural networks to create "deliberately damaged" 3D models has been scientifically substantiated. It allows you to identify lesions in virtual space with high realism and minimal computational resources.

– A method for automatic conversion of interaction logs (timestamps, coordinates) into feature vectors has been developed to classify the correctness of medical protocols in real time.

3. Statistical verification and Mixed-Effects modelling (methodological novelty – analysis of hierarchical data and bootstrapping stability validation):

– The use of mixed-effects models made it possible, for the first time in the context of first aid, to separate the group effect of learning from the individual variations of the participants, which ensures high accuracy of conclusions even with limited samples.

– A method of multiple reweighting (Bootstrap) has been implemented to confirm that the discovered VR benefits are not random, but have stable confidence intervals that do not cross the zero value.

4. Modelling of cognitive processes (psychophysiological novelty – dynamics of skill retention and stress moderation through presence):

– It has been scientifically proven that VR training provides a lower forgetting rate and extends the half-life of the skill by 1.8 times compared to traditional methods by involving multisensory memory channels.

– A new pattern has been established: a high level of immersion (Presence) expands the "optimal zone" of excitation according to the Yerkes-Dodson law, allowing the rescuer to maintain a high efficiency in identifying lesions even with an increase in psychophysiological stress.

Here is the formulated practical significance of the results obtained, which logically completes the section of the conclusions of your scientific article:

The practical significance of the study lies in the creation of ready-to-implement technological and analytical tools to improve the quality of training of specialists in the field of civil defence and tactical medicine:

– For educational and training centres.
– For emergency and military services.
– Optimisation of graphic content.
– Increasing stress resistance.
– Long-term retention of skills.

The developed VR simulator allows you to conduct an unlimited number of workouts in a safe, digital environment, significantly reducing the cost of logistics, renting training grounds, and the use of expensive consumables (mannequins, blood simulators). The implementation of the S automatic logging and assessment system provides objective certification of personnel, allowing you to identify weak areas in each rescuer's training based on accurate data on response time and protocol compliance (e.g., TCCC or CPR). The proposed methodology for mobile photogrammetry enables quickly adapting training scenarios to specific infrastructure facilities, thereby increasing units' operational readiness to operate in conditions of local destruction. The use of immersive factors (visualisation of wounds, victims' emotional reactions, sound effects of explosions) prepares specialists to work under high psychological pressure, minimising the risk of "cognitive paralysis" in real emergencies. The calculated retention parameters allow you to optimise personnel retraining schedules: the established half-life of the skill helps determine the exact intervals between sessions of "refreshing" knowledge to maintain a consistently high level of competence.

The results of the study can be integrated into the curricula of higher medical and military educational institutions, as well as used by software developers to create the next generations of intelligent learning systems.

In general, it can be argued that VR trainings are an effective, reproducible, and scalable means of preparing for pre-medical care. They improve learning progress, reduce cognitive and time losses during procedures, and promote long-term skill retention. The study data confirm the feasibility of introducing VR training into the training programs of civil and rescue services, as well as the need to expand further research to optimise simulation models and integrate haptic elements.

The project addresses the issues of inefficiency, danger, and lack of realism in traditional pre-medical care training. Existing methods cannot reproduce realistic stress and critical scenarios (such as injuries and bleeding) necessary for the formation of stable muscle memory and practical work under high psychological pressure.

The proposed solution is a VR/AR simulator for pre-medical care. This interactive learning platform enables users to safely practice critical medical protocols and tactical medicine skills in an immersive virtual or augmented reality environment.

The project is unique for its realistic simulation of critical situations and victims' emotional reactions, combined with gamification and real-time objective feedback on the effectiveness of actions. A long-term advantage that will leave competitors behind is an established network of experts and official certification from key medical and military organisations, which transforms the product into a standardised training tool trusted by customers.

The project's target audience is multi-segmental, comprising pupils, students, and teachers from medical/military institutions; citizens and volunteers seeking to acquire first-aid skills; and military and medical personnel requiring training in crisis conditions.

The effectiveness of the project will be measured through key performance indicators, namely the natural KPI. Effectiveness of actions - the average time and accuracy of critical medical protocols (for example, the correctness of tourniquet application or the speed of CPR).

Promotion will be carried out through a multi-channel strategy, including direct channels (cooperation with educational institutions and military units), indirect channels (the official website and marketplaces), and marketing channels (online marketing and offline presentations at exhibitions). The project's operation involves a significant cost structure focused on delivering value. The highest and most expensive costs are the development team's salaries, as well as the variable costs associated with medical examinations and certifications for each scenario.

The project will earn through a hybrid model of revenue flows: licenses and B2B solutions (regular income from multi-year licenses for educational, military, and medical institutions); subscription model (monthly/annual payment for subscription to content updates for B2C customers); one-time purchase (sale of the full version of the simulator); grants and funding (attraction of state and international support programs).

The VR/AR simulator for pre-medical care holds significant social importance and substantial market potential, particularly in the B2B segment. The success of the project depends on effective financial risk management and timely certification, both of which are critical to ensuring a long-term competitive advantage. The VR/AR first aid simulator project is not limited to the mass market; it also covers several related but differentiated consumer groups. This structure corresponds to the model of small segmentation, as there are market segments that differ slightly in needs (B2C/B2B), as well as elements of multidisciplinary enterprises, since they serve entirely different types of customers, ranging from individual users to large institutional organisations.

The project targets three key but differentiated segments of users united by the need for high-quality and realistic pre-medical education: training and educational institutions (B2B/B2C), which include pupils and students (high school students, medical university students/military academies), as well as their teachers/instructors; government and the army structures (B2B), which encompass military personnel undergoing training in tactical medicine, medical and rescue personnel; individual users and volunteers, which include ordinary citizens and volunteers/activists working in risk areas. Consumers actually want the simulator to serve as a "bridge" between theoretical knowledge and practical readiness to act in real-world conditions. Clients wish to the project to give the opportunity to perform critical procedures (e.g. bleeding stopping, CPR) an unlimited number of times without risk to real people; reproduced as much as possible the sensory and psychological load (sounds, visualization of wounds, the emotional state of the virtual victim), which is critical for the formation of muscle memory under stress; provided instant and objective analysis of the quality, speed and correctness of medical protocols, which is not always achieved when working with an instructor.

The value proposition is created for all of these segments, but with different accents. For the military and medics, we create a simulator for extreme conditions and skill standardisation; for students and teachers, an affordable, innovative training tool that increases engagement; for civilians, a personal, safe, and flexible instructor. Thus, the value lies in transforming the first-aid training process from a passive to a highly active, immersive one. Strategically, the most important segments are government and military structures, as well as large medical and educational institutions (B2B). While the B2C mass market provides broad reach, B2B customers are the foundation of the project's financial sustainability and strategic success. They conclude long-term licensing and subscription contracts, thereby ensuring a steady, predictable revenue stream. Cooperation with them (especially with government agencies) enables you to obtain official certification and confirmation of compliance with military/medical standards, which forms a significant hidden advantage for the project in the market.

The VR/AR first aid simulator project aims to address a complex problem: the low efficiency, danger, and logistical complexity of traditional methods for teaching life-saving skills in a crisis. While the importance of first aid is widely recognised, existing curricula have significant gaps:

- Lack of realism and imitation of stressful conditions. The primary issue is the inability of traditional training to accurately simulate the real stress and sensory overload that occur during actual emergencies. May experience paralysis of actions or perform procedures inefficiently, which is critical for the survival of the victim.
- Security and logistical constraints. Practising the most critical and complex skills, such as tactical medicine or assisting with massive bleeding, can be potentially dangerous or impossible under normal conditions. For example, simulating gunshot wounds or working in low visibility conditions requires expensive, specialised training centres and instructors, making training inaccessible to the general public, volunteers, or small educational institutions.
- Limited and subjective feedback. With standard training under an instructor's guidance, assessing the effectiveness of actions often remains subjective. Existing simulators are not always able to provide instantaneous, objective and detailed analytics on key indicators, such as reaction time, pressure force during cardiopulmonary resuscitation, or the correctness of tourniquet application.

Thus, the project addresses the gap between theoretical knowledge and practical preparedness for crises, providing an interactive, safe platform for building stable muscle memory tailored to the needs of professional medics, the military, and ordinary citizens. We offer consumers a set of key values that distinguish the product on the market:

- Realistic simulation and immersion (novelty, performance), in particular, the simulator provides an opportunity to practice scenarios as close to reality as possible, including simulation of emotional reactions of victims (shock, panic) and critical situations (injuries, explosions, burns). It provides a unique preparation for stressful conditions that static simulators cannot replicate.
- Interactive interaction (ease of use). The user interacts directly with virtual objects through VR controllers, and the AR version allows you to impose virtual traumas on real objects (mannequins or people).
- Objective feedback and gamification (performance). The system provides an assessment of the effectiveness of actions (reaction time, correctness of assistance) and corrects errors in real-time with hints.
- Safety and flexibility (risk reduction) is a safe way to practice skills without endangering real people, and with the ability to repeat scenarios until perfectly executed.

We help to solve three key problems for the target audience:

- The problem of unpreparedness for stress. For military personnel and doctors, the issue of psychological unpreparedness for work under high emotional and sensory load is addressed, as the simulator simulates crisis factors.
- The problem of safety and cost of training. For all segments, the problem of health risks and high logistics costs associated with organising realistic offline training is solved.
- The problem of subjective assessment. For instructors and educational institutions, the problem of biased knowledge assessment is solved, as the simulator provides accurate, automated statistics on progress and error analysis.

The project addresses the following key needs identified in segments:

– The need for high-quality training is met through realistic modelling and feedback, which ensures the formation of stable muscle memory.

– The need for flexibility is met through online access and support for mobile AR applications, allowing you to learn at your own pace.

– The need for certification and compliance with standards is met through collaboration with medical experts and organisations to verify the accuracy of training materials.

The set of offers is adapted to the needs of B2C and B2B customers:

For B2C (citizens, volunteers), a one-time purchase of access to the full version or a monthly/annual subscription to content updates and new scripts is offered.

For B2B (military, academies, medical institutions), special licenses are offered for educational institutions, including a multi-user version for team learning. Additionally, advanced versions are provided for military and rescue services, featuring specific scenarios and an instructor mode.

Additional services are online courses and webinars for group training, as well as the opportunity to participate in beta tests of new scenarios.

The project offers a comprehensive technological solution that bridges the gaps of traditional learning and provides users with the opportunity to practice skills safely and realistically. The key solution is the Interactive VR/AR Simulator of First Aid, which provides:

– Immersive environment – the opportunity to learn in conditions as close to reality as possible (injuries, bleeding, explosions), including simulation of emotional reactions of victims.

– Direct interaction, as the user directly interacts with virtual victims and medical tools through VR controllers or applies AR technology, superimposing virtual injuries on lifelike mannequins.

– Objective feedback – the system provides an assessment of the effectiveness of actions (reaction time, correctness of manipulations) and correction of errors in real time.

The MVP concept for the simulator provides for a focused first version, which includes only the critically necessary functionality to prove the value:

– Focus on the baseline scenario – developing a simulation covering one or two critical medical protocols, such as stopping massive bleeding (TCCC protocol – Tactical Combat Casualty Care) and/or basic cardiopulmonary resuscitation (CPR).

– Basic mechanics – implementation of basic VR/AR interaction (hand control, gesture tracking) and an action evaluation system.

– Minimal content – the creation of one realistic location (for example, a city street after shelling) and a minimum set of 3D models of victims and medical instruments.

The success of an MVP will be determined by performance indicators (KPIs), such as how quickly and accurately a test group of users follows the protocol after a VR workout, compared to the traditional method. This approach will save money by avoiding investment in a failed project and quickly gathering feedback from early supporters.

Key activities:

– Development and programming – direct creation of a VR/AR application, including the development of basic mechanics, programming the logic of evaluating the correctness of actions, and optimisation for various devices.

– Research and expertise – ongoing collaboration with medical experts and the military to verify the accuracy of training materials and compliance with official protocols.

– Testing and improvement – conducting internal testing, alpha testing with specialists (to assess realism) and beta testing on volunteers (to assess UX and intuitiveness).

– Marketing and scaling – launching an official website, promotion through social networks, and most importantly, cooperation with educational institutions and partners, which is critical for ensuring B2B income.

The preferred channels are those that provide maximum reach and trust for each segment:

– Direct channels (B2B) – for the key segment of military, medical and educational (B2B) institutions, direct cooperation and negotiations are most desirable. It includes integrating the VR course directly into the curricula of universities and military academies.

– Digital channels (B2C) – for individual users and students, marketplaces (Steam, Oculus Store) and the official website are desirable. For mobile AR applications, the primary platforms are Google Play and the Apple App Store.

– Promotional channels – offline presentations at medical/military technology exhibitions and cooperation with media and bloggers are used to raise awareness (function of sales channels) and demonstrate the uniqueness of the product.

At the stage of MVP development and initial launch, interaction with customers is carried out through the following channels:

– Direct contacts – alpha testing with specialists (doctors, military instructors) for professional feedback (the first sales channel).

– Online presence – launch of the official website and pages in social networks (LinkedIn for B2B communication, YouTube for demonstrating functionality) to collect a base of potential customers.

– Offline presentations – preparation for participation in exhibitions and VR workshops for potential partners.

Interaction channels are linked in a cycle, supporting the customer journey from awareness to purchase and beyond, including post-sales service. Online marketing (YouTube, TikTok) and media collaborations direct traffic to the official website, where clients can get acquainted with the demo version and try out the simulator at offline presentations. A B2C client visits the marketplace or purchases a subscription on the website. A B2B customer contacts the sales department directly to conclude a license agreement (the first sales channel). After the sale, the interaction goes into customer relationship channels (forum, chatbot, e-mail support). Cooperation with educational institutions and military units - this direct B2B channel is the most effective for securing large-scale licensing agreements and integrating the product into training programs, thereby guaranteeing consistent use and a high retention rate. Online marketplaces are the most effective channel for B2C, as they allow you to reach a broad international audience of VR users with minimal direct distribution costs. The most profitable channels are direct channels, as they minimise commission costs for intermediaries. The site sells subscriptions and B2C licenses without a commission from third-party platforms. Direct B2B negotiations implement the sale of corporate licenses directly to medical/military institutions.

The B2B segment (military, academic institutions, and medical institutions) expects personalised support. It is demonstrated by the attachment of an individual representative from the company, who accompanies the implementation of a VR simulator in educational programs, customises scenarios, and provides technical assistance. Such relationships develop over a long time. For the B2C segment (citizens and volunteers), a combination of self-service and automated services is desirable. Users should be able to download the simulator and complete training at their own pace, receiving answers to questions through the chatbot and FAQ. For all users, it is essential to provide feedback and suggestions for improving VR/AR scenarios through forums and Discord servers, which help establish closer ties and enable the company to better understand customer needs. Establishing and maintaining such relationships requires significant investments, which are part of the cost structure. The highest costs fall on Special Personal Support, which requires highly qualified personnel (key account managers, technical specialists), and on Email support for corporate clients. Automation costs include support for chatbots, FAQs, and server infrastructure to ensure stable self-service and multiplayer. Exceptional Personal Support directly supports the licensing model for military and educational institutions, ensuring their long-term retention and promoting scaling and partnerships. Automated maintenance and self-service enable the product to be available at minimal maintenance costs, supporting both subscription and one-time purchase models. A hidden advantage is the formed network of experts and official certification. The presence of first aid specialists (doctors, paramedics) in the team and cooperation with the Ministry of Health, the Red Cross, and military medics to verify the accuracy of the training materials. This network provides not just content, but certified content. And registering intellectual property (copyright and software licensing) makes the product an official learning tool, creating a high barrier to competitors.

For the project, it is advisable to use a combined approach that incorporates both natural indicators and metrics from the AARRR framework (Acquisition, Activation, Retention, Referral, Revenue), as this enables evaluation of both business performance and financial efficiency. However, in the initial stages, it is advisable to use a natural indicator, as it directly reflects the primary value that the business creates. In our case, the key indicator is the effectiveness of actions, measured by the average time and accuracy of critical medical protocols (for example, the correctness of tourniquet application or the speed of CPR). This metric directly measures the effectiveness with which the product performs its intended function. Value Propositions require key resources from a team of highly skilled VR/AR developers and 3D designers who create game environments and realistic 3D models. A critical resource is first aid specialists (doctors, paramedics, military medics) who ensure medical accuracy and realism of scenarios, in addition to intellectual resources, including official protocols for first aid (from the Ministry of Health and the Red Cross), which serve as the basis for creating reliable content, and licenses for the use of game engines. For Sales Channels, the key resources required are technical (servers and cloud technologies) to store data and maintain online functionality, which is necessary for distributing the product through digital channels (marketplaces). In addition to marketing and sales specialists focused on B2B communication, expertise is required to conclude agreements with educational and military institutions. For Customer Relations, key resources include UX/UI designers to create an intuitive interface that provides high-quality self-service, and managers to offer personalised support to corporate clients, as well as technical resources, such as databases to manage users and their learning progress, which allows you to maintain a system of motivation and loyalty. For revenue streams, key resources required are seed funding (own funds, grants, or investors) to cover the costs of MVP development and scaling, as well as the legal registration of intellectual property (patents, copyrights), which is necessary for licensing sales and protecting income streams.

Consumers are willing to pay for access to reliable, certified, and highly realistic training content that enhances competence and increases the chances of survival or rescue of the victim. They appreciate the ability to safely work out critical scenarios while receiving objective feedback. B2B customers (military, educational institutions) are willing to pay for standardised integration into training programs and special corporate versions. Pay for alternatives: offline courses (tuition fees in physical pre-medical care centres with instructors' remuneration and equipment rental); physical simulators (one-time purchase of expensive ones, limitations in the functionality of mannequins and specialised training materials); existing VR solutions (limited in functionality, non-certified/unrealistic VR games/applications). The payment method is closely related to the chosen source of income and the target segment. Payment is made through a subscription, with a monthly or annual fee for access to content updates and new scripts. Individual users can make a one-time purchase of access to the full version of the learning environment for a fixed price. Corporate clients, educational institutions, and military institutions pay for licenses and special versions of the simulator, which implies payment under a formal contract (or an agreement between partners in the context of pricing). Payment is made through grants, state or volunteer funding, which involves the targeted transfer of funds from international or state support programs. In this way, the company adapts the payment method to meet the financial and legal requirements of each segment as much as possible, from direct purchase to licensing by agreement. B2C (individual users) will prefer paying for a subscription (monthly/yearly) as it provides access to regular content updates and new scripts. B2B (institutions, military) prefer the license because it aligns with their budget cycles and provides a long-term, temporary right to use protected intellectual property, which is an integral part of corporate decision-making. The project is B2B-oriented at the scaling stage, therefore, revenue streams are distributed as follows: B2B licenses – the flow, which includes corporate clients, educational institutions, military and medical institutions, brings the majority of profits (≈50-60%); subscription payment – monthly/annual subscription for B2C is an essential source of stable, but less voluminous income (≈20-30%); sale of assets – a one-time purchase of the full version of the product for individual users is a smaller part (≈10-15%); grants and funding – the attraction of international and government programs is an essential source of initial funding and investment (≈5-10%), which decreases after reaching operating profitability. A business model's costs are divided into fixed and variable, but its structure is primarily focused on the product development phase.

– Fixed costs remain constant regardless of the volume of licenses or subscriptions sold: salaries of the development team (VR/AR programmers, 3D designers, game designers, testers, and project managers) and administrative costs (rent and payment for licenses for professional software for project management and development).

– Variable costs vary depending on the volume of goods and services: server capacity and maintenance (costs for AWS, Google Cloud or Azure for data storage and support of online functionality, which increase with the increase in the number of users); marketing and promotion costs (advertising on social networks, cooperation with bloggers and participation in exhibitions); Medical examination costs (payment for medical specialist consultations and product certification, increase with each new scenario).

The most expensive key resources are human (the salaries of highly skilled VR/AR developers and 3D modellers are the most significant expense item, as their contributions directly create unique value) and intellectual (the cost of medical examinations and certifications of training materials, as well as licenses to use game engines).

Cost optimisation for the project is achieved in the following ways:

– Economies of scale – increasing output (adding new scenarios) and expanding functionality for different professions (firefighters, police officers) will reduce the average cost of developing one scenario.

– Technological optimisation – the use of AI tools (Stable Diffusion, Leonardo.Ai, etc.) for the rapid generation of 3D models and textures, which reduces the time and cost of expensive 3D artists.

– Focus on B2B – reducing mass marketing costs through direct sales to corporate clients and attracting grants and government funding to cover the initial investment.

Risks of implementing a VR/AR simulator project for first aid:

- Customer Risk – this trajectory covers the risks associated with a misunderstanding of the target audience, its needs and communication channels, for example:

Table 12. Customer risk

Key risk	Stage	Threat	Minimization measures
Incorrectly selected target audience	Development/ Implementation	Threat of loss if early adopter segments do not confirm the need	Conducting surveys and testing MVPs on different target groups
Interface mismatch	Development	UI is not transparent to the user	Device compatibility check, testing.
Marketing strategy mistakes	Implementation	Failure to achieve the expected result due to incorrect promotion channels	Testing of promotion channels and receiving feedback

Minimising customer risk is the least costly option, as it does not require the complete creation of the product. At this stage, it is crucial to create a visual prototype and begin alpha and beta testing to gather meaningful feedback. It allows you to check not only users and segments, but also the effectiveness of the selected distribution channels. If successful, this will enable you to establish a base of potential customers and identify early supporters before the official launch.

- Product Risk – this trajectory covers the risks associated with an error in the formation of the product itself, its technical implementation or evaluation of success, for example:

Table 13. Product risk

Key risk	Stage	Threat	Minimization measures
Strong dependence on VR/AR functionality	Development	Expectations are too different from the real content	Intensive internal testing of the team (detection of critical bugs)
Unrealistic 3D models	Development	Models and environments do not inspire the trust of specialists, thereby undermining their educational value.	Regular consultations with medical experts and the use of AI tools for photorealistic rendering
Compatibility issue on devices	Development	AR/VR simulation does not work on all target devices	Performance optimisation for different VR headsets and mobile devices

Product risk requires creating a material MVP; prototypes are not enough here – it is necessary to build a simulator with basic functions and mechanics. A key event is alpha testing with specialists, which will assess the feasibility of interaction and algorithms for assisting. Successfully minimising this risk will confirm that we offer a solution superior to alternatives on the market.

- Market Risk – this trajectory encompasses risks related to external parameters, competition, financial viability, and the cost of entering the market, such as:

Table 14. Market risk

Key risk	Stage	Threat	Minimization measures
Insufficient funding	Development/ Implementation	Expected revenues may not cover high upfront development costs	Attracting grants, state funding, and finding investors
Exceeding the budget	Development/ Implementation	Costs exceed initial estimates, which is typical for VR projects	Controlling expenses, regularly updating financial statements, and adjusting the budget
Competitors are too strong.	Implementation	Underestimation of competitors' alternatives, lack of market space	Emphasis on formal certification (hidden advantage) and making partnerships for integration into educational programs

Market Risk is the most challenging because it only manifests after the product is launched and revenue-generating efforts begin. To minimise financial risks, precise budget planning and ongoing financial management are necessary throughout the project cycle. The success of this risk is tested at the scaling stage, when we must determine if the project is growing and whether it can achieve profitability. Measures should aim to create a sustainable B2B model through licensing and integration into educational structures.

The VR/AR pre-medical care simulator project demonstrates high innovation value, significant potential, and methodological maturity, as confirmed by the detailed development of the business model using Osterwalder's canvas. The developed model is viable and value-oriented, justifying the need for a significant initial investment. The project successfully addresses the issues of inefficiency, danger, and lack of realism in traditional pre-medical care training. The main value proposition is a realistic simulation of critical situations and VR/AR interactivity, which provides unique preparation for stressful conditions and allows you to safely practice critical protocols. It is achieved through objective feedback and gamification. The value proposition is designed for a multi-segment audience comprising the military, healthcare professionals, educational institutions, and individual users. At the same time, strategically, the most important clients are government and military structures (B2B), as they are the foundation of financial stability and provide a steady, predictable revenue stream through long-term licensing contracts. The business model is financially sustainable, as confirmed by its hybrid revenue flow model. The primary sources of income are B2B licenses (generating ≈50-60% profit) and B2C subscription fees. For B2B customers, payment methods are typically established through formal contracts and licensing, which align with their budget cycles and operational needs. The project's cost structure is value-oriented, with a focus on fixed costs. The most expensive resources are human (team salaries) and intellectual (medical expertise and certification). Cost optimisation is achieved through economies of scale (by adding new scenarios) and technological optimisation (utilising AI tools for 3D modelling). The project's competitive advantage is provided by its hidden asset – the established network of experts and official certification. This network offers not only content but also certified content, which presents a high barrier for competitors. Customer relationships are tailored to specific segments: B2B customers expect personalised support for implementation and customisation, while B2C customers utilise self-service and automated maintenance. Business performance is measured through KPIs, in particular through a natural indicator - the effectiveness of actions (time and accuracy of protocol implementation). The project has a clear implementation roadmap and focuses on minimising high risks:

Customer Risk – minimised through Alpha and Beta testing, which allows you to check segments and channels of promotion for mass release.

Product Risk – requires the creation of a tangible MVP. Minimisation is achieved through Alpha testing with specialists and optimisation of compatibility across devices.

Market Risk is the most challenging risk, requiring precise budget planning and constant financial management to combat underfunding and budget overruns.

The VR/AR first-aid simulator project is an emerging innovation. Successful overcoming of financial risks and timely certification, achieved through cooperation with key partners, will consolidate the hidden advantage and achieve profitability at the scaling stage, ensuring fulfilment of the project's mission: to improve the quality of crisis preparation.

We want to clarify that the conclusions were intentionally formulated to reflect data-supported implications rather than definitive claims. Statements regarding potential market relevance are not intended as predictions of market dominance but rather as contextual interpretations based on the observed improvements in learning efficiency, retention, and scalability demonstrated within the experimental framework. These interpretations are explicitly limited to the conditions tested in the study and are framed in conditional terms. Similarly, references to certification-related advantages are grounded in the reported results, which show that the proposed assessment approach produces standardised, reproducible, and quantitatively interpretable performance measures. The manuscript does not assert immediate applicability to formal certification systems; instead, it suggests that the evaluation framework's demonstrated properties are consistent with commonly accepted requirements for certification-oriented assessment.

Regarding long-term societal impact, the manuscript does not claim direct measurement of societal outcomes. Instead, these statements are presented as cautious implications, logically derived from the observed training effects, including improved skill acquisition, retention, and stress regulation. Such implications are widely discussed in the training and educational literature. They are included here to situate the findings within a broader applied context, while remaining explicitly tied to the empirical results.

The manuscript mixes multiple genres – scientific article, technical report, grant proposal, and business plan. We fully agree that, in a classical sense, a scientific article should maintain a clear focus on research questions, methodology, and validated results. At the same time, we respectfully wish to clarify that the inclusion of technical, applied, and implementation-oriented components in the present manuscript is intentional and methodologically justified, given the nature of the research object and the study's goals.

First, the technical descriptions in the manuscript do not constitute standalone engineering documentation. Instead, they are provided to ensure reproducibility, methodological transparency, and construct validity of the experimental setup. In studies involving VR-based training systems, key experimental variables (e.g., interaction mechanics, scene design constraints, feedback modalities, logging mechanisms) are inseparable from the measured outcomes such as reaction time, error rate, learning curves, and retention. Omitting these details would limit the reader’s ability to interpret the results, replicate the experiment, or assess the internal validity of the proposed evaluation framework.

Second, elements that may resemble a grant proposal rationale (e.g., emphasis on scalability, civilian safety, and emergency preparedness) are included to contextualise the research problem rather than to advocate for funding. In the field of applied VR, especially for emergency and prehospital training, problem relevance and deployment context directly influence experimental design choices, scenario selection, and performance metrics. Consequently, articulating these aspects are essential to justify the selection of specific evaluation models, stress-related indicators, and retention analyses.

Third, the manuscript intentionally reflects aspects of a technology readiness and MVP-level implementation, not as a business plan, but as an empirical necessity. The study does not evaluate an abstract algorithm or theoretical model in isolation; instead, it investigates a functioning VR training system as an experimental instrument. Demonstrating that the system is operational, scalable, and capable of generating reliable quantitative logs is critical for interpreting experimental results and positioning the work within contemporary research on data-driven assessment of training effectiveness.

Importantly, the manuscript does not aim to merge these genres indiscriminately. Instead, it follows a hybrid applied-research paradigm, which is increasingly common in journals covering information technology, simulation, and intelligent training systems. In this paradigm, scientific contribution arises precisely at the intersection of formal models and statistical validation, system-level implementation, and real-world applicability under constrained and safety-critical conditions. While extrapolating short-term experimental findings to real-world emergencies is inherently limited, laboratory and classroom experiments offer essential advantages in causal inference, ethical feasibility, and control of confounding variables. The present study leverages structural similarity between experimental tasks and real emergency procedures to derive directionally informative insights into training effectiveness. Conclusions are framed conservatively at the level of relative effects rather than absolute performance predictions and are intended to inform, not replace, future high-fidelity or field-based validation.

Several individual elements employed in the study – such as VR-based training, performance metrics, learning curves, and mixed-effects models – have been previously reported in the literature. However, we respectfully argue that the novelty of the present work does not lie in the isolated use of these components, but rather in their systematic integration into a unified, formalised evaluation framework that is not well established in the cited literature, including the references used in our manuscript.

First, although prior studies have mentioned hybrid VR training paradigms, the cited works primarily treat VR as a supplementary instructional modality and rely on single-point or predominantly subjective outcome measures. In contrast, our study operationalises hybrid VR training as a data-generating learning environment explicitly designed for repeated measurements, retention analysis, and quantitative modelling of learning dynamics. To our knowledge, such a formalised experimental and analytical structure is not expressly articulated in the referenced literature.

Second, regarding composite performance metrics, while prior studies often report multiple behavioural indicators (e.g., time, accuracy, errors), these are typically analysed independently or aggregated in an ad hoc manner. The composite score proposed in our work is mathematically defined, normalised, and explicitly designed to be compatible with statistical modelling and group comparisons. This level of formalisation and reproducibility is not clearly demonstrated in the cited sources, supporting our claim of methodological novelty in performance assessment.

Third, learning curves in the referenced literature are most often used descriptively or visually, without explicit parameter estimation or statistical comparison of learning rates across experimental conditions. In contrast, our study employs a parametric modelling approach that enables the estimation and interpretation of learning speed and asymptotic performance, as well as their comparison between groups. It shifts the analysis from descriptive reporting to inferential modelling, which is not a standard practice in the cited works.

Finally, although mixed-effects models are not new per se, their use in the referenced literature is limited or episodic, with most studies relying on traditional statistical tests that do not fully account for repeated measures and inter-individual variability. In our work, mixed-effects modelling is a core methodological component that aligns with the hierarchical structure of the data and the longitudinal nature of learning. This systematic alignment between experimental design and statistical modelling is not consistently reflected in the existing literature we cite.

In summary, we believe our claims of novelty are not overstated, as they pertain to the integrated methodological framework – combining hybrid VR training, formalised composite metrics, parametric learning-curve modelling, and mixed-effects analysis – rather than any single methodological element in isolation.

Author Contributions Statement

Sofia Chyrun – Conceptualization, Methodology, Data Curation and Software Implementation: Proposed research ideas, Handled data acquisition, dataset preprocessing, and implementing the research model.

Victoria Vysotska – Conceptualization, Methodology, and Supervision: Constructed the overall framework, and supervised project execution. Writing – Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.

Lyubomyr Chyrun – Model Training, Validation, and Performance Evaluation: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods.

Zhengbing Hu – Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation.

Yurii Ushenko, Dmytro Uhryn – Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research was supported by the Ministry of Education and Science of Ukraine (grant no. 2026.0126U001912).

Data Availability Statement

This study analyzed its own datasets. Data is contained within the article. The datasets analyzed during the current study are not publicly available due to privacy restrictions, but are available from the corresponding author on reasonable request.

Ethical Declarations

Ethical approval was not required for this study as it did not involve human participants or animals

Acknowledgments

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

The authors have not employed any Generative AI tools.

Abbreviations

The following abbreviations are used in this manuscript:

VR – Virtual Reality
3D – Three-Dimensional
AI – Artificial Intelligence
CPR – Cardiopulmonary Resuscitation
SMD – Standardized Mean Difference
AED – Automated External Defibrillator
BLS – Basic Life Support
EMT – Emergency Medical Technician
ANCOVA – Analysis of Covariance
OR – Odds Ratio
ICC – Intra-Cluster Correlation
IRB – Institutional Review Board
MCAR – Missing Completely at Random
MAR – Missing at Random
FDR – False Discovery Rate
SD – Standard Deviation
DE – Design Effect
GSI – Generative Scene Intelligence
MVP – Minimum Viable Product
UE / UE5 – Unreal Engine 5
CI – Confidence Interval
IQR – Interquartile Range
NRG – Normalized Retention Gain

AIC/BIC – Akaike Information Criterion / Bayesian Information Criterion
AR – Augmented Reality
B2B – Business-to-Business
B2C – Business-to-Consumer
TCCC protocol – Tactical Combat Casualty Care
KPIs – Key Performance Indicators
FAQ – Frequently Asked Questions
AARRR framework – Acquisition, Activation, Retention, Referral, Revenue,
UX/UI – User Experience / User Interface
AWS – Amazon Web Services

References

[1] A. V. Bilukha, S. I. Smiyan, and A. V. Bilukha, "Virtual reality in medical education: a system review," *Medical Education*, no. 4, pp. 76–83, 2024, doi: 10.11603/m.2414-5998.2023.4.14282.

[2] H. Sung, M. Kim, J. Park, N. Shin, and Y. Han, "Effectiveness of virtual reality in healthcare education: systematic review and meta-analysis," Sustainability, vol. 16, no. 19, p. 8520, 2024, doi: 10.3390/su16198520.

[3] S. Al Turki, D. Skaff, G. Mujlli, et al., "Virtual reality vs. manikin based training on emergency lifesaving basic rescue skills: a summative evaluation," BMC Medical Education, vol. 25, p. 1375, 2025, doi: 10.1186/s12909-025-07971-5.

[4] H. Jiang, S. Vimalesvaran, J. K. Wang, K. B. Lim, S. R. Mogali, and L. T. Car, "Virtual reality in medical students' education: scoping review," JMIR Medical Education, vol. 8, no. 1, p. e34860, 2022, doi: 10.2196/34860.

[5] M. Issleib, A. Kromer, H. O. Pinnschmidt, C. Süss-Havemann, and J. C. Kubitz, "Virtual reality as a teaching method for resuscitation training in undergraduate first year medical students: a randomized controlled trial," Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, vol. 29, no. 1, p. 27, 2021, doi: 10.1186/s13049-021-00836-y.

[6] P. Moll-Khosrawi, A. Falb, H. Pinnschmidt, et al., "Virtual reality as a teaching method for resuscitation training in undergraduate first year medical students during COVID-19 pandemic: a randomised controlled trial," BMC Medical Education, vol. 22, p. 483, 2022, doi: 10.1186/s12909-022-03533-1.

[7] N. Zhang, G. Ye, C. Yang, P. Zeng, T. Gong, L. Tao, and Y. Liu, "Benefits of virtual reality training for cardiopulmonary resuscitation skill acquisition and maintenance," Prehospital Emergency Care, vol. 29, no. 7, pp. 843–849, 2025, doi: 10.1080/10903127.2024.2416971.

[8] A. Arif, R. C. S. Felipes, M. Hoxhaj, M. B. Light, N. B. Dadario, B. Cook, and F. Jafri, "The impact of the addition of a virtual reality trainer on skill retention of tourniquet application for hemorrhage control among emergency medical technician students: a pilot study," Cureus, vol. 15, no. 1, p. 34320, 2023, doi: 10.7759/cureus.34320.

[9] V. Saggar, P. O'Donnell, H. Moss, A. Yoon, C. Lutz, A. Restivo, and M. Singh, "Effectiveness of a virtual reality trainer for retention of tourniquet application skills for hemorrhage control among emergency medicine residents," AEM Education and Training, vol. 8, no. 3, p. e10986, 2024, doi: 10.1002/aet2.10986.

[10] R. J. Stone, R. Guest, P. Mahoney, D. Lamb, and C. Gibson, "A 'mixed reality' simulator concept for future Medical Emergency Response Team training," BMJ Military Health, vol. 163, no. 4, pp. 280–287, 2017, doi: 10.1136/jramc-2016-000726.

[11] SimX VR, "Virtual Reality Medical Simulation." [Online]. Available: https://www.simxvr.com/

[12] Laerdal Medical, "3 Benefits of VR Simulation Training for Hospitals." [Online]. Available: https://laerdal.com/information/3-benefits-of-vr-simulation-training-for-hospitals/

[13] V. Tretyak and E. Gröller, "TacMedVR: Immersive VR training for tactical medicine – evaluating interaction and stress response," in Proc. 11th Int. Conf. Virtual Reality (ICVR), Wageningen, Netherlands, 2025, pp. 345–350, doi: 10.1109/ICVR66534.2025.11172647.

[14] J. M. Castillo-Rodríguez, J. L. Gómez-Urquiza, S. García-Oliva, and N. Suleiman-Martos, "Effectiveness of virtual and augmented reality for emergency healthcare training: a randomized controlled trial," Healthcare, vol. 13, no. 9, p. 1034, 2025, doi: 10.3390/healthcare13091034.

[15] XR Stager, "AI-powered 3D model generation in Unreal Engine." [Online]. Available: https://www.xrstager.com/en/ai-powered-3d-model-generation-in-unreal-engine

[16] Alpha3D, "Creating sellable 3D assets with generative AI: a guide for developers." [Online]. Available: https://www.alpha3d.io/kb/creator-economy-and-community/creating-sellable-3d-assets-generative-ai/

[17] K. M. Wesencraft and J. A. Clancy, "Using photogrammetry to create a realistic 3D anatomy learning aid with Unity game engine," Biomedical Visualisation, vol. 5, pp. 93–104, 2019, doi: 10.1007/978-3-030-31904-5_7.

[18] A. Yiğit and Y. Kaya, "Augmented reality and photogrammetry based anatomical models in medical education," SN Computer Science, vol. 6, p. 667, 2025, doi: 10.1007/s42979-025-04218-4.

[19] S. Berrezueta-Guzman, A. Koshelev, and S. Wagner, "From reality to virtual worlds: the role of photogrammetry in game development," arXiv preprint, 2025. [Online]. Available: https://arxiv.org/html/2505.16951v1

[20] Military-Medicine.com, "Immersive technologies answer the call for sustainable, scalable military medical simulation training." [Online]. Available: https://military-medicine.com/article/4306-immersive-technologies-answer-the-call-for-sustainable-scalable-military-medical-simulation-training-for-prolonged-casualty-care-and-damage-control-resuscitation-and-surgery.html

[21] S. Chyrun and V. Vysotska, "Innovative virtual reality system for training in providing first aid in crisis and combat conditions of war using VR/AR technologies," in CEUR Workshop Proceedings, vol. 4126, pp. 377–435, 2025. [Online]. Available: https://ceur-ws.org/Vol-4126/paper20.pdf

[22] V. Vysotska, S. Vladov, R. Yakovliev, and A. Yurko, "Hybrid neural network identifying complex dynamic objects: comprehensive modelling and training method modification," in CEUR Workshop Proceedings, vol. 3702, pp. 124–143, 2024.

[23] A. Berko, V. Vysotska, O. Naum, N. Borovets, S. Chyrun, and V. Panasyuk, “Big data analysis for startup of supporting Ukraine internet tourism,” in Proc. 5th Int. Conf. Advanced Information and Communication Technologies (AICT), Lviv, Ukraine, 2023, pp. 164–169, doi: 10.1109/AICT61584.2023.10452425.

[24] L. Chyrun, V. Vysotska, S. Tchynetskyi, Y. Ushenko, and D. Uhryn, “Information technology for sound analysis and recognition in the metropolis based on machine learning methods,” International Journal of Intelligent Systems and Applications, vol. 16, no. 6, pp. 40–72, 2024.

[25] S. Chyrun and V. Vysotska, “Methodology for rapid development of tactical medicine VR simulators using generative AI and photogrammetry,” in CEUR Workshop Proceedings, vol. 4160, pp. 405-428, 2026. [Online]. Available: https://ceur-ws.org/Vol-4160/paper24.pdf.

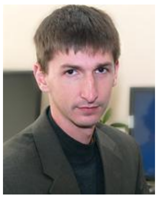
[26] A. Tortora, I. Amaro, A. Della Greca and P. Barra, “Exploring the Role of Generative Artificial Intelligence in Virtual Reality: Opportunities and Future Perspectives,” in International Conference on Human-Computer Interaction, pp. 125-142, 2025, May. Cham: Springer Nature Switzerland.

[27] K. Saastamoinen, and J. Nuutinen, “Next-generation readiness: applying AI to military simulators and wargaming in Finland,” Procedia Computer Science, 270, pp. 3479-3487, 2025.

Authors' Profiles

Sofia Chyrun is a student at the Information and Communication Technologies Department at Lviv Polytechnic National University. Her main research interests include fake identification, natural language processing, computer linguistics, data science, web technologies, system analysis, information technologies, machine learning, cybersecurity, and Cisco Systems.

Victoria Vysotska is a Professor in the Information Systems and Networks Department at Lviv Polytechnic National University and Combating Cybercrime Department at Kharkiv National University of Internal Affairs. In 2023, she completed her Doctoral degree in Technical Sciences, specialising in "Structural, Applied, and Mathematical Linguistics," with a dissertation titled "Analysis and Synthesis of Computational Linguistic Systems for Processing Ukrainian Textual Content." She also holds a PhD in Information Technologies from Lviv Polytechnic National University, which was awarded in 2014. World’s Top 2% Research Scientist in 2020-2025 (https://topresearcherslist.com/Home/Profile/1095113). Victoria has authored over 700 publications, and her primary research interests focus on identifying disinformation, fake news, and propaganda, as well as detecting disinformation sources and inauthentic behaviour (e.g., bots) in coordinated groups through the use of artificial intelligence. Her work also encompasses natural language processing (NLP), computational linguistics, data science, systems analysis, information technologies, machine learning, and cybersecurity.



Lyubomyr Chyrun is an Associate Professor in the Information Systems and Networks Department at Lviv Polytechnic National University, Computer Science Department at Yuriy Fedkovych Chernivtsi National University and the Department of Applied mathematics at the Ivan Franko National University of Lviv. Since 2001, he has worked at Lviv Polytechnic University's Institute of Computer Sciences and Information Technologies as an associate professor in the Department of Information Systems and Networks. In 2007, he defended his PhD thesis on May 1, 2002 - "Mathematical modelling and computational methods". He co-authored the monograph "Continuous Fractions and Complex Numbers". World’s Top 2% Research Scientist in 2021-2022 (https://topresearcherslist.com/Home/Profile/550263). He is the author of more than 200 scientific publications. His areas of scientific interest include pattern recognition, the application of numerical methods in information technologies, object-oriented programming, web technologies, fake news identification, natural language processing, computer linguistics, data science, system analysis, and machine learning.

Dmytro Uhryn graduated from Yuriy Fedkovych Chernivtsi National University, Chernivtsi. He is a Doctor of Technical Sciences and an professor at Yuriy Fedkovych Chernivtsi National University. He has currently published over 250 articles. His research interests include data mining, information technology for decision support, swarm intelligence systems, artificial neural networks, and industry-specific geographic information systems.

Zhengbing Hu: Prof., Deputy Director, International Centre of Informatics and Computer Science, Faculty of Program Systems and Applied Mathematics, National Technical University of Ukraine "Kyiv Polytechnic Institute", Ukraine. Adjunct Professor, School of Computer Science, Hubei University of Technology, China. Visiting Prof., DSc Candidate at the National Aviation University (Ukraine) from 2019. Primary research interests: Computer Science and Technology Applications, Artificial Intelligence, Network Security, Communications, Data Processing, Cloud Computing, Education Technology.

Yurii Ushenko: World's Top 2% Research Scientist (2020, 2021, 2022, 2024). Nominee of The Photonics100 2025 list by ElectroOptics. Winner of 1000 Talents Plan Program, PhD, DSc, Prof., at Shaoxing University, Shaoxing, Zhejiang Province 312000, China. Professor, Head of Computer Science Department, Chernivtsi National University, Ukraine. His research interests include intelligent information systems design, data mining, pattern recognition and digital image processing, artificial neural networks, laser polarimetry and interferometry.

How to cite this paper: Sofia Chyrun, Victoria Vysotska, Lyubomyr Chyrun, Dmytro Uhryn, Zhengbing Hu, Yurii Ushenko, "Information Technology for VR Training Evaluation with First Aid Skills Improvement to Detecting Human Resource Damage in Emergencies based on Behavioural Methods", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.2, pp. 151-207, 2026. DOI:10.5815/ijigsp.2026.02.10