•Article•

Interaction design for paediatric emergency VR training

TJ MATTHEWS[1], Feng TIAN[2*], Tom DOLBY[3]

1. *Centre for Digital Entertainment, Bournemouth University, UK*
2. *Faculty of Science & Technology, Bournemouth University, UK*
3. *Head of Simulations, AiSolve, UK*

*** Corresponding author,** ftian@bournemouth.ac.uk

Received: 17 April 2020 **Accepted:** 8 July 2020

Supported by Centre for Digital Entertainment (EP/L016540/1, EPSRC, UK)

Citation: TJ MATTHEWS, Feng TIAN, Tom DOLBY. Interaction design for paediatric emergency VR training. Virtual Reality & Intelligent Hardware, 2020, 2(4): 330—344
DOI: 10.1016/j.vrih.2020.07.006

Abstract **Background** Virtual reality (VR) in healthcare training has increased adoption and support, but efforts are still required to mitigate usability concerns. **Methods** This study conducted a usability study of an in-use emergency medicine VR training application, available on commercially available VR hardware and with a standard interaction design. Nine users without prior VR experience but with relevant medical expertise completed two simulation scenarios for a total of 18 recorded sessions. They completed NASA Task Load Index and System Usability Scale questionnaires after each session, and their performance was recorded for the tracking of user errors. **Results and Conclusions** Our results showed a medium (and potentially optimal) Workload and an above average System Usability Score. There was significant improvement in several factors between users' first and second sessions, notably increased Performance evaluation. User errors with the strongest correlation to usability were not directly tied to interaction design, however, but to a limited 'possibility space'. Suggestions for closing this 'gulf of execution' were presented, including 'voice control' and 'hand-tracking', which are only feasible for this commercial product now with the availability of the Oculus Quest headset. Moreover, wider implications for VR medical training were outlined, and potential next steps towards a standardized design identified.

Keywords Virtual reality; Medical training; Human-Centred design; Interaction design

1 Introduction

Virtual reality (VR) technology has found widespread adoption for use as medical simulations and training, particularly since the off-the-shelf modern generation arrived in 2016[1]. There are strong validity studies supporting the usage of VR technology to increase/strengthen learning outcomes and incur improved knowledge retention[2–4]. With a current majority of users introduced to VR training systems using VR for the first time[5], as well as usability's correlation with learning outcomes, robust, intuitive interactions are crucial to retain users through onboarding and the further sessions.

Overwhelmingly VR training applications and research projects in the healthcare domain are focused on surgery[2,4], which fits within VR being suitable for a heavily procedure-based practice. Other applications include diagnosis[6] and nursing[7], and clinical decision-making skills[8–10] such as the paediatric emergency VR application, which this paper focuses on.

VR in this application and study refers to 'desktop' or 'full' VR hardware (such as using the Oculus Rift as here) with positional and rotational head and hand tracking, as opposed to 'mobile' or 'limited' VR hardware (such as the Oculus Go) which only has rotational head and hand tracking.

This study builds from previous research by Chang et al.[11] that uses the same VR resuscitation training application developed by AiSolve[12]. The VR training application, as designed and produced with paediatric medicine subject experts, presents trainees with an emergency medicine scenario in which they have to diagnose, provide intervention, and stabilize the patient. More information on scenario design and interaction is given in Section 2.1.

The Chang et al. study drew comparison between the stress levels experienced by residents in both real-life resuscitation events and in simulated emergency medicine events as within the VR application[11]. Findings indicate this VR application places trainees in an 'optimal' level of stress that could show greater learning outcomes than low-stress simulations.

However, there is a recommendation for "strategies to mitigate the novelty and "foreign" feel of the VR system are needed if VR is a viable simulation modality"[11], to improve the onboarding process for trainees, particularly those without prior VR experience. This indicates a need for a standardised, intuitive interaction design for VR training applications. Due to the relative novelty of widespread adoption of virtual reality (VR) as a training application, the evidences for best practices of interaction designs, for training purposes, are lacking.

Jason Jerald has outlined a selection of VR interaction designs[13], inspired by the Human-Centred Design of Norman[14], but these are primarily based on designs originating prior to the modern generation-2016 onwards-of VR hardware, and furthermore of the current generation-2018 onwards-that brings standalone headsets and additional interaction affordances[1]. The same is true of the recommendations set out by Alger[15], which focus on user interface interactions in VR.

There are some VR interaction studies within this current generation, and they tend to explore locomotion controls[16–18] and otherwise don't advance beyond the previously defined interaction affordances; no study could be identified that investigated interaction design for the specific tracking features of the Oculus Quest[19] or Valve Knuckles[20], for example. There is a risk of difference between this potential wave of VR interaction research and its precursors, namely that studies predating this current iteration may cite negative usability results due to issues that current hardware possibilities can negate[21–23].

With this gap in mind, this study aims to evaluate the usability of the current interaction design of the VR resuscitation training application[12] which, as it shares a common design with several VR medical training applications, should also identify design suggestions to further close the 'gulf of execution'[14] via standardised interactions for the purposes of VR training, and will compare particularly explore where usability could be enhanced with alternative, current generation VR hardware. The 'gulf of execution' here, as defined by Norman[14], means the gap between a user's goal or intent (i.e. 'give 20mg of medication X') and the user's understanding of how to execute that goal (i.e. 'point at and select medication group, point at and selection medication X').

This study defines usability as with ISO 9241-11: "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use", with a specific focus on interaction design (aligned with Human-Centred Design) for the scope of this research. It uses standard usability metrics for consistency with Chang et al.[11], and the focus is given to identifying user errors during training performances, including such as relating to the aforementioned 'gap of execution' in order to identify not just the usability of the combined hardware-software training system, but the usability potential separated from any possible hardware restrictions.

The results of this user testing found similar Workload scores to Chang et al. [11], and an acceptable usability score. Some similar but infrequent user errors were recorded, and the only error group with major occurrence (>40% average) was directly tied to the VR hardware interface.

This paper concludes with identifying connections between the other user errors and usability metrics to outline the gaps that theoretically exist across the standard for VR medical training, and thus would be most suited for future research and development efforts, taking into particularly consideration current generation VR hardware affordances.

2 Methods

Users recruited for our study were all physician employees of Children's Hospital Los Angeles, US, deemed to have the applicable existing medical knowledge to know the procedures needed to solve the resuscitation scenarios as presented. These users were targeted as they have equivalent knowledge and experience to the target trainees of the VR application.

Some had prior knowledge/awareness of VR technology, but no direct experience with either the paediatric emergency VR training used in this study, or any other VR training application. Any users with prior experience with the paediatric emergency VR training used in this study, including those involved in the previous Chang et al. research[11], were excluded from this study.

2.1 Simulation

The paediatric emergency VR training used in this study is an application available for the Oculus Rift[24] and Oculus Go[25] hardware, developed using the Oculus Utilities plugin with the Unity game engine[26] . Trainees are placed in an immersive resuscitation room environment and tasked with completing simulation resuscitation emergency scenarios. The two initial scenarios are infant status epilepticus and paediatric anaphylactic shock: high-risk, low-frequency paediatric resuscitation scenarios. These were designed with the input of subject matter experts and physician authors of the previous study[11] and require observation, diagnosis, and intervention knowledge and skills. These scenarios were chosen for consistency with the Chang et al. study[11], in which they were initially designed to fulfil a perceived gap in high-risk, low-frequency resuscitation exercises. They are also suitable for the test group users, whose existing medical knowledge covers the prerequisites for the scenarios.

Test sessions took place using Oculus Rift DK2 with two tracked Oculus Touch controllers, running through a VR-ready desktop PC. These headsets were chosen for consistency with Chang et al. [11], but no major difference in results would be expected in using alternative VR headsets with similar specifications and functionality, such as the HTC Vive, although this would require further investigation. The application ran at a steady 90 frames per second and third-party screen-capture software was used to record user performances (which had no discernable impact on frame rate).

Within the VR scenario the trainee acts as 'code captain' and directs the other medical staff on which actions to perform. In the environment there are the same persons (avatars) as would be expected in a real-life resuscitation room during an emergency scenario including (Figure 1):

- EMT—Provides initially known pretext information and introduction to scenario symptoms.
- Nurse—Commits instrument-based actions and provides ongoing feedback on patient state.
- Respiratory Therapist—Commits airway-based actions.
- Patient—Focus of scenario, can be examined and has realistic physiological state changes in response to trainee decisions.

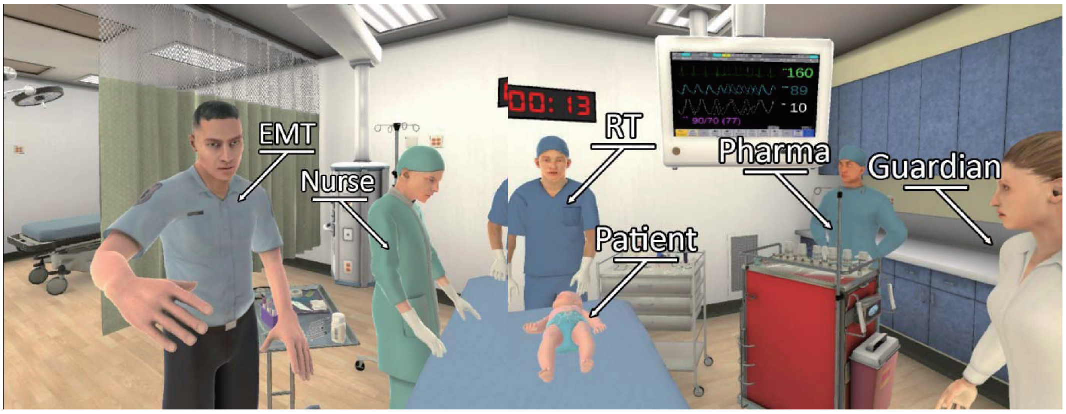


Figure 1　Avatars in virtual simulation environment.

- Pharmacist—Commits medication-based actions.
- Guardian—Responds emotionally to patient state (included for immersion/realism).

Scenarios have patient symptoms programmed as finite-state machines, in which trainees commit the appropriates step in a pre-defined prioritised order, as prompted by patient changes and stimuli; this is to support this training to function similarly to script concordance tests[27], as these two included scenarios have pre-defined decision algorithms that trainees should be following.

In the 'infant status epilepticus' scenario, the virtual patient is having an epileptic seizure, difficulty breathing, and vomiting. This is audio-visually shown, communicated by other virtual agents, and shows on medical examinations (i.e. shallow/reduced breathing when listening to patient breath). Correct procedure is determined as: clearing the airway, providing oxygen to stabilize, and then setting up an IV with a seizure medication (localized as Lorazepam in this build), as well as regular checks of patient pupils and capillary reflex.

In the 'pediatric anaphylactic shock' scenario, the virtual patient is having a life-threatening allergic reaction, and difficulty breathing. This is audio-visually shown, communicated by other virtual agents, and shows on medical examinations (i.e. shallow/reduced breathing when listening to patient breath). Correct procedure is determined as: providing medication to increase blood pressure (localized as Albuterol), providing medication as per anaphylaxis treatment algorithm (localized as Epinephrine, Methylprednisolone, Diphenhydramine, and Ranitidine), and then setting up advance airway intubation, as well as regular checks of patient breathing, pulse, and capillary reflex.

Each scenario takes between 3−5 minutes to complete, and has two variation axes available:

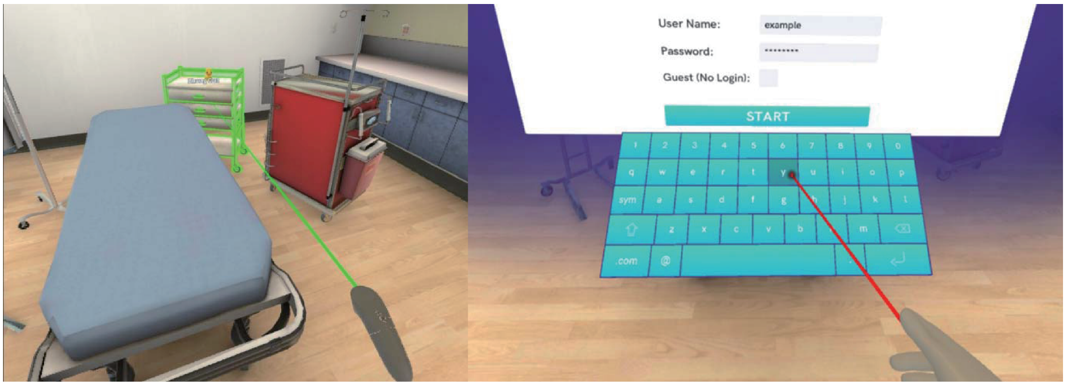
- Difficulty: On Beginner difficulty the nurse avatar provides strongly guided suggestions for next steps, and the standard (most common) procedure for the symptoms of the scenario is sufficient. On Advanced difficulty the nurse avatar will not provide hints, and additional steps are required in response to non-standard symptoms of the scenario.
- Distraction: Higher distraction levels include more audio-visual stressors and external stimuli in the training environment, e.g. background noise levels, evocative language, visual clutter.

For the purposes of this study, all users completed scenarios on Beginner difficulty and Low distraction. Unlike the previous study[11], in which Advanced difficulty and High distraction were chosen to induce higher stress levels, for this study these low settings were chosen to minimise the impact of scenario difficulty and stressors inadvertently affecting usability scores.

2.2　Interaction design

Within the VR environment, the trainee directs the other medical staff on which actions to perform by navigating to and selecting the appropriate instrument or medication. Users do not directly interact with the virtual avatars other than to select which tools and objects are to be used, and similarly do not perform the actions with the tools and objects directly. Selections are considered as instructions for the virtual avatars, mimicking directions given by the code captain in the real-world resuscitation scenarios and in mannequin-based training. Procedures such as pupil check and capillary reflex can be and are performed directly by selecting the appropriate hotspot on the patient.

Interaction uses the Pointing Pattern with semi-realistic hands[13] in which a ray is extended from the user's dominant hand (as chosen in the application menu) akin to the user 'pointing' at the object they wish to select (Figure 2). Raytracing is used to determine the closest interactable object intersected by this selection ray, which is then highlighted, and selection is completed with the controller trigger button.



User Name: example

Password: ********

Guest (No Login):

Figure 2 Selectable object in virtual simulation environment, and pointer controllers for Oculus Go and Oculus Rift.

The application is a 'standing VR' experience[28] in which users have a small area, positioned relative to their starting physical position, to move, pivot and look around in, but cannot physical walk around the entire virtual environment. This approach was taken, as opposed to 'room-scale VR'[29], to support both Oculus Rift and Oculus Go hardware (as requested by clients), and to support limited training space available to learners in the target medical institutions.

To make an instruction on which action to perform, users first select the relevant item holder (instrument tray, airway cart, supply cart) and are moved via 'blink teleportation'[17] to a closer vantage point near the item holder. From there they select the appropriate object that corresponds to the action they will to be committed, i.e. selecting the lorazepam vial to administer lorazepam to the patient.

These interactions are built using the Oculus Utilities plug-in for the Unity game engine[30], which provides prefabbed components for VR applications, including the virtual hands and pointer ray used in this application.

Each object selected (and thus decision made) by the user can prompt one of the following outcomes:

- Positive: This decision is defined as correct for the current state of the scenario, will be acted upon.
- Neutral: This decision is defined as unnecessary, but not harmless, for the current state of the scenario, will be acted upon.
- Negative: This decision is defined as incorrect and harmful for the current state of the scenario, will not be acted upon and feedback given.
- Undefined: This decision has not been defined for the current state of the scenario, will not be acted upon and feedback given.

A tutorial scenario to introduce users to the interaction control scheme is included. It involves learning how to select objects and navigate around the resuscitation room. Non-medical prompts are used here to not impact diagnostic ability in the full training scenarios.

2.4 Outcome measures

To record usability, this study uses the NASA Task Load Index (NASA-TLX)[31] rating scale and the System Usability Scale (SUS)[32].

NASA-TLX is a proven scale[31] designed for workload estimates of tasks and systems, consisting of six subscales (Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance). It was chosen for consistency with Chang et al.[11] and to determine correlations between user errors and associated TLX scores.

SUS is an established list of ten Likert subscales to provide a "measure of people's subjective perceptions of the usability of a system"[32]. It was chosen for a quick, reliable measurement of both learnability and usability[32], and for the provision of specific usability scores to correlate user errors against.

To identify potential gaps in usability that lead to user error, performances in each simulation scenario were recorded and each error was recorded and coded within the Generic Error Modeling System[33] as follows:

- Skill-Based Errors:
- Controller Issue (CI)—Issues caused by incorrect handling of the physical controllers / confusion over how to select or navigate the environment. Example: Difficulty with directing ray pointer resulted in unintended object selection.
- State Misidentify (SM)—Dentified gap between the understood state of the scenario and the actual state. Example: Thought medications had been given when they had not.
- Environment Awareness (EA)—User unable to locate object they wish to select. Example: Unable to locate specific medication to select.
- Rule-Based Errors:
- Uncoded Medication Request (UMR)—A medication was requested that was not made available for this scenario by the subject expert co-authors.
- Uncoded Airway Request (UAR)—An airway tool was requested that was not made available for this scenario by the subject expert co-authors.
- Uncoded Procedure Request (UPR)—A medical procedure was requested that was not made available for this scenario by the subject expert co-authors. Example: User expects to be able to order specific test unnecessary for the scenario.
- Uncoded Action Order (UAO)—A defined decision was made that required a precursor decision to be made before being feasible. Example: An IV medication was selected before an IV-line start had been selected.

Note that, different from Skill/Rule-Based Errors above listed, Knowledge-Based Errors (such as selecting a medication intentionally that is defined as harmful for the situation) were not tracked as these are in the remit of the training outcomes and thus shouldn't be quantified as linked to usability.

2.5 Test design

Before beginning the test session, users were given an overview of the study and prompted to complete the included tutorial to familiarise themselves with the basic controller interactions. After this, in order to ensure all users begin with the same base knowledge, users were shown the main object holders in the virtual environment on physical cue cards, and which items they could expect to find on each. Specific objects were not listed, so as to not prime users towards particular options. Cue cards were presented, in consistency with Chang et al.[11], as in usual proceedings the trainees would be familiar with the real-world environment that is replicated in the VR simulation. To reduce any impact of room layout/object placement (which is dictated by the real-world environment) on usability, cue cards were presented to trainees regardless of prior experience.

They were also onboarded on what stimuli to expect as the scenario begins specifically that the EMT would be giving information to them to reduce their mental load at scenario start and allow information provided to be processed, which not doing so would affect the remaining scenario decisions and thus inadvertently affect usability scores. Onboarding of this nature plays a crucial part of mannequin-based simulation (that this VR application is supplementing/replicating), and it was previously found, during Chang et al.[11] studies, users not given some familiarity before starting the VR simulation were seen to not register information or stimuli being presented at the very beginning of the scenario.

Users completed both available training scenarios, Status Epilepticus and Anaphylaxis. As there was no feasible way in this study to validate individual users' scoring in these scenarios beforehand without revealing the symptoms of the scenarios involved (and thus invalidating test data), consistent scenario order was chosen.

Each session was observed by the researcher who noted errors (as defined by Outcome Measures) and both voice and screen recordings were made of users' performances. Users were requested to adhere to a think-aloud protocol[34], in which they vocalised their decision-making process and the scenario as they understood it at each stage. This captured included, for example: self-identified user errors, vocalisations of controller issues, and incorrect state analyses.

Data logs from the Resuscitation VR application were also captured, containing, as relevant to this study, time entries for object selections and stimuli triggers in the virtual environment. The transcripts of these voice recordings were later cross-referenced with both the screen recordings and data log to extrapolate further errors as defined by Outcome Measures.

Test sessions took place in the following order:

(1) Non-VR: Study introduction

(2) VR: Controls and interaction tutorial

(3) Non-VR: Environment onboarding

(4) VR: Scenario #1 (Status Epilepticus)

(5) Non-VR: User questionnaire and post-performance debriefing

(6) VR: Scenario #2 (Anaphylaxis)

(7) Non-VR: User questionnaire and post-performance debriefing

Debriefing consisted of unprompted feedback gathering, followed by prompted discussion of observed errors. These were included in the cross-reference involving voice recordings as outlined above to identify instances of user errors (section 2.4).

3 Results

Nine users completed the study for a total of eighteen performances (not including tutorial scenarios). These were all the available residents of Children's Hospital Los Angeles who answered to a recruitment call during a three-week visit and had existing medical knowledge that would allow them to complete the scenarios successfully.

One user's SUS scorings were identified as outliers (gave the lowest scores possible for all scales, outside 2 standard deviations from the mean) and were removed from analysis. This participant also expressed a direct dislike of virtual reality simulations prior to conducting the study, and wished to end their session before concluding the second scenario, which the researchers determined to also invalidate their scoring.

Example questions (for TLX Mental Demand, TLX, Physical Demand, SUS Frequency (1), SUS Complexity (2), SUS Easy (3) respectively) are shown in Figure 3.

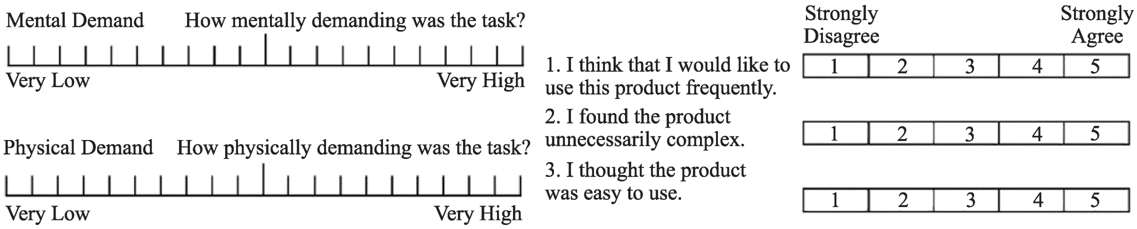

Mental Demand
How mentally demanding was the task?
Very Low
Very High
Physical Demand
How physically demanding was the task?
Very Low
Very High
1. I think that I would like to use this product frequently.
2. I found the product unnecessarily complex.
3. I thought the product was easy to use.
Strongly Disagree
Strongly Agree
1
2
3
4
5

Figure 3 Example questions from NASA TLX and SUS questionnaires.

Each subscale on NASA TLX is scored between 0 and 100, with a higher score indicating a higher task-load (except Performance, which is inverse). A total Workload (Raw TLX) scale is the average calculation and thus uses the inverse of Performance (a positive scale) to match the other TLX scores (all negative scales).

Each SUS scale is scored between 0 and 5, and the higher/lower usability direction alternates with each factor (i. e. SUS Frequency is a 'positively' scored subscale, and SUS Complex is a 'negatively' scored subscale). An additional Bangor Rating of "overall rating" is graded on a Likert scale from 1 to 7 and provides a subjective quality rating to anchor user perception against.

As shown in the results in Table 1, Physical task-load scored low, with an average of 13.82±6.76. Other NASA TLX factors averaged close to medium task-load, with the highest as Effort with 61.63±13.93. The overall Workload was also medium task-load, with 47.96±13.11.

Positive SUS factors (Frequency, Easy, Integrated, Learn Quickly) averaged closest to Agree, with the

Table 1 NASA TLX and SUS scores

Score	Scenario #1	Scenario #2	Change	Average
TLX Mental[1]	57.22 ± 22.37	58.13 ± 24.99	+ 2%	57.67 ± 23.65
TLX Physical[1]	13.89 ± 6.98	13.75 ± 6.50	− 1%	13.82 ± 6.76
TLX Temporal[1]	57.22 ± 26.89	61.88 ± 28.28	+ 8%	59.55 ± 27.65
TLX Performance[2]	43.33 ± 26.03	61.88 ± 26.33	+ 43%	52.60 ± 27.76
TLX Effort[1]	63.89 ± 13.29	59.38 ± 14.24	− 7%	61.63 ± 13.93
TLX Frustration[1]	57.78 ± 25.29	58.75 ± 27.92	+ 2%	58.26 ± 26.57
Workload (Raw TLX)[1]	51.11 ± 8.74	48.33 ± 12.92	− 5%	49.72 ± 10.99
SUS Frequency[2,3]	4.38 ± 0.48	4.43 ± 0.73	+ 1%	4.40 ± 0.61
SUS Complexity[1,3]	2.25 ± 0.97	2.00 ± 0.76	−11%	2.13 ± 0.88
SUS Easy[2,3]	3.88 ± 0.78	3.86 ± 0.99	− 0.5%	3.87 ± 0.88
SUS Support[1,3]	3.13 ± 0.93	2.86 ± 0.99	− 9%	2.99 ± 0.97
SUS Integrated[2,3]	3.88 ± 0.33	3.71 ± 0.45	− 4%	3.79 ± 0.40
SUS Inconsistency[1,3]	2.75 ± 0.66	2.43 ± 0.49	− 12%	2.59 ± 0.61
SUS Learn Quickly[2,3]	3.88 ± 0.60	3.86 ± 0.83	− 0.5%	3.87 ± 0.72
SUS Awkward[1,3]	2.13 ± 1.05	2.00 ± 1.07	− 6%	2.06 ± 1.06
SUS Confident[2,3]	3.50 ± 1.12	3.57 ± 1.05	+ 2%	3.54 ± 1.09
SUS Learn Before[1,3]	2.25 ± 1.09	1.86 ± 1.12	− 17%	2.05 ± 1.12
System Usability Score[2]	67.50 ± 14.68	70.71 ± 15.22	+ 5%	69.11 ± 15.05
Bangor Rating[2]	5.63 ± 0.70	5.57 ± 0.73	− 2%	5.60 ± 0.71

Notes: [1]Lower is better; [2]Higher is better; [3]1=strongly disagree, 5=strongly agree; Values are presented as mean ± SD.

highest as Frequency with 4.40±0.61.

Negative SUS factors (Complexity, Support, Inconsistency, Awkward, Learn Before) averaged between Disagree and Neutral, with the highest as Support with 2.99±0.97. No negative SUS factor averaged above Neutral.

The overall System Usability Score of 69.11 is above average as per Sauro & Lewis[35], and would fall within the 'C' letter grade.

There was significant improvement in Performance (+43%), Complexity (−11%), Support (−9%), Inconsistency (−12%) and Learn Before (−17%) scores in the second session compared to the first session. There was no significant difference in overall Workload between sessions, nor for Bangor Rating.

Error findings are below (Table 2), demarcated into count (the number of sessions in which this error occurred, split per scenario) and sum (the total times this error occurred, as multiple errors can occur in a scenario session, also split per scenario), and the average per scenario (in which this error took place at all).

Table 2 User error findings

	CI[1]		SM[2]		EA[3]		UMR[4]		UAR[5]		UPR[6]		UAO[7]	
Count S1	6	67%[8]	3	33%[8]	2	22%[8]	1	11%[8]	4	44%[8]	1	11%[8]	3	33%[8]
Count S2	5	56%[8]	1	11%[8]	2	22%[8]	5	56%[8]	3	33%[8]	1	11%[8]	1	11%[8]
Count Total	11	61%[8]	4	22%[8]	4	22%[8]	6	33%[8]	7	39%[8]	2	11%[8]	4	22%[8]
Diff Count	–17%		–67%		0%		400%		–25%		0%		–67%	
Sum S1	8		5		2		3		4		2		3	
Sum S2	5		2		2		7		3		3		1	
Sum Total	13		7		4		10		7		5		4	
Diff Sum	–38%		–67%		0%		133%		–25%		50%		–67%	
Avg Per Scenario	1.18		1.75		1.00		1.67		1.00		2.50		1.00	

Notes: [1]Controller Issue; [2]State Misidentify; [3]Environment Awareness; [4]Uncoded Medication Request; [5]Uncoded Airway Request; [6]Uncoded Procedure Request; [7]Uncoded Action Order; [8]Percentage of scenario sessions in which error occurred.

Most users had a Controller Issue (CI) during at least one of their scenario sessions (61% overall). Of these Controller Issues, the specific breakdown (compared to sum) was thusly:

- 38%—Interface confusion: Examples include forgetting where buttons where physically located on the controller, or how to 'point' with the hand to aim at objects.
- 31%—Selection error: Examples include accidentally selecting the wrong object / object because of imprecise aiming.
- 23%—Accidental press: Examples include accidentally pressing the back or select button, typically due to confusion between the two.
- 8%—Assumed functionality: Examples include assumption that the app had voice control.

For most error categories, there was a reduction in both error count and sum comparing the first and second scenario session. The exception to this is Uncoded Medication Requests, which saw a 400% increase in count for the second scenario session. As the second scenario (anaphylaxis) codes 4 mandatory and 2 optional medications, as opposed to the first scenario (seizure epilepticus) with 1 mandatory medication code, it is likely that this is a false positive due to the increased scenario focus on medications.

The highest ratio of errors per scenario was for Uncoded Procedure Requests, with 2.5 average. However, the only two instances of this error are from the same user (in both their scenario sessions), and with high frequency, so this could be considered an outlier.

Finally, to understand the correlations between NASA TLX/SUS scores and user errors, a bivariate (Pearson) correlation analysis was conducted between these two factors. This was conducted to determine if and what user errors types had an impact on which elements of the user's perceptions of usability, which in turn could guide future development and research efforts. Table 3 below shows only where there were significant correlations:

Table 3 Bivariate Pearson correlation analysis

	TLX	TLX	SUS	SUS	SUS	SUS	SUS	SUS	SUS	SUS	SUS	Bangor
	Perf.	Effort	Comp.	Easy	Sup.	Incon.	LQuick	Awk.	Conf.	LBefore		
SM[3]			0.568[1]		0.514[1]	0.515[1]				0.555[1]		–0.605[1]
EA[4]						–0.592[1]	–0.518[1]					
UMR[5]	0.580[1]		0.533[1]	–0.533[1]					–0.542[1]	0.524[1]	–0.577[1]	–0.695[2]
UPR[6]		0.493[1]	0.636[1]	–0.549[1]				0.698[2]	–0.541[1]	0.660[2]	–0.676[2]	–0.862[2]

Notes: [1]Correlation is significant at the 0.05 level; [2]Correlation is significant at the 0.01 level; [3]State Misidentify; [4]Environment Awareness; [5]Uncoded Medication Request; [6]Uncoded Procedure Request

Obviously, there is significant correlation between State Misidentify (SM) error occurrences and negative scoring for Complexity, Support, Inconsistency, Learn Before and Bangor Rating.

It can also be noticed from Table 3 that there is significant correlation between Environment Awareness (EA) error occurrences and negative scoring for Learn Quickly. However, there is also significant correlation with positive scoring for Inconsistency, which implies that users with EA errors felt more that "there was [not] too much inconsistency in the VR simulation", which appears antithetic. There was a low sum (4) of EA error occurrences and so this could be an outlier.

For Uncoded Medication Request (UMR), there is significant correlation with negative scoring for Complexity, Easy, Confidence, Learn Before, System Usability and Bangor Rating. However, this is also a significant correlation between UMR and Performance and positive scoring for Performance. This suggests that users who conducted UMR errors felt that they were "more successful in accomplishing" their task, which is antithetical. A possible explanation for this outlier is that Scenario 2 (anaphylaxis) is where the majority of UMR errors took place (Table 2), and also had a 43% increase in Performance (Table 1), likely due to it being the second scenario. Thus, it is suggested that this correlation is not the result of causation.

Uncoded Procedure Requests (UPR) also had significant correlations with negative scoring for Effort, Complexity, Easy, Awkwardness, Confidence, Learn Before, System Usability, and Bangor Rating. Unfortunately, as outlined in regard to Table 2, both UPR errors were conducted by the same user so limited extrapolation can be made about these correlations.

4 Discussion

Overall NASA-TLX factors, including Workload, scored around medium task-load across both scenarios, which revalidates that this simulation falls within the 'optimal' stress levels for resuscitation training,, consistent with the findings of Chang et al.[11]. 'Optimal' stress here is defined as user mental stress load (as matched with TLX scores) that is not low enough for the user to disengage because of a lack of interest or boredom, and not high enough for the user to disengage because of frustration or being overwhelmed.

Regarding System Usability, the 69.11 score (which correlates to a 'C' grade) indicates that this training system is suitable for VR newcomers, but that there is still room for improvement. The user errors found to have valid correlation with usability are State Misidentify (SM) and Uncoded Medication Requests (UMR) errors.

This suggests that the 'gulf of execution'[14] lies not in the interaction designs directly, but in the 'possibility space'[36] afforded. 'Possibility space' is a term adopted from game design research, and concepts "all of the gestures made possible by a set of rules" [36], which in this study specifically means both the range of and type of interactions available by the user. These results suggest that the source of user errors is not from difficulties with the existing interaction control, but from limitations of what information can be given/received using these interactions.

The wicked problem of adequate 'possibility space' to meet both designer and user/player intent has been explored well in general games studies[36,37] and furthermore recently in gameful pedagogies[38,39]. Some research into 'possibility space' and affordances exists for virtual reality training[40–42] but, much like this study, have been focused on physical-interface interaction design rather than a holistic review.

The 'gulf of execution' limits what decision-making data can be captured by the VR application, and as such can serve well as supplementary to mannequin-based training (as it currently is within the partner institution[11]), but cannot fully replace mannequin-based training for learning outcomes and debriefing capabilities.

It should be noted that, despite Controller Issues (CI) constituting the majority of all user errors, no significant correlation was found between CI and usability scores. This could indicate that users perceive the usability aspects of the software system as separate from the hardware with which issues occurred; an explanation for this could be that CI causes breaks-in-presence[43] and thus shifts mental workload effort away from the training application, affecting immersion (and therefore learning outcomes) but not usability scoring. Exploring this possibility warrants further user testing.

Improving affordances by increasing the quantity of options (in the form of virtual objects/interfaces) available for medications and state information requests with the existing physical interaction could have negative effects. There are two design strategies: the first is to add virtual objects/interfaces to the existing environment, which could increase visual clutter, and could incur user errors (particularly Controller Issues and Environment Awareness); the second is to replace virtual objects/interfaces is those requested, which could 'prime' users to make selections they would not have otherwise.

Additionally, some actions described during think-aloud were optional and aren't typically included in mannequin training-examples include wanting to comfort the guardian, wanting to reposition the patient, wanting to provide after-care, etc. It has been frequently discussed that, in general, users and trainees have high expectations of VR applications[44–46] and anticipating these affordances could match that high expectation.

It is argued, therefore, that these findings draw attention to an acknowledged but open problem of how to best replicate Closed-Loop Communication, as would be used in medical mannequin simulations, in Virtual Reality[47,48]. With physical interfaces such extensive options aren't feasible for the reasons outlined above. It is suggested that a possible functionality to allow this would be a natural language interface, or 'voice control'. Voice control, the ability to speak commands to the simulation, has already found implementation in augmented and virtual reality medical applications, particularly in surgery[49] and diagnostics[50], but further research is required to impact and strategies for adoption.

This would allow for both increasing the quantity of options without the issues previously outlined and would also allow for user feedback when an uncoded action is requested (an Uncoded Medication Request, for example). This may also alleviate State Misidentify errors, as this opens an interface for users to vocalise requests for state updates.

Some physical interaction with virtual objects would still be required though, and there remains high Controller Issues frequency. There are two main issues per Results: the physical controller (Interface confusion, Accidental press) and the selection method (Selection error).

A frequently developed alternative to the physical controller is to use a 'hand-tracking' interface. This involves visual algorithms that can track and understand the position, rotation, and gestures of a user's hand and digits, which allows, for example, natural pointing and grabbing motions to be replicable in virtual space. Until recently, these have used costly third-party hardware[51,52] or bespoke, limited computer vision algorithms[53–55]. However, as of December 2019, hand-tracking functionality is available off-the-shelf for the Oculus Quest headset[19], and is available through the same Oculus Utilities plug-in that this project uses[56]. This supports an exploration into converting the current physical interface to a hand-tracking interface.

Hand-tracking doesn't alleviate the second main issue of Selection errors, however. The current Pointing Pattern interaction design is required due to the distance between the user and environment objects to select. An alternative would be to have Proximity Selection or direct Hand Selection[13], which requires closer proximity to objects. As 'voice control' outlined above should reduce the need for physical representations of object selections, Hand Selection could be supported by positioning the user nearest the remaining physical objects, which in this application is the patient interactions, like checking pulse.

Exploring other VR medical simulations with decision-making processes[8,9,57,58] indicate that this 'gulf of execution' may be common, as similar interaction designs are found. Therefore, these next steps outlined above should have further reaching implications for a standardised design of VR medical training applications as a whole.

4 Future work

The above analysis support a future development effort to integrate and evaluate the impact of 'voice control' and 'hand tracking' functionality on this VR simulation, both in terms of usability impact and or error reduction (and furthermore any determinable impact on training performance).

Due to the relatively small sample size used in this study, the results here do have a reduced statistical power, and should be seen as indicative rather than fully conclusive. Future research and user tests following these efforts should involve a larger sample.

Furthermore, whilst all users involved in this study were applicable as the target audience for the VR application training outcomes, conclusions regarding general usability of specific designs and functionalities would require an abstracted task exercise that does not require pre-requisite knowledge. Initial validation of the future functionality developments outlined above should involve these task designs.

With a focus on interaction design this study does not capture the full extent of the usability of the system, and future efforts will be made to evaluate holistically the integration of the VR simulation within it's intended training purposes, particularly examining holdover elements from mannequin-based training such as the training environment and user processing/grading.

Finally, it could also provide useful outcomes to perform a metanalysis on the evaluation methods of VR.

5 Conclusion

This study evaluated the workload and usability of an in-use paediatric VR training application using the NASA-Task Load Index and System Usability Scale measures. Users completed two VR scenarios in sequence, and improvements in self-reporting for Performance as well as some SUS factors were seen. The application was found to have medium (and thus potentially optimal) Workload, and a 'C' grade usability ranking, which suggest that the current interaction design for this training application is suitable as supplementary to mannequin-based training.

However, despite highest frequency of Controller Issues, other user errors linked to Norman's[14] 'gulf of execution' had the only significant correlation to usability scores, and demonstrates a limitation to fully replicating decision-making affordances of mannequin-based training with existing VR interaction design.

Potential next steps were identified within the remit of Human-Centred Design[13], namely 'voice control' and 'hand-tracking' to close the 'gulf of execution'. Widespread adoption of these functionalities has been previously limited by hardware affordances, but with the current Oculus Quest headset available, development of these interaction designs would be a suitable next step. It is hoped that these findings and recommendations have wider implications for VR medical training as a whole, and that future interaction patterns can address the underlying issues found in the current standard design.

References

1 Virtual Reality Society. https://www.vrs.org.uk/virtual-reality/history.html

2 Vaughan N, Dubey V N, Wainwright T W, Middleton R G. A review of virtual reality based training simulators for orthopaedic surgery. Medical Engineering & Physics, 2016, 38(2): 59–71
DOI:10.1016/j.medengphy.2015.11.021

3 Vaughan N, Gabrys B, Dubey V N. An overview of self-adaptive technologies within virtual reality training. Computer Science Review, 2016, 22: 65‒87
DOI:10.1016/j.cosrev.2016.09.001

4 Moglia A, Ferrari V, Morelli L, Ferrari M, Mosca F, Cuschieri A. A systematic review of virtual reality simulators for robot-assisted surgery. European Urology, 2016, 69(6): 1065–1080
DOI:10.1016/j.eururo.2015.09.021

5 Cohen L, Duboé P, Buvat J, Melton D, Khadikar A, Shah H. Augmented and virtual reality in operations. 2018

6 Indhumathi C, Chen W, Cai Y Y. Multi-modal VR for medical simulation. International Journal of Virtual Reality, 2009, 8(1): 1‒7
DOI:10.20870/ijvr.2009.8.1.2707

7 Kleven N F, Prasolova-Førland E, Fominykh M, Hansen A, Rasmussen G, Sagberg L M, Lindseth F. Training nurses and educating the public using a virtual operating room with Oculus Rift. In: 2014 International Conference on Virtual Systems & Multimedia (VSMM). Hong Kong, China, IEEE, 2014, 206–213
DOI:10.1109/vsmm.2014.7136687

8 Harrington C M, Kavanagh D O, Quinlan J F, Ryan D, Dicker P, O'Keeffe D, Traynor O, Tierney S. Development and evaluation of a trauma decision-making simulator in Oculus virtual reality. The American Journal of Surgery, 2018, 215 (1): 42‒47
DOI:10.1016/j.amjsurg.2017.02.011

9 Crosby L E, Real F J, Cruse B, Davis D, Klein M, McTate E, Hood A M, Brinkman W, Hackworth R, Quinn C T. An immersive virtual reality curriculum for pediatric providers on shared decision making for hydroxyurea. Blood, 2019, 134(Supplement_1): 3402
DOI:10.1182/blood-2019-128661

10 Jacklin S, Chapman S, Maskrey N. Virtual patient educational intervention for the development of shared decision-making skills: a pilot study. BMJ Simulation and Technology Enhanced Learning, 2019, 5(4): 215‒217
DOI:10.1136/bmjstel-2018-000375

11 Chang T P, Beshay Y, Hollinger T, Sherman J M. Comparisons of stress physiology of providers in real-life resuscitations and virtual reality-simulated resuscitations. Simulation in Healthcare, 2019, 14(2): 104–112
DOI:10.1097/sih.0000000000000356

12 Resuscitation VR. Version 1.0. Luton: AiSolve. 2017

13 Jerald J. The VR Book: Human-centered design for virtual reality. New York, ACM Press, 2016

14 Norman D A. The design of everyday things. Massachusetts: MIT Press, 2013

15 Alger M. Visual design methods for virtual reality. Dissertation for the Masters' Degree. London, Ravensbourne University, 2015

16 Boletsis C, Cedergren J E. VR locomotion in the new era of virtual reality: an empirical comparison of prevalent techniques. Advances in Human-computer Interaction, 2019, 1‒15
DOI:10.1155/2019/7420781

17 Ntokos K. Techniques on multiplatform movement and interaction systems in a virtual reality context for games. In: Advances in Multimedia and Interactive Technologies. IGI Global, 2019, 199‒216

DOI:10.4018/978-1-5225-5912-2.ch009

18 Calandra D, Lamberti F, Migliorini M. On the usability of consumer locomotion techniques in serious games: comparing arm swinging, treadmills and walk-in-place. In: 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin). Berlin, Germany, IEEE, 2019, 348−352
DOI:10.1109/icce-berlin47944.2019.8966165

19 Oculus. Thumbs up: hand tracking on oculus quest this week. 2019

20 Valve. Controllers-Valve Index. 2019

21 Freina L, Ott M. A literature review on immersive virtual reality in education: state of the art and perspectives. In: The International Scientific Conference E-learning and Software for Education. 2015, 1(133): 10−1007

22 Jensen L, Konradsen F. A review of the use of virtual reality head-mounted displays in education and training. Education and Information Technologies, 2018, 23(4): 1515−1529
DOI:10.1007/s10639-017-9676-0

23 Carruth D W. Virtual reality for education and workforce training. In: 2017 15th International Conference on Emerging ELearning Technologies and Applications (ICETA). Stary Smokovec, Slovakia, IEEE, 2017, 1–6
DOI:10.1109/iceta.2017.8102472

24 AiSolve. Resuscitation VR on Oculus Rift. 2019

25 AiSolve. Resuscitation VR on Oculus Go. 2019

26 Unity Technologies. Unity. 2020

27 Fournier J P, Demeester A, Charlin B. Script concordance tests: guidelines for construction. BMC Medical Informatics and Decision Making, 2008, 8(1): 1−7
DOI:10.1186/1472-6947-8-18

28 Virtual Reality and Augmented Reality Wiki. Standing VR. 2017

29 Gepp M. Roomscale 101-An Introduction to Roomscale VR. 2017

30 Oculus. Oculus Utilities for Unity. 2020

31 Hart S G. Nasa-task load index (NASA-TLX); 20 years later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2006, 50(9): 904−908
DOI:10.1177/154193120605000909

32 Brooke J. SUS: a retrospective. Journal of Usability Studies, 2013, 8(2): 29–40

33 Reason J. Human error. Cambridge: Cambridge University Press, 1990

34 Dumas J. Usability testing methods: think-aloud protocols. In: Design by people for people: Essays on usability, Usability Professionals' Association, 2001, 119−130

35 Lewis J R, Sauro J. Quantifying the user experience: Practical statistics for user research. Amsterdam: Elsevier, 2012

36 Bogost I. The rhetoric of video games. In: The Ecology of Games: Connecting Youth, Games, and Learning, MIT Press, 2008, 117−139
DOI: 10.1162/dmal.9780262693646.117

37 Jones S E. The meaning of video games: gaming and textual strategies. Oxfordshire, Routledge, 2008

38 Cooke L, Dusenberry L, Robinson J. Gaming design thinking: wicked problems, sufficient solutions, and the possibility space of games. Technical Communication Quarterly, 2020, 1−14
DOI:10.1080/10572252.2020.1738555

39 Caravella E. Teaching gamefully: proceduralizing the classroom through possibility space pedagogy. Dissertation for the Doctoral Degree. Virgina, George Mason University, 2019

40 Miller N, Willemsen P, Feyen R. Comparing interface affordances for controlling a push broom in VR. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). Reutlingen, Germany, IEEE, 2018, 635–636
DOI:10.1109/vr.2018.8446510

41 Gordon C L, Shea T M, Noelle D C, Balasubramaniam R. Affordance compatibility effect for word learning in virtual reality. Cognitive Science, 2019, 43(6): e12742
DOI:10.1111/cogs.12742

42 Ruffaldi E, Bardy B T, Gopher D, Bergamasco M.Feedback, affordances, and accelerators for training sports in virtual environments. Presence: Teleoperators & Virtual Environments, 2011, 20(1): 33−46

DOI:10.1162/pres_a_00034

43 Slater M, Steed A. A virtual presence counter. Presence: Teleoperators and Virtual Environments, 2000, 9(5): 413–434
DOI:10.1162/105474600566925

44 Keskitalo T. Students' expectations of the learning process in virtual reality and simulation-based learning environments. Australasian Journal of Educational Technology, 2012, 28(5): 841–856
DOI:10.14742/ajet.820

45 Wang H G, Ma Z H, Cao M M. VR shooting training system interaction design. In: 2018 International Conference on Virtual Reality and Visualization (ICVRV). Qingdao, China, IEEE, 2018, 160–161
DOI:10.1109/icvrv.2018.00057

46 Maravilla M M, Cisneros A, Stoddard A, Stretching D, Murray B, Redmiles E. Defining virtual reality: Insights from research and practice. In: iConference 2019 Proceedings. iSchools, 2019, 1–5
DOI:10.21900/iconf.2019.103338

47 Cordar A, Wendling A, White C, Lampotang S, Lok B. Repeat after me: Using mixed reality humans to influence best communication practices. 2017 IEEE Virtual Reality (VR), 2017, 148–156
DOI:10.1109/vr.2017.7892242

48 Balint B N. [DC] designing VR for teamwork: the influence of HMD VR communication capabilities on teamwork competencies. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). Osaka, Japan, IEEE, 2019, 1365–1366
DOI:10.1109/vr.2019.8798147

49 Pratt P, Ives M, Lawton G, Simmons J, Radev N, Spyropoulou L, Amiras D. Through the HoloLens™ looking glass: augmented reality for extremity reconstruction surgery using 3D vascular models with perforating vessels. European Radiology Experimental, 2018, 2(1): 1–7
DOI:10.1186/s41747-017-0033-2

50 Silva J N A, Southworth M, Raptis C, Silva J. Emerging applications of virtual reality in cardiovascular medicine. JACC: Basic to Translational Science, 2018, 3(3): 420–430
DOI:10.1016/j.jacbts.2017.11.009

51 Wozniak P, Vauderwange O, Mandal A, Javahiraly N, Curticapean D. Possible applications of the LEAP motion controller for more interactive simulated experiments in augmented or virtual reality. In: SPIE Optical Engineering + Applications. Proc SPIE 9946, Optics Education and Outreach IV, San Diego, California, USA, 2016, 9946
DOI:10.1117/12.2237673

52 Strazdins G, Pedersen B S, Zhang H X, Major P. Virtual reality using gesture recognition for deck operation training. OCEANS 2017-Aberdeen, 2017, 1–6
DOI:10.1109/oceanse.2017.8084584

53 Schlattman M, Klein R. Simultaneous 4 gestures 6 DOF real-time two-hand tracking without any markers. In: Proceedings of the 2007 ACM symposium on Virtual reality software and technology-VRST '07. Newport Beach, California, New York, ACM Press, 2007, 39–42
DOI:10.1145/1315184.1315188

54 Pan Z G, Li Y, Zhang M M, Sun C, Guo K D, Tang X, Zhou S Z. A real-time multi-cue hand tracking algorithm based on computer vision. In: 2010 IEEE Virtual Reality Conference (VR). Waltham, MA, USA, IEEE, 2010, 219–222
DOI:10.1109/vr.2010.5444787

55 Wang R Y, Popović J. Real-time hand-tracking with a color glove. ACM Transactions on Graphics, 2009, 28(3): 1–8
DOI:10.1145/1531326.1531369

56 Oculus. Hand Tracking. 2020

57 Latham K, Kot P, Wariach A, Al-Jumeily D, Puthuran M, Chandran A. A review on the development of a virtual reality learning environment for medical simulation and training. In: 2019 International Conference on Applications and Systems of Visual Paradigms. Rome, IARIA, 2019, 1–5

58 Schild J, Misztal S, Roth B, Flock L, Luiz T, Lerner D, Herkersdorf M, Weaner K, Neuberaer M, Franke A, Kemp C, Pranqhofer J, Seele S, Buhler H, Herpers R. Applying multi-user virtual reality to collaborative medical training. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). 2018, 775–776
DOI:10.1109/VR.2018.8446160