JMIR Publications
The Leading eHealth Publisher

Clinical Competency Assessments: A Comparative Study of Virtual-Reality-Based and Traditional Physical OSCE Stations

Tobias Mühling, Verena Schreiner, Marc Appel, Tobias Leutritz, Sarah König

Submitted to: Journal of Medical Internet Research
on: December 01, 2023

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint
JMIR Publications

Clinical Competency Assessments: A Comparative Study of Virtual-Reality-Based and Traditional Physical OSCE Stations

Tobias Mühling[1] MD; Verena Schreiner[2]; Marc Appel[2]; Tobias Leutritz[2] PhD; Sarah König[2] MD

[1]Institute of Medical Teaching and Medical Education Research University Hospital Würzburg Würzburg DE

Corresponding Author:
Tobias Mühling MD

Abstract

Background: Objective structured clinical examinations (OSCEs) are a widely recognized and accepted method to assess clinical competencies but are often resource-intensive. Moreover, they may not comprehensively capture the complexity of emergency scenarios.

Objective: This study aimed to evaluate the feasibility and effectiveness of a virtual reality (VR) station compared to traditional physical stations in an already established curricular OSCE.

Methods: Fifth-year medical students participated in an OSCE that included ten stations in total, with one station dedicated to emergency medicine offered in two formats and featuring scenarios of septic and anaphylactic shock in each format. Participants in the study were randomly divided into two groups, participating either in the virtual-reality station (VRS) or the physical station (PHS). Student performance and item characteristics were analyzed focusing on the one emergency and the five other case-based stations; four technical-skills-oriented stations were excluded from this study. Student perceptions were recorded as part of a post-examination online survey to assess the acceptance and usability of VR.

Results: Following randomization and exclusions of invalid datasets, 57 and 66 participants were assessed for the VRS and PHS, respectively. The two VRS scenarios (septic and anaphylactic shock) integrated well and demonstrated a balanced level of difficulty (P = 0.67 and 0.58, respectively) with an average difficulty of 0.68 across all stations. They exhibited above-average values with respect to item discrimination (r' = 0.40/0.33, overall = 0.30) and discrimination index (D = 0.25/0.26, overall = 0.16). VRS participant responses emphasized the realistic portrayal of medical emergencies and the fair assessment conditions provided. However, there was some hesitancy towards its broader application in future practical assessments, highlighting the need both for further familiarization as well as maintaining physical interaction with simulated patients.

Conclusions: Integration of the VRS into the current OSCE framework proved feasible both technically and organizationally, even within the strict constraints of short examination phases and schedules. The VRS was accepted and positively received by students across various levels of technological proficiency, including those with no prior VR experience. Notably, the VRS demonstrated comparable or even superior item characteristics, particularly in terms of discrimination power. While challenges remain, such as technical reliability and some acceptance concerns, VR remains promising in applications of clinical competence assessment.

(JMIR Preprints 01/12/2023:55066)
DOI: https://doi.org/10.2196/preprints.55066

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint JMIR Publications

Original Manuscript

Preprint JMIR Publications

Clinical Competency Assessments: A Comparative Study of Virtual-Reality-Based and Traditional Physical OSCE Stations

Tobias Mühling[1], Verena Schreiner[1], Marc Appel[1], Tobias Leutritz[1], Sarah König[1]
[1]Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Würzburg, Germany

Abstract:

Background: Objective structured clinical examinations (OSCEs) are a widely recognized and accepted method to assess clinical competencies but are often resource-intensive. This study aimed to evaluate the feasibility and effectiveness of a virtual reality (VR) station compared to traditional physical stations in an already established curricular OSCE.

Methods: Fifth-year medical students participated in an OSCE that included ten stations in total, with one station dedicated to emergency medicine offered in two formats and featuring scenarios of septic and anaphylactic shock in each format. Participants in the study were randomly divided into two groups, participating either in the virtual-reality station (VRS) or the physical station (PHS). Student performance and item characteristics were analyzed focusing on the one emergency and the five other case-based stations; four technical-skills-oriented stations were excluded from this study. Student perceptions were recorded as part of a post-examination online survey to assess the acceptance and usability of VR.

Results: Following randomization and exclusions of invalid datasets, 57 and 66 participants were assessed for the VRS and PHS, respectively. The two VRS scenarios (septic and anaphylactic shock) integrated well and demonstrated a balanced level of difficulty (P = 0.67 and 0.58, respectively) with an average difficulty of 0.68 across all stations. They exhibited above-average values with respect to item discrimination (r' = 0.40/0.33, overall = 0.30) and discrimination index (D = 0.25/0.26, overall = 0.16). VRS participant responses emphasized the realistic portrayal of medical emergencies and the fair assessment conditions provided. However, there was some hesitancy towards its broader application in future practical assessments, highlighting the need both for further familiarization as well as maintaining physical interaction with simulated patients.

Discussion/Conclusion: Integration of the VRS into the current OSCE framework proved feasible both technically and organizationally, even within the strict constraints of short examination phases and schedules. The VRS was accepted and positively received by students across various levels of technological proficiency, including those with no prior VR experience. Notably, the VRS demonstrated comparable or even superior item characteristics, particularly in terms of discrimination power. While challenges remain, such as technical reliability and some acceptance concerns, VR remains promising in applications of clinical competence assessment.

Keywords:
Virtual reality (VR), Objective structured clinical examination (OSCE), medical education, technological proficiency, assessment of clinical competence, item characteristics, discrimination power, student acceptance, technical feasibility.

1. Introduction

Objective structured clinical examinations (OSCEs), first described in 1975 [1], have long since been recognized and accepted as a reliable, valid, and objective method to assess clinical competencies in medical education. They are organized in a circuit format and employ stations featuring standardized cases and predefined assessment criteria using checklists or global rating scales to objectify the evaluation by the assessor. By breaking down the clinical tasks into multiple subtests, various skills aligned with learning objectives can be evaluated simultaneously [2]. Presently, this examination format is administered using either standardized patients (SPs) or simulators. SPs are trained actors who are provided with specific cases and then present themselves as patients to students [3]. Students, using this method, can exhibit their proficiency at the third level of Miller's competency framework “shows how” [4]. This surpasses traditional formats such as oral or written examinations, which generally focus on Miller’s second competency level, “knows how”, linked to the application of knowledge. Nevertheless, OSCEs are subject to a number significant limitations, with one of the primary issues being their resource-intensive nature, both in terms of materials and personnel [2]. In addition, the time-intensive training and deployment of SPs also cause significant financial expense. In terms of content, it should be added that OSCEs may not comprehensively capture the complexity of emergency scenarios. Employing SPs often proves inadequate at simulating the complex pathophysiology found in living organisms accurately. In particular, healthy actors may struggle to depict diseases in all their nuances, and invasive procedures cannot be executed.

Virtual reality (VR) simulation as a supplementary method in medical education has received a high degree of acceptance among learners and shown promising results in terms of learning outcomes [5,6]. Importantly, the use of VR-based scenarios in assessments may potentially overcome the aforementioned limitations: Virtual patients can display symptoms and findings that realistically represent illnesses, and a computed dynamic physiology can replicate appropriate responses to medication or invasive interventions such as endotracheal intubation or administration of catecholamines. Additionally, digitally assisted assessment offers the possibility of relieving examiners through automated and objective recording of the results. Considering these advantages, employing VR scenarios of medical emergencies in OSCE stations appears promising, but evidence relating to technical reliability and cost effectiveness remains limited [7]. Indeed, to ensure smooth examinations, a sufficient level of hardware/software maturity is essential to avoid interruptions resulting from technical issues. Furthermore, the technically available scenarios must align with the learning and assessment objectives of the respective curriculum.

Given these prerequisites, application examples in this domain are limited. A pilot study did showcase the successful use of VR-based training of cardiopulmonary resuscitation in an

examination setting [8]. Another study reported that skills assessment by using 360° videos delivered via VR head-mounted displays can yield valid outcomes [9]. We also introduced a VR-based simulation training course on complex medical emergencies into the curriculum at our institution in 2020 [10]. It has been continuously refined to meet the above-mentioned technical requirements and is now a suitable candidate for use as an examination tool.

While VR holds promise for medical assessment, research into large-scale OSCEs and technical feasibility is limited. Additionally, only few studies address the didactic requirements, such as test quality and consistency of results. In this study, we aimed to determine whether the theoretical benefits of VR are realized within the tight schedule of an already established routine OSCE in the curriculum with a full cohort of students.

With this basis, we aimed to address the following questions:

1. Whether it is both organizationally and technically feasible to integrate VR-based stations focused on emergencies within an existing curricular OSCE framework
2. Whether VR-based stations display item characteristics (such as item difficulty, item discrimination, discrimination index) comparable to their physical counterparts that test identical content, and
3. How students perceive and to what extent they accept the VR-based stations.

Preprint JMIR Publications

2. Materials and Methods

2.1 VR-based simulation training

STEP-VR (version 0.13b) was used as the VR-based simulation of complex emergencies together with the hardware setup and head-mounted displays essentially as described previously [10].

2.2 Study Design

The study was conducted at a medical school in Germany at the end of the fifth year of study towards the degree of medicine. The already established curricular OSCE was designed traditionally as a circuit with ten stations, with two circuits running parallel to one another to increase throughput and reduce the examination duration in total to two days. Five stations were case-based focusing on a number of specialties, and four were skills-based. Central to the study was the tenth station, the medical emergencies station. This station was the only one available in two separate modes: either a VR-based or a real-world mode, which we designated as VR-based station (VRS) or physical station (PHS), respectively. The VRS and PHS were designed to be visually and functionally as similar as possible (Figure 1). At each station, students were given one minute to read the case description and task, followed by nine minutes to complete the examination. To mitigate the consequences of students potentially sharing information, scenarios in all stations were switched at differing intervals.

Throughout the semester, students received comprehensive preparation for the OSCE, with particular emphasis on its implementation of emergencies in VR. They were given a script detailing the acute treatment of different types of shock. Additionally, a tutorial video was made available to familiarize students with the VR equipment and software. All students participated in STEP-VR as a part of mandatory three-hour small-group sessions. However, during these sessions not all students were active in the VR; some were observing through a screen displaying the first-person perspective. Furthermore, each student had the opportunity to practice using the system beforehand in a voluntary training session.

Figure 1: Representative scenes from the VRS (A) and its physical counterpart PHS (B). The case description and task assignment for the students, case dynamics during simulation (expressed through changes in vital parameters), and functionality of the emergency room environment were identical in both scenarios.

In the VRS and PHS on day one, students encountered a patient with an initial diagnosis of fistulizing Crohn's disease, leading to septic shock. The assessment focused on the managing measures of the 'One Hour Bundle' (i.e., actions to treat sepsis to be executed within one hour) and the decision-making process for either interventional or surgical abscess drainage. On day two, students faced the challenge of stabilizing a patient suffering an anaphylactic shock triggered by the painkiller metamizole. In addition to managing the associated respiratory distress, they were required to advise the patient on measures to prevent recurrence. All medical content was based on

established guidelines and reviewed by experienced faculty members. At the outset of the OSCE, participants were randomly assigned to one of the two parallel and simultaneous circuits. Ultimately, each student had to undertake one of the two scenarios either within the VRS or the PHS. A backup VR setup was always on hand to address any technical issues.

In both the VRS and PHS, students were provided with all the essential information for the case in the task description (as a sign on the door prior to entering the station). This encompassed medical history, physical examination, and diagnostic test results. The subject/goal of the examination was focused on acute treatment (taking immediate actions) and making decisions for the next steps (correct indication for intervention/surgery or providing recommendations). Students were provided with assistance when donning the head-mounted display and controllers. The first-person perspective of the students was transmitted to a screen, allowing assessors to view the students' actions. The performance was rated with standardized candidate assessment forms.

For students assigned to the VRS but who chose not to use VR technology owing to reservations (such as past instances of simulation sickness), a tutor took over the operation of the headset and controllers. The student could observe the scenario on the screen and had to guide the tutor with the appropriate actions and measures through verbal commands. However, data from these students were not included in the study.

The five case-based stations (internal medicine, surgery, family medicine, paediatrics, gynaecology) were comparable to the VRS and PHS. In the case-based stations, the scenarios were changed every half day, so that four different scenarios were utilized for each specialty. These also assessed students' management and clinical decision-making, especially with regard to diagnostic and therapeutic measures. The other four stations, which focused on procedural competence in highly standardized scenarios (e.g. postoperative blood transfusion), were incomparable with the other stations and thus excluded from subsequent analysis in this study. Figure 2 depicts the layout of the OSCE and data collection.

Preprint JMIR Publications

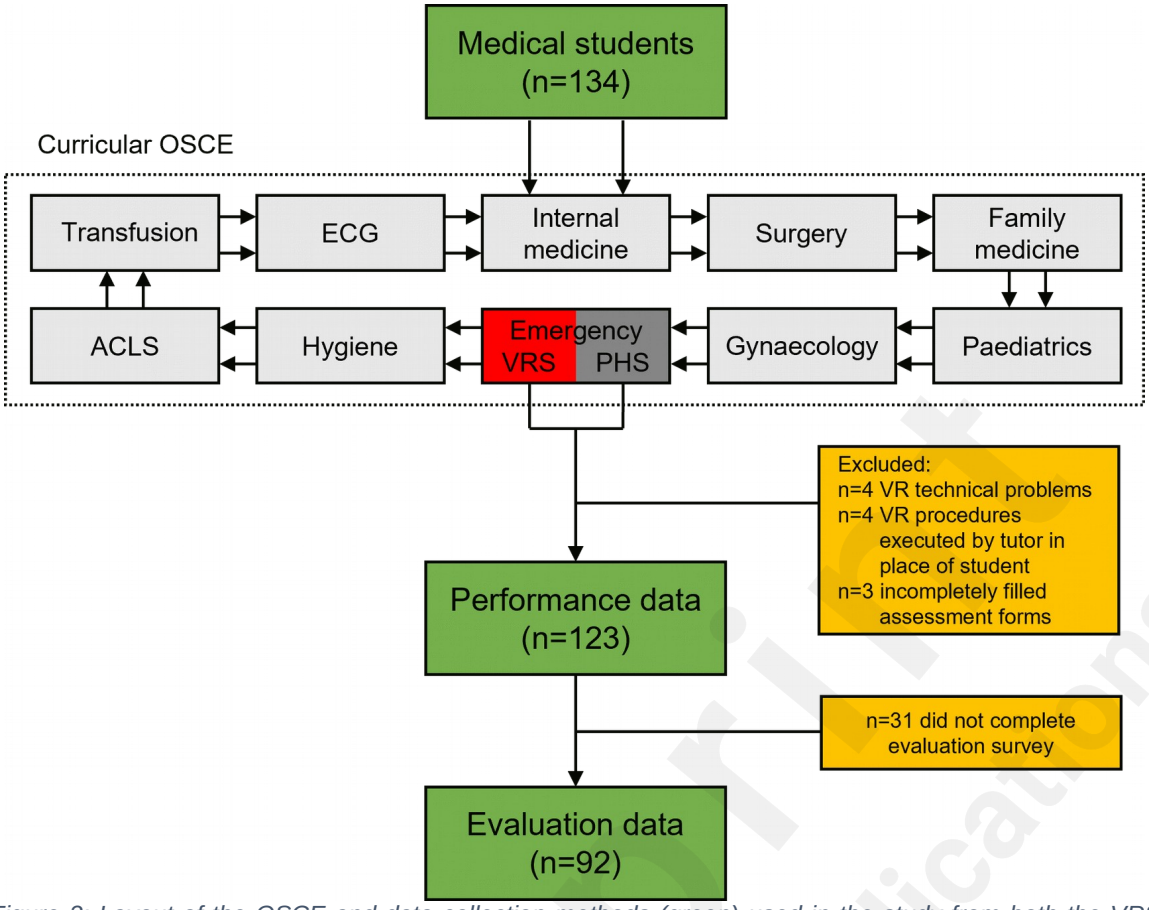


Figure 2: Layout of the OSCE and data collection methods (green) used in the study from both the VRS and PHS. The number of participants, along with those excluded (yellow), is indicated. ECG: electrocardiogram, ACLS: advanced cardiac life support.

2.3 Collection of performance and evaluation data

2.3.1 Assessment of student performance

To assess students in both the VRS and PHS as equally as possible, identical candidate assessment forms were utilized. They consisted of ten items (septic shock) and 13 items (anaphylactic shock) from the categories of 1) additional monitoring and diagnostics, 2) treatment and definitive diagnosis, and 3) subsequent actions and advice. Scoring for each item was either binary (criteria met or not met, corresponding to 1 or 0 points) or ternary (criteria fully met, partially met, or not met, corresponding to 2, 1, or 0 points). In calculating the total sum, all items were equally weighted and normalized to a maximum of 1 point. The candidate assessment forms are outlined in Supplement 1.

2.3.2 Online survey for evaluation and feedback from students

Right after the OSCE, students were invited to participate in an online survey accessible via a QR code to be scanned and performed following the examination. The survey encompassed demographic parameters such as age and prior experience with VR technology. It also featured a total of 18 items, divided into various topics: stress experience (3 items), usability (2 items),

preparation (2 items), acceptance (5 items), subjective performance (3 items), and general rating (3 items to share views on the OSCE's value, its personal relevance, and its future inclusion). Items were rated on a five-point Likert scale from “strongly agree” (5 points) to “strongly disagree” (1 point). Two open-ended questions allowed students to provide qualitative feedback.

2.4 Analysis of item characteristics

The OSCE's item characteristics for both the VRS and PHS across the two scenarios, as well as for the other five case-based stations, each comprising four distinct scenarios, were detailed. Parameters were computed for all OSCE stations, such as difficulty P (average participant score in the station), discrimination r’ (correlation of station scores with overall scores excluding that station), and discrimination index D (difficulty difference between high and low performers based on top and bottom 33rd percentiles) as suggested by Möltner *et al*. [11]). The computation of item statistics was carried our using R 4.3.1 [12]. For the sake of clear presentation, the average was provided for each of the other five case-based stations.

2.5 Analysis of survey data

Student responses from the VRS and PHS were analyzed by employing descriptive statistics such as counts (both absolute and percentages), means, and standard deviation (SD). Nominally scaled variables were compared with the chi-square test. The Wilcoxon rank-sum test was used to compare between the groups (VRS/PHS). A "thematic analysis" method was employed to summarize the open-ended responses from VRS participants only [13].

2.6 Ethics approval

The local institutional review and ethics board judged the project as not representing medical or epidemiological research on human subjects and as such adopted a simplified assessment protocol. The project was approved without any reservation under the proposal number 20230323-03. Survey data from the questionnaires were retrieved anonymously using the EvaSys® platform (Lüneburg, Germany). Students were informed about the study and their participation was voluntary. The decision to participate or not had no consequences on the students’ academic progress. Data were processed and stored in accordance with local data protection laws.

Preprint JMIR Publications

3. Results

3.1 Student participation, performance data, and item characteristics

In total, 134 students participated in the OSCE examination. Eleven students were excluded from the final analysis for various reasons: technical problems with the VR equipment (n = 4), VR procedures executed by a tutor instead of the student (n = 4), and incompletely filled assessment forms (n = 3). Thus, 123 participants completed either the VRS (n = 57) or PHS (n = 66).

The characteristics of items across all stations are presented in Figure 3. The overall average item difficulty P was 0.68. In VRS, septic shock and anaphylactic shock exhibited similar item difficulties at 0.67 and 0.58, respectively, comparable to the same scenarios in PHS (0.71 and 0.64, respectively). Additionally, the mean item difficulty for scenarios from the other five case-based stations consistently aligned at 0.71. Concerning item discrimination, the overall average was determined as r' = 0.30. The VRS scenarios proved to be above average with values of 0.40 (septic shock) and 0.33 (anaphylactic shock), in contrast to then PHS scenarios, for which below-average values of 0.12 and 0.25, respectively, were determined. The scenarios from surgery, pediatrics, and gynecology demonstrated values that were comparable or even better than those for the VRS, while others (internal medicine, family medicine) exhibited significantly lower values. When calculating the discrimination index D, an average of 0.16 was observed across all stations. The VRS scenarios outperformed all other stations, achieving the highest values of 0.25 and 0.26. Conversely, the PHS scenarios fell below this average, with values of 0.10 and 0.12. The other case-based scenarios presented a diverse range of values, mirroring the subject-specific trends seen in item discrimination, with an average of around 0.15.

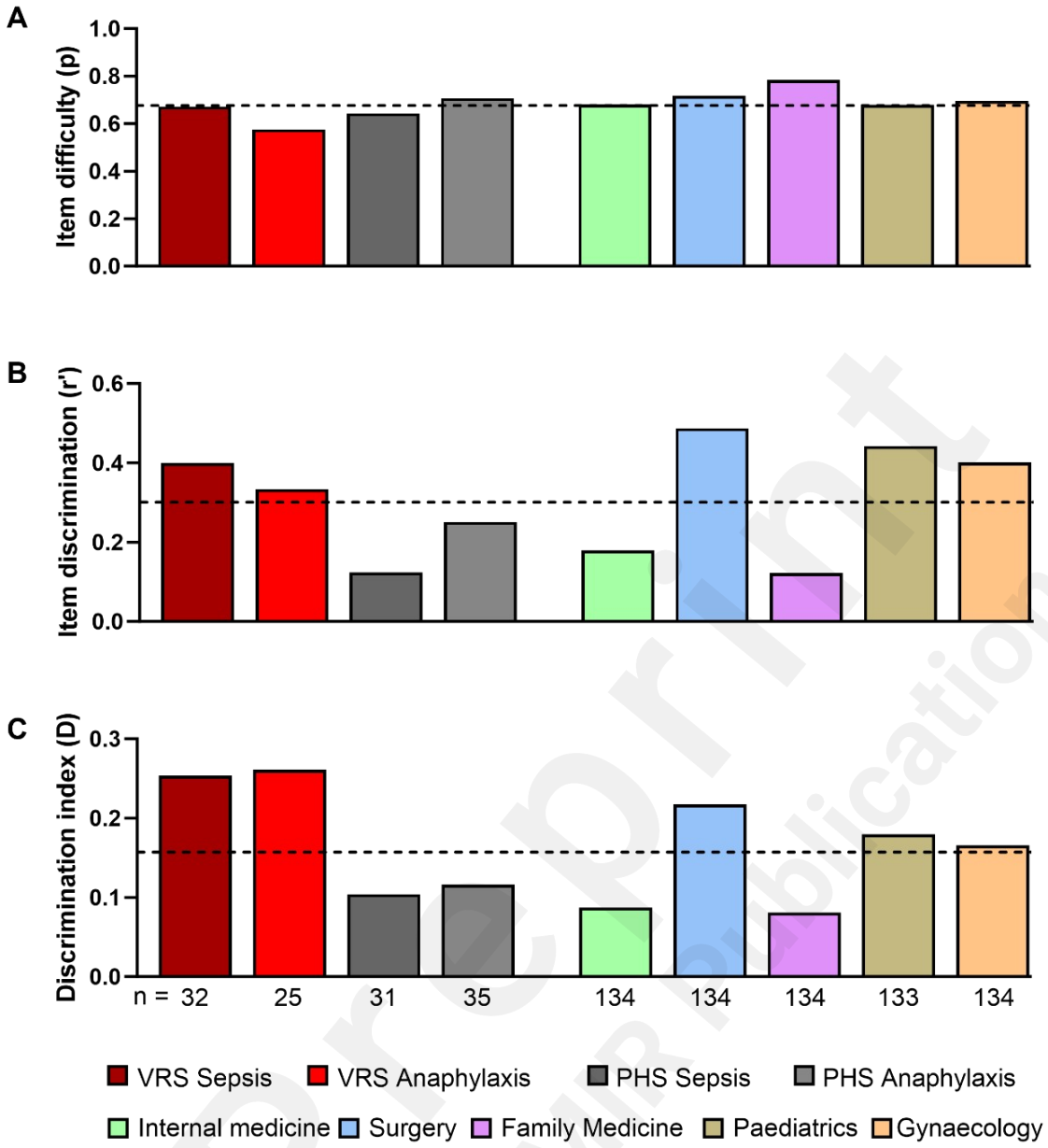


Figure 3: Item characteristics of the OSCE stations for the VRS and PHS in the specific scenarios, as well as the other five case-based stations – each comprising four different scenarios of which the average is shown: item difficulty (A), item discrimination (B), and discrimination index (C). Dashed lines represent the average across all stations. The number of students at each station is provided. sepsis = septic shock, anaphylaxis = anaphylactic shock

3.2 Quantitative and qualitative results of survey data

Ninety-two participants completed the online survey following the OSCE, resulting in a 75% response rate. Table 1 depicts the gender distribution of the participants, as well as their previous experience with emergency medicine and VR. No notable disparities were identified between the VRS and PHS groups.

		Total (n=92)		VRS (n=43)		PHS (n=49)		
Characteristics	**Subgroup**	**n**	**%**	**n**	**%**	**n**	**%**	***P***
Gender	Male	39	42	18	42	21	43	.86
	Female	52	57	25	58	27	55	

	Diverse	1	1	0	0	1	2	
Experience in emergency medicine (e.g., volunteer service)?	Yes	20	22	10	23	10	20	.74
	No	72	78	33	77	39	80	
Cumulative experience with VR applications	None	23	25	14	33	9	18	.29
	0-5h	67	73	28	65	39	80	
	6-10h	2	2	1	2	1	2	
	>10h	0	0	0	0	0	0	
Utilization of VR Lab for preparation	Yes	28	30	13	30	15	31	.97
	No	64	70	30	70	34	69	

Table 1: Participant gender and experience with emergency medicine, as well as characteristics of experience with VR categorized by total, VRS and PHS.

Students rated their perceptions and attitudes towards the VRS or PHS. (Table 2). Moderate scores were recorded for the perception of stress, with no significant difference between the VRS and PHS. Usability was scored favorably, and both formats were viewed as effective in demonstrating acquired skills. Students felt they only had a moderate level of preparation from the general curriculum for the station they completed. However, the specific materials provided enhanced their readiness. Both students from the VRS and PHS groups found their respective scenario as a realistic portrayal of medical emergencies and clinically relevant. They rated the scenarios as manageable within the given time. Students believed that the station allowed them to demonstrate their skills and felt they handled the station well. Overall, students found the OSCE as an examination format in medical school worthwhile. They perceived the type of examination station they completed as meaningful. Notably, compared to the VRS, there was a significant preference among students for the increased use of the PHS in future assessments.

	VRS (n=43)		**PHS (n=49)**		
Theme / Item	**Mean**	**SD**	**Mean**	**SD**	***p***
- Stress					
1. I felt stressed because many aspects of the scenario were beyond my control.	2.28	1.26	2.51	1.19	.30
2. I felt stressed because I lacked sufficient medical knowledge to handle the case.	2.56	1.14	2.39	1.06	.61
3. The presence of the examination staff put pressure on me.	1.61	0.98	1.71	1.17	.87
- Usability					
4. The equipment worked without any issues.	3.63	1.18	3.74	1.17	.75
5. I was able to perform the medical procedures as I had envisioned.	3.40	1.20	3.37	1.24	.92
- Preparation					
6. I felt adequately prepared for the station through the curriculum lectures and courses.	2.67	1.06	2.69	1.23	.87
7. I felt adequately prepared for the station through the preparation materials.	3.23	1.09	3.61	1.27	.06
- Acceptance					
8. The scenario felt realistic.	3.21	1.32	3.25	1.32	.89
9. I found the scenario manageable within the given time.	3.88	1.12	4.08	0.89	.54

Preprint JMIR Publications

10. The content of the scenario was clinically relevant.	4.47	0.59	4.45	0.82	.68
11. This type of station should be used more frequently as an examination format.	3.07	1.20	3.57	1.40	**.04**
12. Overall, I would rate the station as:	3.58	1.10	3.78	1.10	.32
- Performance					
13. The station provided me with an opportunity to demonstrate my learned skills.	3.61	0.90	3.61	1.22	.70
14. I was able to handle the examination station well.	3.40	0.88	3.53	1.00	.35
15. I would rate my performance on this station as:	3.44	0.80	3.57	0.84	.40
- General Rating					
16. I believe OSCE examinations in medical school are generally worthwhile.	3.49	1.32	3.55	1.21	.92
17. I find the type of examination station I completed to be meaningful.	3.05	1.27	3.53	1.24	.07
18. This type of station should continue to be a part of the OSCE examination.	3.07	1.37	3.65	1.18	**.03**

Table 2: Students' perceptions and attitudes towards the two specific OSCE modalities (VRS and PHS). Items were rated on a five-point Likert scale in which 5 was full agreement. Statistically significant values (p < 0.05) are highlighted in bold.

We subsequently performed a thematic analysis on the answers to the open-ended questions from the VRS participants to provide some context to the quantitative findings and to gain deeper insights into students' perspectives on using VR technology in examinations. From the open-ended responses, we identified both positive and negative themes (Table 3) encompassing usability, difficulty/fairness, preparation, practical relevance, and general feedback. Of note, there were nearly twice as many positive (n=58) as negative comments (n=30). Students predominantly praised the realism and fairness the examination presented. These themes also emerged less frequently in a negative context, often from students who faced challenges with time constraints or found some interactions in VR as abstract. Eight participants pointed out difficulties with the technology, experiencing minor issues or challenges. Other notable positive feedback included a generally positive outlook towards the VRS and a strong appreciation for the relevance of the medical content.

Theme	n	%	Quote
Positive themes			
Realism	20	34%	"The setting was realistic and provided ample room for action."
Fairness	17	29%	"The station was manageable within the time frame. There was [technical] assistance available when needed."
General	7	12%	"I had fun during the examination."
Relevance	6	10%	"Responding in an emergency situation is a highly relevant content."
Usability	5	9%	"The operation was reliable and intuitive."
Preparation	3	5%	"The preparation served as good practice."
Total	**58**	**100%**	
Negative Themes			
Fairness	9	30%	"I would have needed more time and more support."
Realism	9	30%	"The medications and procedures could be made somewhat more realistic."

Usability	8	27%	"The handling could be optimized. Additionally, I experienced technical disruptions, which could also be addressed."
Preparation	2	7%	"Better preparation through teaching is needed for medical emergencies."
General	2	7%	"The VR station should not be part of a graded examination."
Relevance	0	0%	-
Total	**30**	**100%**	

Table 3: Summary of positive and negative feedback on VRS from open-ended questions. Multiple responses were allowed.

Preprint JMIR Publications

4. Discussion

In our study, we incorporated a VR-based station specifically designed to accommodate two intricate emergency scenarios into the already established OSCE aimed at advanced medical students at our institution. This integration was carefully planned and executed, involving the adaptation of VR technology to simulate complex emergency scenarios in a controlled educational environment. The VR stations were designed to augment the existing OSCE framework, ensuring compatibility with the current educational objectives and assessment criteria of clinical competence.

Integration of the VRS into the current OSCE framework proved feasible both technically and organizationally. The effective participation of 134 students in the OSCE, with 57 students completing the tasks on the VRS without substantial issues clearly supports this. This study demonstrates that the VRS can be implemented both practically and efficiently [14], even within the strict constraints of short examination phases and schedules.

Analysis of all the data revealed that the item statistics for the VRS were not only comparable to, but in some cases even superior to those of the PHS and the physical case-based scenarios from the five medical disciplines. Such good item characteristics and high discrimination values align well with observations from other studies. Recently, the efficacy of VR-mediated 360° videos in differentiating various skill levels during assessments was highlighted [9]. Studies into telemedicine examinations during the COVID-19 era also suggest that the validity and reliability of digital examination tools can be strong [15,16]. Of note, the discrimination r' and discrimination index D of the VRS were found to be superior to those of the PHS, despite the randomization of participants and rotation of examiners at the stations. One possible explanation for these findings is the more uniform setting offered by VR. This environment minimizes or even eliminates the variability [17] that might be introduced unintentionally by simulated patients, which can occur even when SPs are well-trained and experienced in their roles. Furthermore, the possibility exists that operating the VRS itself may constitute an additional task that high-performing students are more adept at handling, which could lead to greater differentiation in performance at these stations. In other words, there might be confounding variables (e.g. spatial ability) related to achievement in VR environments [18], which could potentially enhance performance outcomes. This hypothesis warrants further investigation, such as examining the correlation between VR handling skills and overall academic performance. Such insights could prove

valuable in the design of more complex digital examinations and case-based assessments.

The study clearly demonstrates that students generally responded positively to the VRS, indicating a favorable attitude and a willingness to engage with and benefit from VR technology. The realism and content relevance of the VR scenarios received above-average ratings and were frequently commended in the qualitative feedback. Fair assessment conditions were emphasized, and even stress experienced during the examinations was considered as manageable. This aligns with recent studies indicating that students exhibit a positive attitude [19] towards VR-based teaching and assessment. Furthermore, they perceive VR as engaging and immersive, affecting learning outcomes positively [9].

Presumably, owing to the lower level of technical refinement in its implementation, another study discovered that students were still more likely to accept VR in classroom settings as opposed to its use in practical assessments [20,21]. Of note, the VRS here were accepted by students across various levels of technological proficiency, including those with no prior VR experience. Interestingly, a substantial majority (70%) of the students chose not to utilize the offers of extra preparation towards the VR examination. Moreover, for 25% of the participants, this was their first-ever experience with a VR application. These findings highlight the viability of VR as an examination tool, accommodating students with a wide range of familiarity with technology. This study did not concentrate on the viewpoint of assessors; another study has explored the feasibility and benefits of using specific VR scenarios in OSCEs, receiving positive evaluation from assessors [15].

Nevertheless, a degree of hesitancy among participants was recorded regarding whether VRS should be used in future examinations: agreement with this statement was clearly less than that for the PHS. This aligns with findings from another study, which indicated that students' reservations were primarily due to their lack of experience with VR technology [22]. The referenced study concluded that practical examinations utilizing VR should only be considered once the technology is firmly established and has demonstrated reliability in educational contexts. Notably, some open-ended comments in our study expressed concerns regarding the potential increase in replacing human assessors and SPs with technology, aligning with findings from prior research [23]. A strategy to counteract this could involve clear communication that VR simulation is intended merely as an additional option to complement, not merely replace, existing examination formats. Reservation and reliance on technology during the early stages of implementation were also the reason why we still opted for manual recording during data collection, regardless of the fact that the VR software provided the capability of automated performance evaluation through a checklist.

Nonetheless, the potential for using such automated features in the future to aid assessors in their demanding role is promising. This approach could represent a significant advancement in the efficiency and objectivity of future assessments, including approaches for formative feedback [24].

Strengths of the study

One advantage of the study is the utilization of hardware and software that has been undergoing continuous evaluation in learning contexts since 2018, providing a high level of realism with minimal simulation sickness. Students were provided with enough practice opportunities in the VR environment beforehand to minimalize operational issues (as demonstrated by good usability results). Nevertheless, the technology needs further rigorous development and enhancement to avoid issues in the examination context. Another strength of this study is the relatively large number of participants drawn from an entire semester cohort, which should generally be viewed as representative of the entire medical student population, including those with critical perspectives. The study's greatest strength, however, lies in the comparison of item statistics and questionnaires directly between the VRS and PHS. This approach allows for the separation of effects related to scenario content and such related to modality, which can influence the overall acceptance of and performance within the examination. Additionally, comparing results with other medical disciplines in the same examination aids in assessing the overall performance of the students.

Limitations of the study

One clear limitation is the restriction to a single institutional site and medical discipline. This is especially relevant in practical training, for which curricula can vary greatly across different faculties, thus limiting study generalizability. Another weakness is the absence of inferential statistical analysis that would correlate performance data with survey results. This was a consequence of the voluntary and anonymous nature of the questionnaires. Making participation in the survey mandatory may have overwhelmed the students and potentially affected the acceptance of the VRS. Nevertheless, this analysis should be conducted in the future to identify students facing specific challenges with the VRS.

Conclusions

Our study successfully demonstrated that complex, VR-based assessment scenarios can be integrated into an established curricular OSCE. Compared to a similar physical

examination station, we noted favorable item characteristics. The degree of acceptance by students was high and we encountered no systematic issues that would preclude the widespread adoption of VR in the assessment of clinical competence.

Preprint JMIR Publications

Acknowledgements

We would like to thank Andrew Entwistle for his critical language review of the manuscript.

Funding

Funding for this work was provided by the German nonprofit “Stiftung Innovation in der Hochschullehre” (Foundation for Innovation in Higher Education) (Grant number FRFMM-776/2022).

Conflict of interest

Tobias Mühling was involved in the software development process of STEP-VR.

Preprint JMIR Publications

References

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975;1(5955):447-451. PMID:1115966
2. Barman A. Critiques on the Objective Structured Clinical Examination. Ann Acad Med Singap 2005;34(8):478-482. PMID:16205824
3. Peters T, Sommer M, Fritz AH, Kursch A, Thrien C. Minimum standards and development perspectives for the use of simulated patients - a position paper of the committee for simulated patients of the German Association for Medical Education. GMS J Med Educ 2019;36(3):Doc31. PMID:31211226
4. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(9 Suppl):S63-7. PMID:2400509
5. Hyeon-Young Kim and Eun-Young Kim. Effects of Medical Education Program Using Virtual Reality: A Systematic Review and Meta-Analysis.
6. Liu JYW, Yin Y-H, Kor PPK, Cheung DSK, Zhao IY, Wang S, Su JJ, Christensen M, Tyrovolas S, Leung AYM. The Effects of Immersive Virtual Reality Applications on Enhancing the Learning Outcomes of Undergraduate Health Care Students: Systematic Review With Meta-synthesis. J Med Internet Res 2023;25:e39989. PMID:36877550
7. Thakker A, Devani P. Is there a role for virtual reality in objective structured clinical examinations (OSCEs)? MedEdPublish 2018;8:180. doi:10.15694/mep.2019.000180.1
8. Manuel Rodríguez-Matesanz, Carmen Guzmán-García, Ignacio Oropesa, Javier Rubio-Bolivar, Manuel Quintana-Díaz and Patricia Sánchez-González. A New Immersive Virtual Reality Station for Cardiopulmonary Resuscitation Objective Structured Clinical Exam Evaluation.
9. Knudsen MH, Breindahl N, Dalsgaard T-S, Isbye D, Mølbak AG, Tiwald G, Svendsen MBS, Konge L, Bergström J, Todsen T. Using Virtual Reality Head-Mounted Displays to Assess Skills in Emergency Medicine: Validity Study. J Med Internet Res 2023;25:e45210. PMID:37279049
10. Mühling T, Späth I, Backhaus J, Milke N, Oberdörfer S, Meining A, Latoschik ME, König S. Virtual reality in medical emergencies training: benefits, perceived stress, and learning success. Multimedia Systems 2023;29(4):2239-2252. doi:10.1007/s00530-023-01102-0
11. Möltner A, Schellerg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Zeitschrift für Medizinische Ausbildung 2006;23(3):11.
12. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria 2023 URL: https://www.R-project.org/.
13. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;3(2):77-101. doi:10.1191/1478088706qp063oa
14. Lee JS. Implementation and Evaluation of a Virtual Reality Simulation: Intravenous Injection Training System. Int J Environ Res Public Health 2022;19(9). PMID:35564835
15. Oscar Arrogante, Eva María López-Torre, Laura Carrión-García, Alberto Polo and Diana Jiménez-Rodríguez. High-Fidelity Virtual Objective Structured Clinical Examinations with Standardized Patients in Nursing Students: An Innovative Proposal during the COVID-19 Pandemic.
16. Lan Y-L, Chen W-L, Wang Y-F, Chang Y. Development and preliminary testing of a virtual reality measurement for assessing intake assessment skills. Int J Psychol 2023;58(3):237-246. PMID:36720650

17. Wendling AL, Halan S, Tighe P, Le L, Euliano T, Lok B. Virtual humans versus standardized patients: which lead residents to more correct diagnoses? Acad Med 2011;86(3):384-388. PMID:21248598

18. Lee EA-L, Wong KW. Learning with desktop virtual reality: Low spatial ability learners are more positively affected. Computers & Education 2014;79:49-58. doi:10.1016/j.compedu.2014.07.010

19. Mahling M, Wunderlich R, Steiner D, Gorgati E, Festl-Wietek T, Herrmann-Werner A. Virtual Reality for Emergency Medicine Training in Medical School: Prospective, Large-Cohort Implementation Study. J Med Internet Res 2023;25:e43649. PMID:36867440

20. Wang N, Abdul Rahman MN, Lim B-H. Teaching and Curriculum of the Preschool Physical Education Major Direction in Colleges and Universities under Virtual Reality Technology. Comput Intell Neurosci 2022;2022:3250986. PMID:35310594

21. Tsekhmister YV, Konovalova T, Tsekhmister BY, Agrawal A, Ghosh D. Evaluation of Virtual Reality Technology and Online Teaching System for Medical Students in Ukraine During COVID-19 Pandemic. Int. J. Emerg. Technol. Learn. 2021;16(23):127-139. doi:10.3991/ijet.v16i23.26099

22. Walter S, Speidel R, Hann A, Leitner J, Jerg-Bretzke L, Kropp P, Garbe J, Ebner F. Skepticism towards advancing VR technology - student acceptance of VR as a teaching and assessment tool in medicine. GMS J Med Educ 2021;38(6):Doc100. PMID:34651058

23. Talbot T, Rizzo A“. Virtual Human Standardized Patients for Clinical Training. In: Rizzo A“, Bouchard S, editors. Virtual Reality for Psychological and Neurocognitive Interventions. New York, NY: Springer New York; 2019. ISBN:978-1-4939-9480-9. p. 387–405.

24. Moore AG, Hu X, Eubanks JC, Aiyaz AA, McMahan RP. A Formative Evaluation Methodology for VR Training Simulations. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW): IEEE; 2020. ISBN:978-1-7281-6532-5. p. 125–132.

Supplements

Supplement 1

Candidate assessment form for the station “septic shock”			
Item	**Result**		
1 Monitoring/Diagnostics			
1.1 O_2 monitoring attached	Yes		No
1.2 Blood cultures taken before antibiotic therapy *(partially met: only one pair or sample taken at the same puncture site)*	Fully met	Partially met	Not met
1.3 Arterial blood gas taken	Yes		No
2 Therapy and final diagnosis			
2.1 Rapid volume replacement initiated *(partially met: volume too slow (<500ml/h))*	Fully met	Partially met	Not met
2.2 Catecholamines (norepinephrine) administered if MAP < 65 mmHg under/after fluid replacement *(partially met: catecholamine administration without naming a specific drug)*	Fully met	Partially met	Not met
2.3 Empirical antibiotics administered (1st choice: piperacillin/tazobactam, Meropenem) *(partially met: antibiotic therapy with a 2nd choice drug)*	Fully met	Partially met	Not met
2.4 Measures performed in correct order with volume administration as primary measure	Yes		no
2.5 Correct suspected diagnosis named: Sepsis / septic shock *(partially met: imprecise naming, needs assistance)*	Fully met	Partially met	Not met
3 Further measures and recommendations			
3.1 Surgical consultation requested	Yes		no
3.2 Intensive care unit transfer requested	Yes		no

Supplement Table 1: Candidate assessment form for actions that should be taken during the examination for the scenario “Septic shock”. Scoring for each item was either binary (criteria met or not met, corresponding to 1 or 0 points) or ternary (criteria fully met, partially met, or not met, corresponding to 2, 1, or 0 points). In calculating the total sum, all items were equally weighted and normalized to a maximum of 1 point.

Candidate assessment form for the station “anaphylactic shock”			
Item	**Result**		
1 Monitoring/Diagnostics			
1.1 O_2 saturation monitoring attached	Yes		No
2 Therapy and final diagnosis			
2.1 Allergen exposure stopped *(partially met: not as first measure)*	Fully met	Partially met	Not met
2.2 Epinephrine administered (0.15 to 0.6 mg IM / 1 µg/kg body weight IV) *(partially met: correct dosage not stated)*	Fully met	Partially met	Not met
2.3 Oxygen administered (5-12 L via mask) *(partially met: insufficient flow rate)*	Fully met	Partially met	Not met
2.4 Volume therapy initiated (> 500 mL/h) *(partially met: insufficient flow rate)*	Yes		no
2.5 Salbutamol administered	Yes		no
2.6 Antihistamine administered	Yes		no
2.7 Corticosteroid administered	Yes		no
2.8 Structured approach with prioritization of administration of epinephrine and oxygen over further medication	Yes		no
2.9 Correct suspected diagnosis named: Anaphylaxis / anaphylactic shock *(partially met: imprecise naming, needs assistance)*	Fully met	Partially met	Not met
3 Further measures and recommendations			
3.1 Monitoring for 24h recommended	Yes		no
3.2 Emergency kit prescribed, recurrence risk mentioned	Fully met	Partially	Not met

(partially met: only one aspect)		met	
3.3 Referral for allergological assessment advised	Yes		no

Supplement Table 2: Candidate assessment form for actions that should be taken during the examination for the scenario "Anaphylactic shock". Scoring for each item was either binary (criteria met or not met, corresponding to 1 or 0 points) or ternary (criteria fully met, partially met, or not met, corresponding to 2, 1, or 0 points). In calculating the total sum, all items were equally weighted and normalized to a maximum of 1 point.

Supplement 2

	N	Item difficulty	Discrimination	Discrimination Index
VRS Sepsis	32	0.67	0.40	0.25
VRS Anaphylaxis	25	0.58	0.33	0.26
PHS Sepsis	31	0.64	0.12	0.10
PHS Anaphylaxis	35	0.71	0.25	0.12
Internal medicine	134	0.68	0.18	0.09
Surgery	134	0.72	0.49	0.22
Family medicine	134	0.78	0.12	0.08
Paediatrics	133	0.67	0.44	0.18
Gynaecology	134	0.70	0.40	0.17
Mean		0.68	0.30	0.16

Supplement Table 3: Item statistics comprising item difficulty, item discrimination and discrimination Index of VRS, PHS, and the other stations form medical disciplines. The statistics for the latter are averaged from four examination stations each. sepsis = septic shock, anaphylaxis = anaphylactic shock

Preprint JMIR Publications

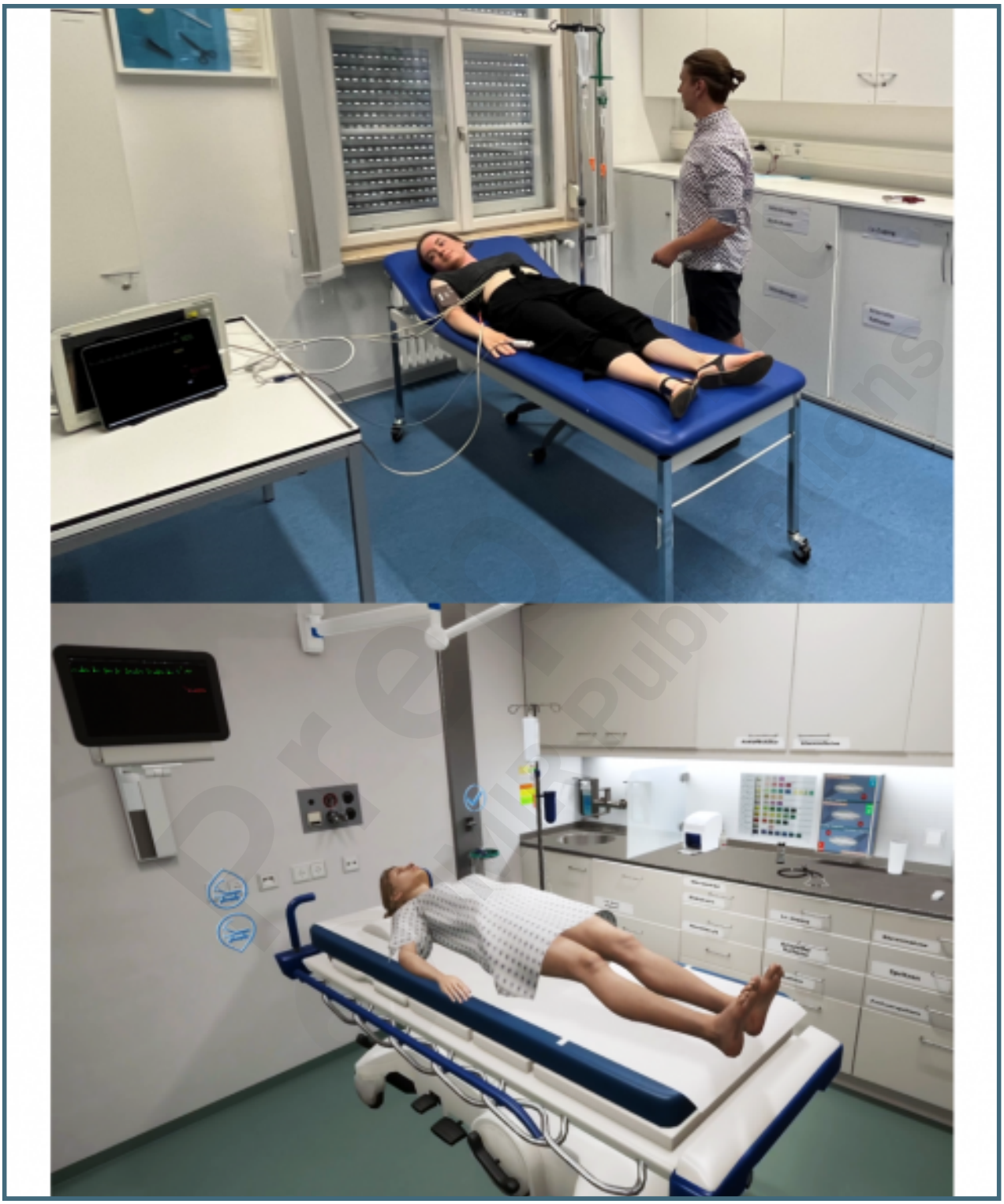
Supplementary Files

Preprint JMIR Publications

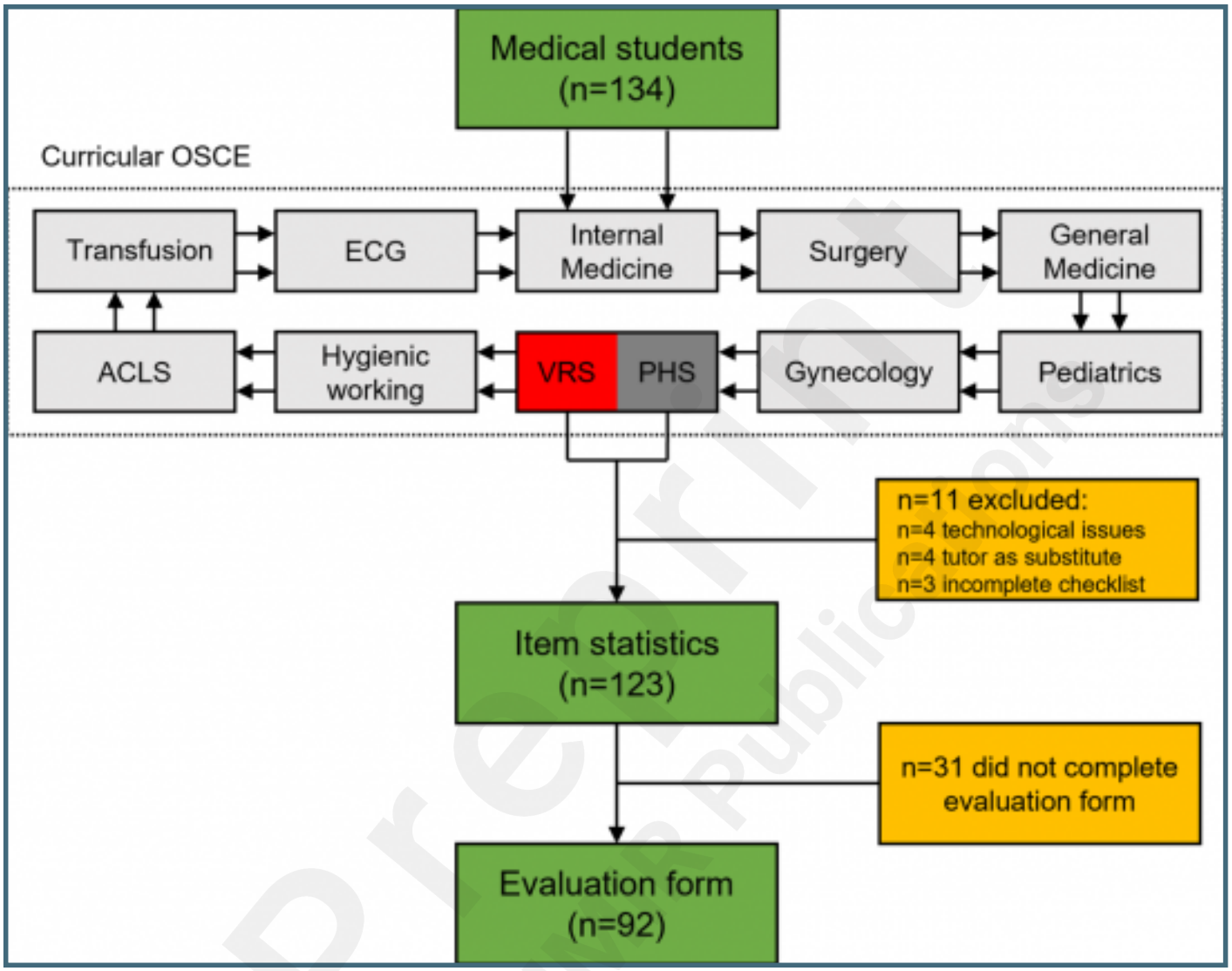
Figures

Preprint JMIR Publications

Representative scenes from the VRS (A) and its physical counterpart PHS (B). The case description and task assignment for the students, case dynamics during simulation (expressed through changes in vital parameters), and functionality of the emergency room environment were identical in both scenarios.



Layout of the OSCE and data collection methods (green) used in the study from both the VRS and PHS. The number of participants, along with those excluded (yellow), is indicated. ECG: electrocardiogram, ACLS: advanced cardiac life support.



Item characteristics of the OSCE stations for the VRS and PHS in the specific scenarios, as well as the other five case-based stations – each comprising four different scenarios of which the average is shown: item difficulty (A), item discrimination (B), and discrimination index (C). Dashed lines represent the average across all stations. The number of students at each station is provided. sepsis = septic shock, anaphylaxis = anaphylactic shock.