

Who carries the load? Role-specific cognitive load and team performance in a multi-role virtual reality emergency simulation: an exploratory pilot study of collaborative clinical reasoning in undergraduate medical education

Julius Josef Kaminski

Charité - Universitätsmedizin Berlin

Dorothea Penders

Charité - Universitätsmedizin Berlin

Robert Vortriede

Charité - Universitätsmedizin Berlin

Sebastian Baer

Charité - Universitätsmedizin Berlin

Robin Schmid

Charité - Universitätsmedizin Berlin

Mareen Machner

mareen.machner@charite.de

Medizinische Universität Lausitz - Carl Thiem

Research Article

Keywords: Virtual reality, Clinical reasoning, Collaborative Clinical Reasoning, Medical education, Emergency medicine, Simulation, Team performance, Collaborative cognitive load theory

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10671015/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Clinical reasoning is increasingly understood as a collaborative, distributed team activity, yet virtual reality (VR) simulation in undergraduate medical education remains predominantly single-user based. Guided by Collaborative Cognitive Load Theory, we developed a structured peer-assisted learning and team-based VR emergency course for one VR-operating student and up to five peers to occupy complementary functional roles (Team Lead, two Guideline Research roles, Documentation, Handover). The aim of this exploratory study was to examine the distribution of cognitive load across the six functional roles, the relation between a role's cognitive load and overall team performance, and the relation between cognitive load and individually perceived competence gain.

Methods

In a single-arm exploratory pilot study, 44 undergraduate medical students participated in nine peer-led, VR sessions with rotating roles. Perceived workload was measured with the NASA Task Load Index (NASA-TLX) in 218 single-role observations and analysed with linear mixed-effects models. Team performance was rated with the Team Emergency Assessment Measure (TEAM), anchored to each participant's first Team Lead rotation, and related to workload using Spearman correlations. Self-assessed competence was rated on five items before and after the sessions and analysed using Wilcoxon signed-rank tests (Holm-corrected).

Results

All five self-assessment items improved significantly (all adjusted $p \leq .001$; rank-biserial $r_{rb} = 0.72-1.00$). Workload differed strongly by role ($\chi^2(5) = 71.98, p < .001$). After removing VR-specific physical demand, the Team Lead carried the highest cognitive load, and the two guideline-research roles diverged significantly, with procedural sequencing more demanding than pharmacological research. Higher Team Lead workload was associated with lower observer-rated team performance ($\rho = -0.37, p = .018$), whereas workload in the other roles showed only weak, non-significant correlations with performance. Individual workload showed no clear association with perceived competence gain.

Conclusions

This study examined a resource-conscious multi-role VR emergency simulation format in undergraduate medical education, associated with competence gains and functionally interpretable, role-specific workload profiles. The Team Lead role showed the highest cognitive demands in distributed reasoning,

suggesting targeted instructional support for this role and offering a measurement model for studying collaborative clinical reasoning in team training.

Background

Clinical reasoning has traditionally been conceptualised and taught as an individual cognitive activity: a learner gathers information, generates diagnostic hypotheses, weighs alternatives, and decides on management. This perspective has contributed substantially to medical education, particularly in understanding expertise development, diagnostic error, and the teaching of reasoning strategies [1]. Emergency and acute care, however, are a good example of how this framing has limits: Decisions must be taken under time pressure and on an uncertain and incomplete evidential basis, and the relevant information is rarely held by any one person at any one moment. Clinical understanding emerges instead while team members exchange observations, retrieve information, document findings, interact with artefacts of the clinical environment, prioritise actions, and coordinate treatment. Accordingly, clinical reasoning has been re-conceptualised as a situated activity in which cognition takes place also outside the individual mind in dynamic systems [2] in team settings, what has been termed collaborative clinical reasoning (CCR) [3–5]. This conceptualisation of clinical reasoning has implications for undergraduate medical education: If CCR is a learnable practice through which teams construct shared clinical understanding in real time [5], learners require opportunities to practise not only individual diagnostic thinking but also the interactional and organisational work through which clinical understanding is shared: communicating observations, aligning interpretations, using guidelines, maintaining a shared working picture, and translating reasoning into coordinated action. Emergency care is a particularly suitable educational context for this, because the conditions described above make collaborative reasoning both visible and necessary. Interestingly, the question of how cognitive demands are actually distributed when reasoning is taking place as distributed practice, is widely under-explored in undergraduate medical education.

Virtual reality (VR) based simulation is a particularly suitable setting in which to rehearse collaborative practices: through immersive patient environments, standardised scenario progression, and emergency decision-making under time pressure, it provides a controlled and repeatable environment before learners encounter high-stakes clinical situations. To date, VR in undergraduate medical education is predominantly used for anatomy and surgery teaching and overwhelmingly in single-user form [6, 7]. Where collaborative configurations exist, they are often vicarious, with team members observing a single player and learning through shared debriefs (as, e.g., described by Mørk et al., (2024)), interprofessional as opposed to monoprofessional [9], and focused on circumscribed procedural skills such as life support [10]. However, monoprofessional VR simulation formats in which one student is immersed while peers take on complementary roles within the same scenario offer several advantages for study purposes: they are resource-friendly as they require only a single headset, can often be integrated into existing teaching routines, and turn the practical constraint that immersive VR is experienced by one user at a time into an opportunity to explore instructional designs in clinical reasoning education.

The format described in this work applies CCR principles to role design, distributing distinct parts of clinical reasoning processes across rotating, functionally complementary roles as detailed in the methods section. The theoretical rationale for role-distributed instructional VR designs is provided by Collaborative Cognitive Load Theory (CCLT) [11], which extends cognitive load theory from individual to collaborative learning. CCLT holds that when a task exceeds an individual learner's working-memory capacity, the interacting information elements can be distributed across the working memories of several group members as a collective working memory, making collaboration advantageous for complex tasks such as emergency management. This advantage comes at the price of transactive costs: communication and coordination are required to share and integrate distributed information, and these activities themselves consume working-memory resources [11]. CCLT further predicts that pre-assigned team roles reduce these coordination costs by clarifying responsibilities and, accordingly, predicts systematic differences in load between functional roles, with transactive costs likely to concentrate in positions where distributed information must be integrated. This makes perceived workload an informative measure of how a VR scenario's cognitive and operational demands are distributed across the team. Workload is also relevant to team performance: high workload in the team lead role may reflect the cognitive cost of coordinating distributed reasoning but may impair the team's ability to maintain direction and anticipate required actions. Finally, cognitive load is consequential for learning itself, as excessive load can impair learning and performance whereas load arising from productive engagement with the learning task may support it [12–17].

Following from these considerations, the aim of this study was to contribute to the design and evaluation of resource-conscious VR simulation formats that support collaborative clinical reasoning in undergraduate emergency medicine training by linking role-specific workload, team performance, and self-assessed competence. Four research questions were addressed:

1. Did perceived workload differ across the six functional roles, using role-specific NASA-TLX profiles as indicators of how task demands were distributed?
2. How was the Team Lead role's workload associated with observer-rated team performance?
3. Did students' self-assessed emergency care competence change after participation in the peer-led small-group teaching?
4. Was an individual's workload during the training sessions associated with their perceived gain (or loss) in competence, both overall and specifically for the Team Lead role?

Methods

Setting and participants

We conducted a single-arm, exploratory pilot study of a collaborative, team-based virtual reality (VR) emergency training, led by trained medical peer students. The format used the STEP.VR application, developed at the University Hospital of Würzburg [18] in a joint process with the developing agency

OrangeWhip Interactive UG (Munich), run on a local computer and operated through Meta Quest 3 headset.

At our institution, we co-developed and evaluated with students a training setup for up to six student learners, taking on different roles with circumscribed tasks and responsibilities as described in Table 1 below. Three guiding ideas led the instructional design: Firstly, to lower the complexity of tasks didactically that a practising physician in emergency settings would usually perform simultaneously. Secondly, to create roles that mirror different parts of clinical reasoning processes and are modelled to cover as many facets of the CanMEDS framework [19] as possible (e.g., leader, collaborator, communicator, medical expert). Thirdly, given the importance of Crew Resource Management (CRM) training in emergency situations for reducing errors related to human factors [20], particular emphasis was placed during the development of the course on ensuring that it supports the practice of CRM behaviours, for example through circumscribed tutor interventions such as calls for team-timeouts to foster team collaboration. Roles rotated in 20-minute intervals so that each participant occupied several roles per session; in smaller teams, roles were occasionally combined. Role sequences were not standardised across teams, and scenarios were switched after the group solved them. Participants were allowed to contribute to solving problems another role was mainly working on.

Each 3-hour training session comprised a peer-tutor-led introduction followed by a short immersive competence [21] training and the choice of a starting scenario out of 12 possible scenarios covering typical and frequent emergency medicine problems such as sepsis, chest pain, dyspnoea, shock, etc., without students knowing the diagnoses in advance. Multimedia Appendix 1 shows the setup and gives an impression of the STEP.VR scenario.

Table 1
Description of the six team roles in the VR training scenario

Role	Description
VR Operator	The VR Operator is immersed in a virtual emergency department (ED) environment and navigates it by moving around in the simulation room while wearing a headset and using a controller. They are within earshot of their teammates, who can follow the action in the VR environment in real time on a large projection screen. All team members can hear sounds from the VR environment (e.g. monitor signals, the telephone ringing and patient responses) via loudspeakers.
Team Lead	The Team Lead is responsible for coordinating communication within the team. They are supported in this by the documentation role. For example, during team time-outs, Team Leads summarise the current status of the work, hypotheses and next steps according to CRM principles. The Team Lead's focus is on prioritising, integrating contributions, maintaining direction, and coordinating action.
Guideline Research I	Guideline Research I researches and proposes relevant workflows and SOPs for the diagnosis and management of the problems encountered by the team in the VR scenarios. This may include, for example, consulting evidence-based sources to select the appropriate imaging modality, assessing the clinical significance of certain laboratory parameters, or calculating relevant scores such as the SOFA score. All team members may consult any reference materials and use the internet and apps as required at any point during the simulation; the use of smartphones for that purpose is encouraged.
Guideline Research II	Guideline Research II is focusing specifically on pharmaceutical issues and problems. For instance, the role involves searching databases and specialist literature to determine which pharmaceuticals are appropriate for administration and the factors that must be considered when administering it. The role also involves guiding the calculation of dosages and dilutions, such as those required for preparing syringes for infusion pumps in the VR scenario.
Documentation	The documentation role supports the entire team, particularly the Team Lead, by recording the team's key considerations on a digital whiteboard. This ensures that differential diagnoses and planned next steps, as decided by the team, are visible to everyone.
Handover	The Handover role involves preparing a case presentation and outlining the issues involved. To this end, the role draws on information provided by the Documentation role, condensing it further to provide the team with an overview of the case, including details of when team time-outs take place, and presenting a final overview of the course of treatment when the case is resolved. The use of handover protocols is permitted but not mandatory.

The study was performed at Charité – Universitätsmedizin Berlin, a big European medical school, following a modular curriculum [22] with an intake of about 600 students per year and was executed in accordance with the decision of the Ethics Committee of the Charité – Universitätsmedizin Berlin, Germany (Reference Number: EA4/105/23).

Instruments

Perceived workload (RQ1, RQ2). After each rotation, participants completed the NASA Task Load Index (NASA-TLX) [23], rating the six subscales Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration (each 0–20 points). Following the unweighted Raw-TLX procedure [24], the primary workload outcome was the sum of the six subscales (0–120 points). Since operating the VR headset imposes physical demands unique to the VR Operator role (spatial navigation, controller interaction), a cognitive load index (CLI; NASA-TLX total minus Physical Demand, 0–100 points) was specified to isolate clinical-cognitive workload from VR-interface-specific physical demands.

Team performance (RQ2). Team performance was rated with the Team Emergency Assessment Measure (TEAM) [25], comprising 11 items (0–4 points) across the domains leadership, teamwork, and task management, plus a global performance rating (item 12, 1–10 points). The TEAM total was calculated as the sum of items 1–11 (0–44 points); the global rating was analysed separately and never included in the total. Each participant was rated once, anchored to their first rotation in the Team Lead role, by the session tutor who underwent a rating calibration session in advance. TEAM item scores are reported descriptively at item level and aggregated to the instrument's three subscales: Leadership (items 1–2), Teamwork (items 3–9), and Task Management (items 10–11), with participant-level subscale scores computed as the mean of the constituent items. For the sum of items 1–11, performance was benchmarked against the categories of Cooper and colleagues (2010): ≤ 33 poor, 34–39 good, ≥ 40 excellent team performance.

Self-assessed competence (RQ3, RQ4). Before and after the sessions, participants rated five statements covering familiarity with emergency department procedures, skill-level of the targeted use of clinical guidelines, knowledge of differential diagnosis of emergency conditions, familiarity with Crew Resource Management principles, and confidence in clinical decision-making in emergencies, each on a 4-point agreement scale (1 = strongly disagree, 4 = strongly agree). The five self-assessment items were developed for this study to reflect the scenario's core learning objectives, including applying crew resource management principles, as well as students' prior knowledge and experience. An English-language version of all items used is provided as Supplementary File 3.

Statistical analysis

For **RQ1**, workload was analysed with linear mixed-effects models to account for the non-independence of up to six observations per participant. The primary model regressed the NASA-TLX total on functional role (fixed effect; reference category Team Lead) and rotation slot (1–6, fixed effect controlling for time-on-task), with a random intercept per participant. The overall role effect was tested with a likelihood-ratio test comparing maximum-likelihood fits of the full model against the model without role; coefficients and 95% confidence intervals are reported from restricted maximum-likelihood estimation. The identical model was fitted to the CLI as the secondary outcome, and a sensitivity analysis re-included the combined-role observations as an additional role category from runs in which participants took on more than one role during a rotation. To examine whether workload changed over the course of a session, rotation slot was additionally entered as a categorical predictor (reference: first slot); in a role-specific

exploratory analysis restricted to VR Operator observations, Physical Demand and total workload were regressed on slot with a random intercept per session, because each participant held this role only once, so that any decline across slots would reflect observational rather than within-person learning. The intraclass correlation was used to quantify the proportion of variance attributable to stable between-participant differences. Finally, a Mann–Whitney U test was conducted to determine whether participants with previous VR experience experienced lower physical demand in the VR operator role than those without.

For **RQ2**, the association between the TEAM total and NASA-TLX workload during the first Team Lead rotation was quantified using Spearman's rank correlations among participants with both measures available ($n = 40$). To assess whether this association was specific to the Team Lead episode rather than reflecting a general person-level reporting tendency, the TEAM total score was additionally correlated with each participant's mean workload in every other role as a descriptive discriminant check. Additionally, the convergent association between the TEAM total and the global rating (item 12) was assessed with a Spearman correlation (ρ). Given the single-rater design and the clustering of participants within sessions, all these analyses were considered hypothesis-generating; p-values are uncorrected.

For **RQ3**, pre–post changes in the five self-assessment items were tested with Wilcoxon signed-rank tests (ordinal 4-point scales) in participants with complete pre- and post-ratings on all five items ($n = 41$). Effect sizes are reported as matched-pairs rank-biserial correlations (r_{rb}), and family-wise error across the five items was controlled for using the Holm–Bonferroni procedure. Paired t-tests were used for parametric sensitivity analysis.

For **RQ4**, aggregate change in self-assessed competence was defined as the five-item post-score total minus the corresponding pre-score total (possible range – 15 to + 15) among participants with complete ratings ($n = 41$). Baseline score was rank-correlated with the change score. The between-participant association of change in self-assessed competence with workload was estimated by extending the RQ1 mixed-effects model, retaining workload as the outcome and adding competence change as a participant-level predictor with a fixed coefficient alongside functional role and rotation slot; a random intercept was included for participant. Models were fitted with and without additional adjustment for the five-item baseline pre-score total and repeated using the cognitive load index as a sensitivity outcome. Change in competence was correlated with NASA-TLX workload during the first Team Lead rotation, as measured by Spearman's ρ . All RQ4 analyses were exploratory, and p-values were uncorrected.

All analyses were performed in Python (version 3.12) using pandas (version 2.2.2), SciPy (version 1.16.3), and statsmodels (version 0.14.6); plots were done with the matplotlib (version 3.10.0) package.

Results

Participants and data completeness

Forty-four medical students participated voluntarily in nine training sessions between May and June 2026, in teams of three (two sessions), five (four sessions), or six participants (three sessions). Participants were in their 4th to 12th semester (median 8). Prior VR experience was low: n = 24 (54%) had never used VR, n = 19 (43%) reported rare use (one to five occasions), and no participant reported more than rare prior use with one missing response. Sessions were peer-led by one of two trained student tutors following a standardised tutor manual.

Complete pre- and post-questionnaires were available for 41 participants (93%). Of 233 complete NASA-TLX observations, 13 lacked a role label that could not be reconstructed and two referred to combined-role assignments that were excluded from the primary role model but re-included in the sensitivity analysis; the primary role analyses therefore comprised 218 single-role observations from all 44 participants. TEAM ratings were available for 42 participants (95%); one missing TEAM item was imputed with the rounded item mean of all available item-3 ratings ($133/41 = 3.24$; imputed value: 3). Three participants in one session completed a second full scenario run; in accordance with the analysis plan, only first-run data were analysed.

RQ1: Workload by functional role

Across the 218 single-role observations, NASA-TLX total scores ranged from 5 to 88 (M = 45.3, SD = 18.0). Descriptively, the VR Operator role showed the highest mean total workload (M = 54.7, SD = 15.5), closely followed by Team Lead (M = 53.8, SD = 12.9), with Guideline Research I (M = 46.3, SD = 16.1), Handover (M = 38.9, SD = 12.1), Guideline Research II (M = 36.8, SD = 18.7), and Documentation (M = 31.7, SD = 18.1) at successively lower levels (Table 2, Fig. 1A).

Table 2
NASA-TLX total and cognitive load index (CLI) by role (218 observations, 44 participants).

Role	n	TLX total M (SD)	CLI M (SD)
VR Operator	44	54.7 (15.5)	44.2 (13.5)
Team Lead	44	53.8 (12.9)	52.2 (12.7)
Guideline Research I	50	46.3 (16.1)	45.2 (15.6)
Handover	14	38.9 (12.1)	37.8 (11.6)
Guideline Research II	29	36.8 (18.7)	35.4 (18.0)
Documentation	37	31.7 (18.1)	29.0 (17.0)

Inspection of the NASA-TLX subscale profiles (Fig. 2) revealed that the elevated total score of the VR Operator role was driven predominantly by Physical Demand (M = 10.5), which was four to ten times higher than in any other role (range of role means 1.1–2.7). This pattern reflects the structural demands unique to headset operation – spatial navigation, controller interaction, and menu-based task execution – rather than the clinical-cognitive demands shared across roles. The Team Lead role, by contrast,

showed the highest scores on Mental Demand ($M = 13.2$), Temporal Demand ($M = 10.8$), Effort ($M = 11.1$), and Frustration ($M = 8.9$) of all roles.

To examine whether the elevated Physical Demand of the VR Operator role reflects VR interface unfamiliarity, VR Operator observations were compared by prior VR experience. Participants who had never used VR reported higher mean Physical Demand as VR Operator ($M = 11.4$, $SD = 3.9$, $n = 24$) than those with rare prior use ($M = 9.4$, $SD = 4.5$, $n = 19$), while their total NASA-TLX load was virtually identical ($M = 54.3$, $SD = 18.2$ vs. 54.8 , $SD = 12.1$). Neither difference reached significance in observation-level Mann-Whitney U tests (Physical Demand: $U = 277$, $p = .233$; total NASA-TLX: $U = 226$, $p = .980$), suggesting caution when regarding Physical Demand as an interface artefact specific to prior headset operation experience.

In the primary linear mixed-effects model (NASA-TLX total; fixed effects: role and rotation slot; random intercept per participant), role was a strong predictor of workload, $\chi^2(5) = 71.98$, $p < .001$. Relative to the Team Lead reference, workload was significantly lower for Guideline Research I ($b = -8.90$, 95% CI -14.48 to -3.31), Guideline Research II ($b = -16.59$, 95% CI -23.15 to -10.03), Handover ($b = -18.69$, 95% CI -27.32 to -10.06), and Documentation ($b = -22.60$, 95% CI -28.69 to -16.51), whereas VR Operator did not differ from Team Lead ($b = -0.08$, 95% CI -5.80 to 5.65). The rotation slot effect was small and not significant ($b = -0.90$ per slot, 95% CI -2.03 to 0.24 , $p = .121$), indicating no substantial time-on-task drift across the session. When slot was modelled categorically, workload did not differ between the first five rotation slots (all $p \geq .215$ vs. slot 1), and only the final sixth slot — reached only in six-person teams — showed lower workload than the first ($b = -7.86$, $p = .038$, uncorrected across five contrasts); this isolated end-of-session deviation should be interpreted cautiously. In the VR Operator role specifically, Physical Demand did not decline across the session ($b = 0.63$ per slot, 95% CI -0.16 to 1.41 , $p = .117$; random intercept per session), providing no indication that (interface-related) physical load diminishes through prior observation of other operators. The intraclass correlation of 0.27 indicates that approximately one quarter of the workload variance was attributable to stable between-participant differences.

The secondary model using the cognitive load index (total minus Physical Demand) separated the two high-load roles: with the VR-interface-specific physical component removed, the VR Operator role fell significantly below Team Lead ($b = -8.99$, 95% CI -14.32 to -3.66), leaving Team Lead as the single highest-load role, while the ordering of the remaining roles was preserved (role effect: $\chi^2(5) = 67.04$, $p < .001$; ICC = 0.29) (Fig. 1B). A sensitivity analysis re-including the two combined-role observations left all role coefficients basically unchanged (differences < 0.1 points).

The two guideline-research roles cover functionally distinct cognitive tasks: Guideline Research I addresses the sequencing of diagnostic and management steps, whereas Guideline Research II addresses pharmacological reasoning (medication choice, dosing, indication, contraindication, side effects, and interactions). In a mixed-effects model restricted to these two roles (79 observations from 41 participants, most of whom rated both), the procedural-sequencing role if Guideline Research I

imposed significantly higher total workload than the pharmacological Guideline Research II role (Guideline Research II vs. I: $b = -8.72$, 95% CI -15.72 to -1.72 , $p = .015$). The difference was most pronounced on the Mental Demand subscale ($b = -2.52$, $p = .005$), where the procedural role had a mean of 11.4 compared with 8.8 for the pharmacological role.

RQ2: Team performance and its relation to Team Lead workload

TEAM ratings, anchored to each participant's first Team Lead rotation, were available for 42 participants. TEAM total scores (items 1–11; possible range 0–44) ranged from 24 to 44 ($M = 35.0$, $SD = 4.8$, $Mdn = 34$). According to the performance benchmarks of Cooper and colleagues (2010), 15 participants (36%) fell into the poor (≤ 33), 19 (45%) into the good (34–39), and 8 (19%) into the excellent range (≥ 40). The global impression rating (item 12; scale 1–10; $M = 7.7$, $SD = 1.0$, range 6–10) showed a strong rank correlation with the TEAM total ($\rho = .77$, $p < .001$), supporting the internal consistency of the ratings in this sample.

The item profile (Fig. 3) showed the lowest mean scores for item 1, direction and command by the team leader ($M = 2.79$), and item 9, anticipation of potentially required actions ($M = 2.88$), whereas item 5, composure and control ($M = 3.71$), and item 6, team morale ($M = 3.62$), showed the highest averages. Aggregated to the instrument's three subscales, participant-level scores were lowest for Leadership (items 1–2; $M = 2.94$, $SD = 0.79$), intermediate for Task Management (items 10–11; $M = 3.04$, $SD = 0.62$), and highest for Teamwork (items 3–9; $M = 3.30$, $SD = 0.44$). Leadership was also the most variable subscale (range 0.50–4.00 versus 2.00–4.00 for both other subscales), indicating that teams differed considerably more in observed leadership behaviour than in teamwork or task management. Within the Teamwork subscale, item scores were heterogeneous: anticipation (item 9) fell markedly below the remaining teamwork items, whereas composure and morale approached the ceiling.

For 40 participants, both a TEAM rating and a role-labelled NASA-TLX observation from the first Team Lead rotation were available. In this exploratory analysis, higher workload during the Team Lead rotation was associated with lower observer-rated team performance (NASA-TLX: Spearman $\rho = -0.37$, uncorrected $p = .018$; cognitive load index: $\rho = -0.33$, $p = .036$). Excluding the participant with the imputed TEAM item did not change this result ($\rho = -0.37$, $p = .020$, $n = 39$). The association was specific to the Team Lead episode and was weak and non-significant for workload in every other role (ρ between -0.25 and 0.16 , all $p \geq .153$; $n = 11$ –37 per role).

RQ3: Changes in self-assessed competence

Scores on all five self-assessment items showed statistically significant pre-post increases ($n = 41$ pairs; Table 3). The largest change was observed for familiarity with emergency department procedures (median 2 to 3; $z = -4.86$, $r_{rb} = 1.00$), where all 31 participants whose rating changed reported an increase. Improvements in the remaining items – targeted use of clinical guidelines ($z = -3.21$, $r_{rb} = 0.72$), differential diagnosis of emergency conditions ($z = -3.42$, $r_{rb} = 0.92$), familiarity with Crew Resource Management principles ($z = -3.86$, $r_{rb} = 0.94$), and confidence in emergency decision-making

($z = -4.01$, $r_{rb} = 1.00$) – were likewise substantial (Table 3). All effects remained significant after Holm-Bonferroni correction across the five items (all adjusted $p \leq .001$). Individual trajectories are shown in Fig. 4.

Table 3
Self-assessed competence before and after the VR-based training sessions ($n = 41$).

Item	Pre Mdn (IQR)	Post Mdn (IQR)	z	r_{rb}	p (Holm)
Familiarity with ED procedures	2 (2–3)	3 (3–4)	-4.86	1	< .001
Targeted use of clinical guidelines	2 (2–3)	3 (3–3)	-3.21	0.72	< .001
Differential diagnosis of emergencies	3 (2–3)	3 (3–3)	-3.42	0.92	< .001
Familiarity with CRM principles	2 (1–3)	3 (2–3)	-3.86	0.94	< .001
Confidence in emergency decision-making	2 (2–3)	3 (2–3)	-4.01	1	< .001

Wilcoxon signed-rank tests; r_{rb} = matched-pairs rank-biserial correlation; p -values Holm-corrected.

RQ4: Workload and perceived competence gain

The aggregate change in self-assessed competence averaged 3.39 points ($SD = 2.49$, $Mdn = 3$, range: -1 to 10) among the 41 participants with complete ratings. Baseline score and change were negatively correlated ($\rho = -0.50$, $p < .001$), participants with higher baseline scores tended to show smaller competence gains, whereas those with lower baseline scores tended to show larger gains, consistent with the bounded scale and with baseline being included in the change score. Across 201 workload observations nested within these 41 participants, neither workload outcome was clearly associated with changes in competence after adjustment for functional role and rotation slot. For the NASA-TLX total, each additional point of competence gain corresponded to 0.12 points higher workload without baseline adjustment ($b = 0.12$, 95% CI [- 1.19, 1.43], $p = .858$) and 1.05 points lower workload after additional adjustment for baseline score ($b = -1.05$, 95% CI [- 2.54, 0.44], $p = .166$). The cognitive-load-index models yielded similarly inconclusive estimates, both without ($b = -0.04$, 95% CI [- 1.32, 1.24], $p = .948$) and with baseline adjustment ($b = -1.19$, 95% CI [- 2.64, 0.26], $p = .107$; Table 4). Workload during the first Team Lead rotation was likewise not rank-correlated with competence gain ($\rho = -0.13$, $p = .451$; $n = 38$). Overall, the mixed-effects models provided no clear evidence of a linear between-participant association between workload and competence change, and the Team Lead-specific Spearman analysis provided no evidence of a monotonic rank association.

Table 4
Linear mixed-effects estimates of the association between perceived competence gain and workload.

Workload outcome	Baseline adjustment	b for competence gain	95% CI	p
NASA-TLX total	No	0.12	[– 1.19, 1.43]	0.858
NASA-TLX total	Yes	–1.05	[– 2.54, 0.44]	0.166
Cognitive load index	No	–0.04	[– 1.32, 1.24]	0.948
Cognitive load index	Yes	–1.19	[– 2.64, 0.26]	0.107

b is the unstandardized fixed-effect coefficient for a one-point greater aggregate perceived competence gain, calculated as the five-item post-score total minus the corresponding pre-score total. All models included functional role and rotation slot as fixed effects and a random intercept for participant. Baseline-adjusted models additionally included the five-item pre-score total. All models were based on 201 workload observations nested within 41 participants. CI = confidence interval. All *p*-values are uncorrected.

Discussion

The aim of this exploratory pilot study was to examine how perceived workload, observer-rated team performance, and self-assessed competence relate to one another in a multi-role VR emergency simulation. Workload differed systematically across functional roles: once VR-interface-specific physical demand was removed, the Team Lead carried the highest cognitive load; SOP-oriented research roles carried higher workload than pharmaceutical research roles. Higher Team Lead workload was associated with lower observer-rated team performance, while individual workload showed no clear association with perceived competence gain, which increased across all five self-assessment domains.

The Distribution of workload. The concentration of mental demand, temporal demand, effort, and frustration in the Team Lead role is consistent with the transactive costs that Collaborative Cognitive Load Theory locates in the position where distributed information must be integrated [11]. The same pattern appears in the empirical literature on resuscitation leadership: leading a resuscitation team is repeatedly described as a high-load cognitive task, and interventions that redistribute leadership functions reduce measured workload for the physician leader [26, 27]. In a cued-recall study of resuscitation leadership, expert team leaders reported lower cognitive load than novices while both groups reported consistently high demands, suggesting that expertise reorganises rather than removes the integrative cognitive burden [28]. Our undergraduate Team Lead means fall between the expert and novice values reported there, though the different instruments, populations, and scenario structures may advise against direct comparisons. Generally, our role means also span a wide range relative to published VR emergency training, where median total workload has fallen in a band described as neither under- nor overstraining [29–31]. Compared to the Team Lead role, our support roles sit toward the lower end of the cognitive demand range. Regarding the Guideline I and II Researcher roles, our study showed that retrieving information on diagnostic and management steps was experienced as more demanding

than retrieving pharmacological information. With the biggest difference seen in the NASA-TLX mental demand subscale, this finding may advise against grouping these cognitively not equivalent tasks in a single guideline research function for training and assessment purposes.

Leadership workload and team performance. Higher workload during the Team Lead rotation was associated with lower observer-rated team performance; this association was not found in other roles. Our design cannot distinguish among at least three readings: that high load impaired coordination, that a struggling team made leading harder, or that a common factor influenced both. Literature on team performance in resuscitation offers grounds for the first interpretation: hands-on involvement by the leader has been associated with poorer team dynamics and structure [32, 33], and cautions against assuming that structural leadership arrangements produce performance benefits, since designated and emergent leadership have shown few differences in some trials [34]. It is also worth noting that a high-fidelity simulation study found no association between NASA-TLX scores and TEAM ratings [35]. Our teams scored lowest on direction and command and on anticipation of required actions, and highest on composure and morale. A validation study in real emergency department resuscitations reports a comparable pattern [36]. In other TEAM applications [37], leadership has been reported as scoring lower than teamwork and task management. These findings may reflect that forward-projecting, directive behaviours are the most variable, and perhaps most demanding, dimensions of emergency team performance across levels of experience. Regarding the application of the TEAM Score in undergraduate populations, it generally can be noted that it was developed and validated with student involvement; furthermore, various studies have shown that novices (i.e. students) and experts achieve similarly reliable results, although novices tend to award slightly higher scores, suggesting it as a reliable instrument in this population. [25, 38, 39].

Physical demand in VR. The elevated physical demand of the VR Operator role motivated the application of our cognitive load index, and removing this component left the Team Lead as the single highest-load role. We had expected physical demand to reflect unfamiliarity with the headset, but participants with and without prior VR experience did not differ meaningfully, as far as can be concluded from our limited sample size. In some cases, headset and controller ergonomics themselves influence physical demands in VR operators [40], so the physical demand subscale may be per se of limited value to draw conclusions regarding interface inexperience.

Perceived competence and its relation to workload. Self-assessed competence increased across all five domains, but no clear association between workload and competence gain (RQ4) was observed. One possible explanation is that the sample size was insufficient to detect a modest association, as reflected by the wide confidence intervals encompassing estimations in both directions. Another explanation relates to the measurement characteristics of the competence scale. Since competence was assessed using five items with four response options, participants with high baseline scores had limited room for further improvement. This restriction in the possible range of change scores may have reduced the ability to detect an association between perceived competence gain and workload. The negative correlation between baseline competence and change ($p = -0.50$) is consistent with this possibility,

although it may also partly reflect the mathematical coupling of baseline and change scores. More generally, direct associations between NASA-TLX scores and learning gains are rarely reported and findings remain inconsistent [35, 41]. Additionally, subjective workload in novices should be interpreted cautiously, as it seems sensitive to affect: in VR surgical training, experienced failure inflated NASA-TLX ratings without corresponding changes in objective measures [42].

Limitations

Several limitations constrain what can be concluded from this work. Our study was a single-arm pilot study at one institution with a limited number of volunteers and no control condition. Changes in self-assessed competence cannot be attributed to specified components of the format, and expectancy effects cannot be excluded; competence was self-assessed on 4-point Likert scales, introducing statistical limitations as discussed in the previous paragraph. Detecting actual competence gains would require follow-up comparison against objective assessment standards. Workload was measured as overall perceived load, and the NASA-TLX does not differentiate intrinsic, extraneous, and germane load, which might restrict interpretability when comparisons are made to studies that operationalise cognitive load differently. Team performance was rated by student tutors who also delivered the sessions, one per session, using a rotation-anchored adaptation that departs from the instrument's validated whole-team administration; no inter-rater reliability could be estimated. Our sample size was small, and role frequencies were unbalanced. Analyses beyond the primary role model were exploratory and uncorrected. Finally, the study design did not allow for assessing reasoning quality in collaborative clinical reasoning or its unfolding in team processes beyond domains depicted in the TEAM score.

Future research and implications for instructional design.

Controlled designs comparing the multi-role format against single-user VR or fixed role assignment would allow the contributions of distribution and rotation to be separated. Following the logic of trials that redistributed leadership functions and observed reduced leader workload [27], an extension would be to test whether adding a second integrating role reduces Team Lead load and whether this affects team performance. Load-differentiated instruments would permit the intrinsic, extraneous, and germane components assumed by CCLT to be measured rather than inferred, and objective outcome measures with delayed retention should complement self-assessment. Multiple independent raters with inter-rater reliability estimation would support the reliability of performance data. Finally, combining workload profiling with observational analysis of teams' reasoning talk and results measures would connect where cognitive cost accumulates to how shared understanding is actually built.

Established role taxonomies in emergency and CRM training are often organised around procedural function, such as airway, compressions, defibrillation, medication, recording [43, 44] or around leadership functions such as information structuring, coordinating, and sense-making [45]. The roles described in our study were additionally derived from stages of the clinical reasoning process: retrieval for sequencing, retrieval for pharmacological decisions, externalisation of the evolving shared picture, synthesis for handover, and integration. Whether this is a useful basis for role design and which roles

promise the biggest educational effect in learning phases of novices cannot be derived by current studies and should be explored in future work. Additionally, role rotation itself is not unusual in simulation design so that learners do not remain in follower positions though we are not aware of controlled evidence for its educational effects and rotation strategies. Some instructional design approaches, however, can be derived from our results: Due to the already high cognitive demand, the Team Lead should probably not take over procedural execution which is supported by other studies [33, 34]. Furthermore, the Team Lead may benefit from more external tutor scaffolding which, indirectly, could increase team performance.

Conclusion

VR-based group simulation trainings with various functional team roles and only one VR operator offer a resource-conscious setup to engage undergraduate medical students in collaborative clinical reasoning. This exploratory pilot study showed differentiated workload profiles of participants across their different functional roles. Individual workload was not associated with competence gain after the training. However, higher workload in the Team Lead role was associated with lower observer-rated overall team performance, and Team Lead positions carried the highest cognitive load after removing physical demand from cognitive load measures.

While our findings should be considered hypothesis-generating, our results suggest that VR formats involving distributed functional roles should be studied in greater detail not only to understand cognitive demands better but also to tap the potential for didactical optimisation that is offered by the theoretical lenses of collaborative cognitive load theory (CCLT) and collaborative clinical reasoning (CCR).

Abbreviations

CanMEDS

Canadian Medical Education Directions for Specialists

CCLT

Collaborative Cognitive Load Theory

CCR

Collaborative Clinical Reasoning

CRM

Crew Resource Management

NASA-TLX

NASA Task Load Index

SOP

Standard Operating Procedure

TEAM

Team Emergency Assessment Measure

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of the Charité – Universitätsmedizin Berlin, Germany (Reference Number: EA4/105/23), participation was voluntary, and participants had the right to withdraw their opt-in at any point without disadvantages. Informed consent to participate was obtained from all participants. The study was conducted in accordance with all relevant ethical guidelines and regulations, including the Declaration of Helsinki.

Consent for publication

All authors reviewed and approved the final manuscript for publication.

Availability of data and materials

The annotated original dataset on which this analysis is based, including the original questionnaire items, can be found in the supplement.

Competing interests

There are no conflicts of interest to declare.

Funding

The study was carried out without using external funding; household funds of our institution were used.

Author's Contributions

Conceptualization: JJK, MM

Methodology: JJK, MM

Formal analysis: JJK

Data curation: RV, SB, RS, JJK, MM

Visualisation: JJK

Writing–original draft: JJK, MM

Acknowledgements

We would also like to thank all the students who voluntarily participated in the study and repeatedly completed the questionnaires.

Authors' information

Julius J. Kaminski, MD, MHPE, is a research fellow at Charité – Universitätsmedizin Berlin and curriculum manager at Medizinische Universität Lausitz – Carl Thiem. ORCID: <https://orcid.org/0000-0002-0310-2956>

Dorothea Penders, MD, is Head of the Learning Centre and Simulation Lab at Charité – Universitätsmedizin Berlin

Robert Vortriede is a medical student at Charité – Universitätsmedizin Berlin. ORCID: <https://orcid.org/0009-0003-2115-368X>

Sebastian Baer is a medical student at Charité – Universitätsmedizin Berlin. ORCID: <https://orcid.org/0009-0008-8697-4194>

Robin Schmid is a simulation technician at Charité – Universitätsmedizin Berlin. ORCID: <https://orcid.org/0009-0007-9147-3987>

Mareen Machner, PhD is research associate at Charité – Universitätsmedizin Berlin and Head of Simulation at Medizinische Universität Lausitz – Carl Thiem. ORCID: <https://orcid.org/0000-0003-4822-986X>

AI Use

Anthropic's Fable 5 language model was used to improve and generate code for the data analysis in Python, as well as to make spelling and stylistic corrections to the manuscript text.

References

1. Mamede S, Schmidt HG. Deliberate reflection and clinical reasoning: Founding ideas and empirical findings. *Med Educ.* 2023;57(1):76–85. 10.1111/medu.14863 . PubMed PMID: 35771936.
2. Boyle JG, Walters MR, Jamieson S, Durning SJ. Distributed cognition: Theoretical insights and practical applications to health professions education: AMEE Guide 159. *Med Teach.* 2023;45(12):1323–33. 2023.2190479 PubMed PMID: 37043405.
3. Hublin L, Koivisto JM, Lyyra M, Haavisto E. Learning Collaborative Clinical Reasoning in Healthcare Education: An Integrative Review. *J Prof Nurs.* 2023;49:126–34. 10.1016/j.profnurs.2023. 09.011 PubMed PMID: 38042546.
4. Lee CY, Lai HY, Lee CH, Chen MM, Yau SY. Collaborative clinical reasoning: a scoping review. *PeerJ.* 2024;12:e17042. 10.7717/peerj.17042 . PubMed PMID: 38464754.
5. Lee CY, Lai HY, Lee CH, Chen MM, Yau SY. Collaborative Clinical Reasoning in the ICU: How Teams Construct Shared Understanding in Real Time [Preprint] [Internet]. 2026 [cited 2026 Jun 12]. Available from: <https://www.researchsquare.com/article/rs-8666674/v1> 10.21203/rs.3.rs-8666674/v1

6. Jiang H, Vimalasvaran S, Wang JK, Lim KB, Mogali SR, Car LT. Virtual Reality in Medical Students' Education: Scoping Review. *JMIR Med Educ.* 2022;8(1):e34860. doi:10.2196/34860 PubMed PMID: 35107421.
7. Wu Q, Wang Y, Lu L, Chen Y, Long H, Wang J. Virtual Simulation in Undergraduate Medical Education: A Scoping Review of Recent Practice. *Front Med.* 2022;9:855403. 10.3389/fmed.2022.855403 . PubMed PMID: 35433717.
8. Mørk G, Bonsaksen T, Larsen OS, Kunnikoff HM, Lie SS. Virtual Reality Simulation in Undergraduate Health Care Education Programs: Usability Study. *JMIR Med Educ.* 2024;10:e56844. doi:10.2196/56844 PubMed PMID: 39560982.
9. Alghanaim N, Hart J, Finn G. Scoping Review: The Effectiveness of Interprofessional Virtual Reality Simulation. *J Med Educ Curric Dev.* 2025;12:23821205251346327. doi:10.1177/23821205251346327 PubMed PMID: 40689380.
10. Ricci S, Calandrino A, Borgonovo G, Chirico M, Casadio M, Viewpoint. Virtual and Augmented Reality in Basic and Advanced Life Support Training. *JMIR Serious Games.* 2022;10(1):e28595. doi:10.2196/28595 PubMed PMID: 35319477.
11. Kirschner PA, Sweller J, Kirschner F, Zambrano JR. From Cognitive Load Theory to Collaborative Cognitive Load Theory. *Int J Comput-Support Collab Learn.* 2018;13(2):213–33. 10.1007/s11412-018-9277-y PubMed PMID: 30996713.
12. Fraser KL, Ma I, Teteris E, Baxter H, Wright B, McLaughlin K. Emotion, cognitive load and learning outcomes during simulation training. *Med Educ.* 2012;46(11):1055–62. 10.1111/j.1365-2923.2012.04355. .x PubMed PMID: 23078682.
13. Fraser KL, Ayres P, Sweller J. Cognitive Load Theory for the Design of Medical Simulations. *Simul Healthc J Soc Simul Healthc.* 2015;10(5):295–307. 10.1097/SIH.000000000000097 PubMed PMID: 26154251.
14. Fraser KL, Meguerdichian MJ, Haws JT, Grant VJ, Bajaj K, Cheng A. Cognitive Load Theory for debriefing simulations: implications for faculty development. *Adv Simul.* 2018;3(1):28. 10.1186/s41077-018-0086-1 . PubMed PMID: 30619626.
15. Hou Y, Xie Q, Zhang N, Lv J. Cognitive load classification of mixed reality human computer interaction tasks based on multimodal sensor signals. *Sci Rep.* 2025;15(1):13732. 10.1038/s41598-025-98891-3 . PubMed PMID: 40259027.
16. Reedy GB. Using Cognitive Load Theory to Inform Simulation Design and Practice. *Clin Simul Nurs.* 2015;11(8):355–60. 10.1016/j.ecns.2015.05.004.
17. Sweller J, Ayres P, Kalyuga S. Cognitive Load Theory [Internet]. New York, NY: Springer New York; 2011 [cited 2026 Jul 9]. Available from: <https://link.springer.com/10.1007/978-1-4419-8126-4> doi:10.1007/978-1-4419-8126-4.
18. Mühling T, Späth I, Backhaus J, Milke N, Oberdörfer S, Meining A, et al. Virtual reality in medical emergencies training: benefits, perceived stress, and learning success. *Multimed Syst.* 2023;29(4):2239–52. 10.1007/s00530-023-01102-0.

19. Frank JR, Snell L, Sherbino J, editors. CanMEDS 2015 physician competency framework. Ottawa: Royal College of Physicians and Surgeons of Canada; 2015.
20. Kersten C, Fink K, Michels G, Busch HJ. Crew Resource Management im Schockraum. *Med Klin - Intensivmed Notfallmedizin*. 2021;116(5):377–88. 10.1007/s00063-021-00808-1 . PubMed PMID: 33830287.
21. Schaal J, Leutritz T, Lindner M, Zamzow A, Backhaus J, König S, et al. Immersive competence as a source of bias in virtual reality clinical assessment. *Npj Digit Med*. 2026;9(1):280. 10.1038/s41746-026-02482-z . PubMed PMID: 41803233.
22. Maaz A, Hitzblech T, Arends P, Degel A, Ludwig S, Mossakowski A, et al. Moving a mountain: Practical insights into mastering a major curriculum reform at a large European medical university. *Med Teach*. 2018;40(5):453–60. 10.1080/0142159X.2018.1440077.
23. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: *Advances in Psychology* [Internet]. Elsevier; 1988 [cited 2026 Jul 9]. pp. 139–83. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0166411508623869> 10.1016/S0166-4115(08)62386-9
24. Hart SG. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proc Hum Factors Ergon Soc Annu Meet*. 2006;50(9):904–8. 10.1177/154193120605000909.
25. Cooper S, Cant R, Porter J, Sellick K, Somers G, Kinsman L, et al. Rating medical emergency teamwork performance: Development of the Team Emergency Assessment Measure (TEAM). *Resuscitation*. 2010;81(4):446–52. 10.1016/j.resuscitation. .2009.11.027 PubMed PMID: 20117874.
26. Jurd C, Barr J. Leadership factors for cardiopulmonary resuscitation for clinicians in-hospital; behaviours, skills and strategies: A systematic review and synthesis without meta-analysis. *J Clin Nurs*. 2024;33(10):3844–53. 10.1111/jocn.17215 . PubMed PMID: 38757400.
27. Pallas JD, Smiles JP, Zhang M. Cardiac Arrest Nurse Leadership (CANLEAD) trial: a simulation-based randomised controlled trial implementation of a new cardiac arrest role to facilitate cognitive offload for medical team leaders. *Emerg Med J EMJ*. 2021;38(8):572–8. 10.1136/emered-2019-209298 . PubMed PMID: 33500268.
28. Popov V, Harmer B, Raphael S, Scott I, Sample AP, Cooke JM, et al. Elucidating cognitive processes in cardiac arrest team leaders: a virtual reality-based cued-recall study of experts and novices. *Ann Med*. 2025;57(1):2470976. 10.1080/07853890.2025.2470976 . PubMed PMID: 40028867.
29. Birrenbach T, Wespi R, Hautz WE, Berger J, Schwab PR, Papagiannakis G, et al. Development and usability testing of a fully immersive VR simulation for REBOA training. *Int J Emerg Med*. 2023;16(1):67. 10.1186/s12245-023-00545-6 . PubMed PMID: 37803269.
30. Birrenbach T, Stuber R, Müller CE, Sutter PM, Hautz WE, Exadaktylos AK, et al. Virtual reality simulation to enhance advanced trauma life support trainings – a randomized controlled trial. *BMC Med Educ*. 2024;24(1):666. 10.1186/s12909-024-05645-2 . PubMed PMID: 38886688.
31. Grier RA. How High is High? A Meta-Analysis of NASA-TLX Global Workload Scores. *Proc Hum Factors Ergon Soc Annu Meet*. 2015;59(1):1727–31. 10.1177/1541931215591373.

32. Cooper S, Wakelam A. Leadership of resuscitation teams: 'Lighthouse Leadership'. *Resuscitation*. 1999;42(1):27–45. 10.1016/S0300-9572(. 99)00080-5 PubMed PMID: 10524729.
33. Kern P, Tschan F, Semmer NK, Marsch S. Effects of team leaders' position in cardiopulmonary resuscitation teams on leadership behavior and team performance: A prospective randomized interventional cross-over simulation-based trial. *Med (Baltim)*. 2023;102(27):e34235. 10.1097/MD.00000000000034235 . PubMed PMID: 37417598; PubMed Central PMCID: PMC10328606.
34. Hunziker S, O'Connell KJ, Ranniger C, Su L, Hochstrasser S, Becker C, et al. Effects of designated leadership and team-size on cardiopulmonary resuscitation: The Basel-Washington SIMulation (BaWaSim) trial. *J Crit Care*. 2018;48:72–7. .001 PubMed PMID: 30172964.
35. Favre-Félix J, Dziadzko M, Bauer C, Duclos A, Lehot JJ, Rimmelé T, et al. High-Fidelity Simulation to Assess Task Load Index and Performance: A Prospective Observational Study. *Turk J Anaesthesiol Reanim*. 2022;50(4):282–7. 10.5152. /TJAR.2022.21234 PubMed PMID: 35979975.
36. Karlgren K, Dahlström A, Birkestam A, Drevstam Norling A, Forss G, Andersson Franko M, et al. The TEAM instrument for measuring emergency team performance: validation of the Swedish version at two emergency departments. *Scand J Trauma Resusc Emerg Med*. 2021;29(1):139. 10.1186/s13049-021-00952-9 . PubMed PMID: 34544459.
37. Cooper S, Cant R, Connell C, Sims L, Porter JE, Symmons M, et al. Measuring teamwork performance: Validity testing of the Team Emergency Assessment Measure (TEAM) with clinical resuscitation teams. *Resuscitation*. 2016;101:97–101. 10.1016/j.resuscitation. .2016.01.026 PubMed PMID: 26875992.
38. Freytag J, Stroben F, Hautz WE, Schaubert SK, Kämmer JE. Rating the quality of teamwork—a comparison of novice and expert ratings using the Team Emergency Assessment Measure (TEAM) in simulated emergencies. *Scand J Trauma Resusc Emerg Med*. 2019;27(1):12. 10.1186/s13049-019-0591-9 . PubMed PMID: 30736821.
39. Hultin M, Jonsson K, Härgestam M, Lindkvist M, Brulin C. Reliability of instruments that measure situation awareness, team performance and task performance in a simulation setting with medical students. *BMJ Open*. 2019;9(9):e029412. 10.1136/bmjopen-2019-029412 . PubMed PMID: 31515425.
40. Kuber PM, Rashedi E. Alterations in Physical Demands During Virtual/Augmented Reality-Based Tasks: A Systematic Review. *Ann Biomed Eng*. 2023;51(9):1910–32. 10.1007/s10439-023-03292-0 . PubMed PMID: 37486385.
41. Tokuno J, Carver TE, Fried GM. Measurement and Management of Cognitive Load in Surgical Education: A Narrative Review. *J Surg Educ*. 2023;80(2):208–15. 10.1016/j.jsurg.2022.10.001 . PubMed PMID: 36335034.
42. Rölfling JD, Nørskov JK, Paltved C, Konge L, Andersen SAW. Failure affects subjective estimates of cognitive load through a negative carry-over effect in virtual reality simulation of hip fracture surgery. *Adv Simul*. 2019;4(1):26. 10.1186/s41077-019-0114-9 . PubMed PMID: 31832243.

43. Reznick M, Emergency Medicine Crisis Resource Management (EMCRM). Pilot Study of a Simulation-based Crisis Management Course for Emergency Medicine. Acad Emerg Med. 2003;10(4):386–9. 10.1197/aemj.10.4.386 . PubMed PMID: 12670855.
44. Shapiro MJ, Gardner R, Godwin SA, Jay GD, Lindquist DG, Salisbury ML, et al. Defining Team Performance for Simulation-based Training: Methodology, Metrics, and Opportunities for Emergency Medicine. Acad Emerg Med. 2008;15(11):1088–97. 10.1111/j.1553-2712.2008.00251. .x PubMed PMID: 18828830.
45. Rosenman ED, Shandro JR, Ilgen JS, Harper AL, Fernandez R. Leadership Training in Health Care Action Teams: A Systematic Review. Acad Med. 2014;89(9):1295–306. 10.1097/ACM.0000000000000413 PubMed PMID: 25006708.

Figures

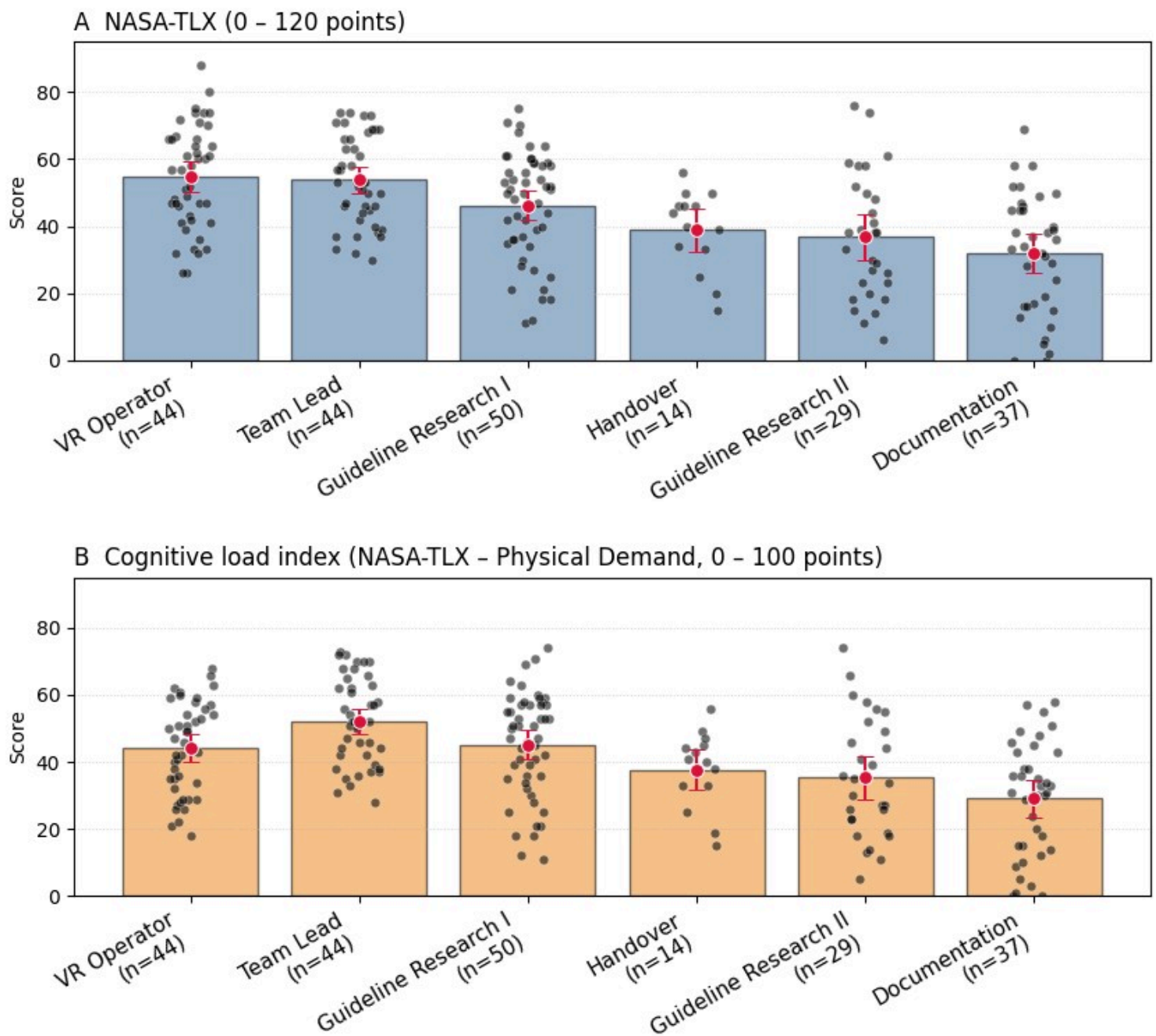


Figure 1

Perceived workload by functional role (mean $\pm$ 95% CI). (A) NASA-TLX total score (0–120). (B) Cognitive load index (total minus Physical Demand, 0–100). Dots represent individual observations; red markers show group means with 95% confidence intervals; n per role as reported in Table 2.

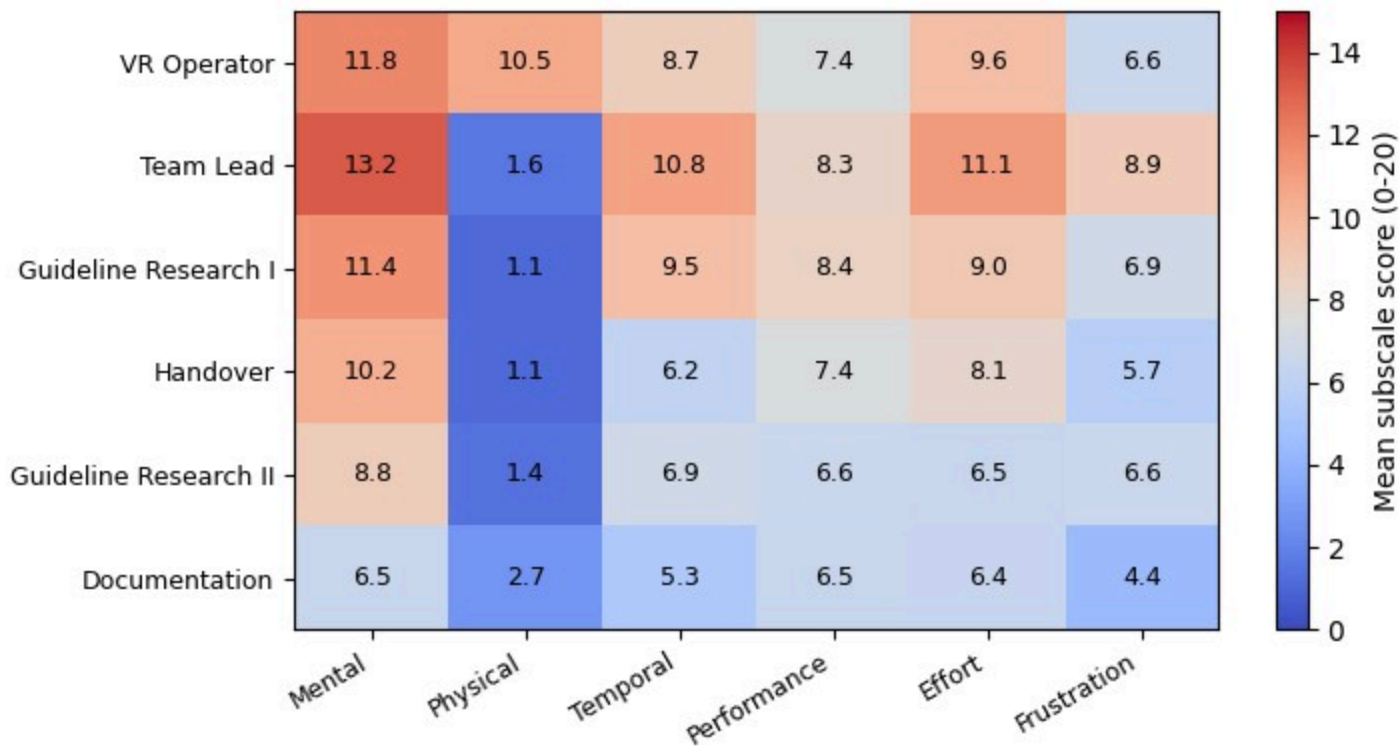


Figure 2

Mean NASA-TLX subscale score profile by role (0–20). Cell values are means across all observations within each role.

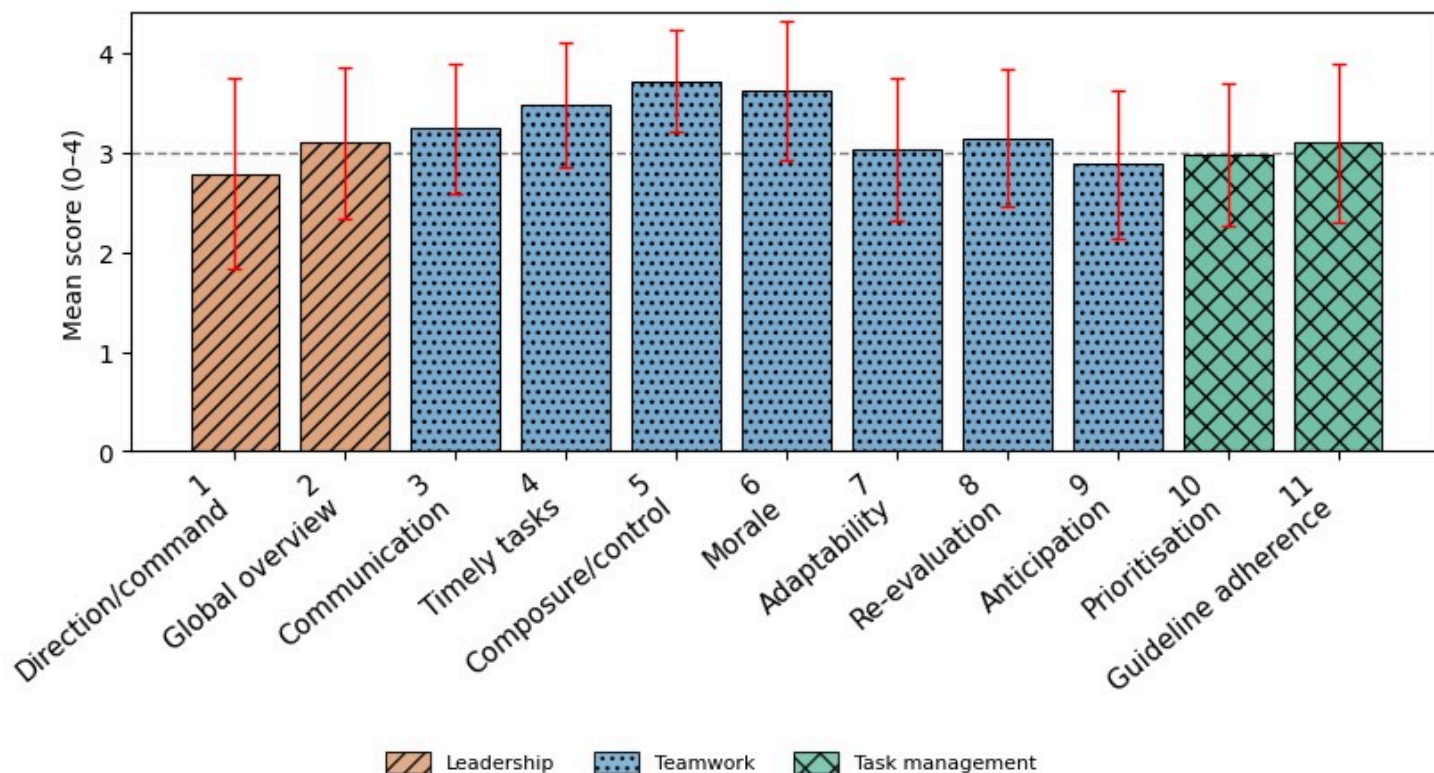


Figure 3

Mean TEAM item scores across participants (n = 42; items 1–11, scale 0–4). Grouped by the instrument's subscales: Leadership (items 1 – 2), Teamwork (items 3 – 9), and Task Management (items 10 – 11). Error bars represent standard deviations; the dashed line marks a score of 3.0.

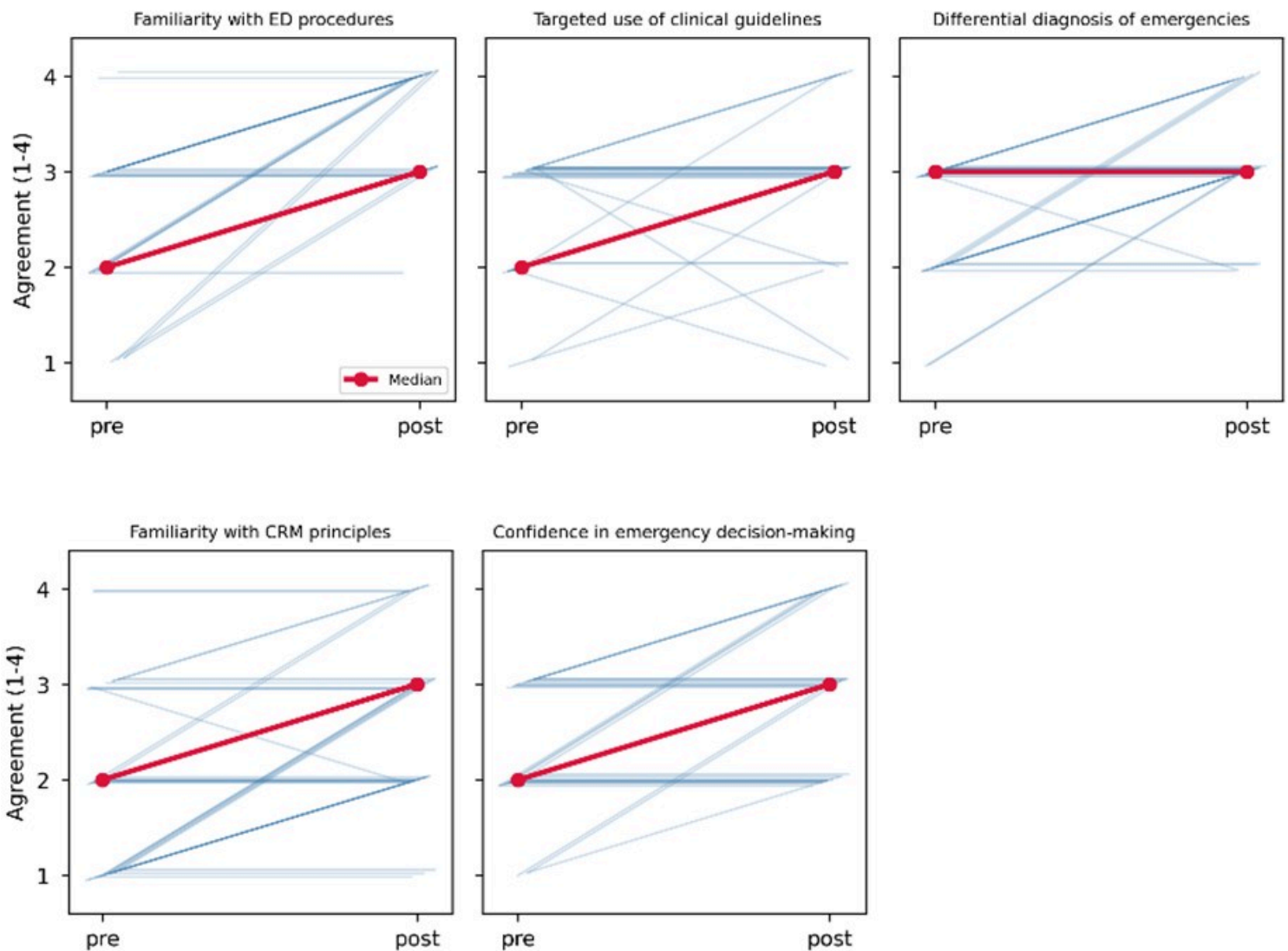


Figure 4

Self-assessed competence before and after the VR-based training sessions (n = 41). Individual pre–post trajectories (blue lines) and medians (red) for the five self-assessment items: (1) familiarity with ED procedures, (2) targeted use of clinical guidelines, (3) differential diagnosis of emergencies, (4) familiarity with CRM principles, (5) confidence in emergency decision-making (n = 41; 4-point agreement scale).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [MultimediaAppendix1.mp4](#)

- [DatasetBMCmedEducWhoCarriesTheLoad.xlsx](#)
- [SupplementFile3QuestionnaireWhoCarriesTheLoad.pdf](#)