

What does immersive virtual reality actually teach? A performance–knowledge dissociation in a blinded cluster randomized trial of catheterization skills in nursing students

María del Carmen Agüera-Bolea

mdcaguera@ucam.edu

Universidad Católica San Antonio de Murcia (UCAM)

Pablo Fernández-Molina

Universidad Católica San Antonio de Murcia (UCAM)

Felipe Jesús Portillo-Castaño

Universidad Católica San Antonio de Murcia (UCAM)

Eloina Valero-Merlos

Universidad Católica San Antonio de Murcia (UCAM)

Paula García-Ibáñez

Universidad Católica San Antonio de Murcia (UCAM)

Sergio Navarro-Sánchez

Universidad Católica San Antonio de Murcia (UCAM)

Research Article

Keywords: virtual reality, simulation-based education, assessment alignment, OSCE, cluster randomized trial, nursing education, catheterization

Posted Date: July 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10410249/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Advances in Simulation on October 3rd, 2026. See the published version at <https://doi.org/10.1186/s41077-026-00492-1>.

Abstract

Background

Immersive virtual reality (VR) reliably produces large simulated-performance effects in health professions education, yet whether such effects reflect transferable competence or alignment between rehearsed scenarios and assessment criteria remains empirically untested. Meta-analyses disagree about whether VR benefits skills or knowledge more—an inconsistency a constructive-alignment account predicts if studies align scenario and assessment to differing degrees. We tested this within one adequately powered, blinded trial by locating the VR advantage at the assessment-component level.

Methods

In a cluster randomized controlled study, 32 intact teaching subgroups (442 second-year nursing students) were allocated 1:1 to immersive VR (Meta Quest Pro; three procedure-choreographed scenarios) or conventional low-fidelity simulation for urinary, nasogastric, and rectal catheterization; both arms then received an identical seminar. Outcomes were a blinded, double-rated three-station OSCE one week after training and a 30-item knowledge test administered immediately and at four weeks. Exploratory analyses decomposed the advantage by checklist item and compared performance and knowledge effect sizes; inference used cluster-robust wild bootstrap.

Results

The VR arm outperformed control on the composite OSCE (87.33 ± 5.05 versus 70.70 ± 8.51 ; difference 16.63, 95% CI 15.32–17.94; Hedges' $g = 2.38$) and on knowledge (immediate difference 4.80 of 30, $g = 1.26$; four weeks 4.03, $g = 1.04$; ANCOVA-adjusted 3.23). The performance effect significantly exceeded the knowledge effect ($\Delta g = 1.34$; $p < 0.001$). The item-level advantage was selective (range -0.11 to $+0.39$): largest for documentation, positioning, verification, and asepsis (steps the scenarios rehearsed), whereas control outperformed VR on material preparation—the one step VR cannot rehearse. Knowledge effects tracked scenario dose (urinary $g = 1.53$; nasogastric 0.25; rectal 0.33).

Conclusions

Immersive VR produced a large but selective advantage that followed what its scenarios rehearsed, alongside a smaller, retained knowledge benefit. To our knowledge, this is the first adequately powered, blinded, within-trial demonstration that immersive-VR effects localize to rehearsed, assessed components—reversing on the one step VR cannot rehearse—helping reconcile why syntheses disagree about VR's benefit to skills versus knowledge. Large VR effects should be read as partly alignment-

dependent; VR is best positioned as a preparatory complement to physical practice, with assessment sampling beyond the rehearsed choreography.

Background

Safe, reproducible psychomotor performance is a defining competency of undergraduate nursing education, and catheterization is among the core procedures students must master before clinical placement. Competent catheterization requires the coordinated execution of a sequenced motor routine together with aseptic or clean technique, patient communication, comfort, and accurate documentation. Simulation-based education offers a psychologically safe environment in which learners can rehearse procedural routines before clinical exposure and receive criterion-referenced feedback, in line with the *Healthcare Simulation Standards of Best Practice* (INACSL Standards Committee, 2021; Cheng et al., 2016). Conventional low-fidelity simulation, however, is constrained by finite equipment, uneven exposure, and limited opportunity to repeatedly observe a standardized procedural exemplar—factors that introduce variability into early skill acquisition.

Immersive VR has been proposed to mitigate these constraints by delivering standardized, reproducible, and visually detailed procedural exemplars within a patient-safe environment that affords repeated, low-risk rehearsal and a strong sense of spatial presence. Contemporary syntheses indicate that VR can improve learning across the cognitive, affective, and psychomotor domains, but that effects are heterogeneous and moderated by immersion level, comparator, scenario design, and outcome measurement (Chen et al., 2020; Efendi et al., 2023; Hsieh et al., 2025; Kyaw et al., 2019; Liu et al., 2023; Park et al., 2024; Vogelsang et al., 2024). Meta-analytic estimates for nursing competency are moderate to large (Hedges' $g \approx 0.88$; Cho & Kim, 2024), and—importantly for the present study—effects on declarative knowledge are typically smaller and less consistent than effects on skill performance (Hsieh et al., 2025; Woon et al., 2021). Whether immersive VR outperforms hands-on simulated practice specifically for the acquisition of clinical psychomotor skills remains only partially resolved, with the existing evidence base described as limited and methodologically variable (Rourke, 2020).

For catheterization specifically, VR—including haptic and game-based approaches—has been associated with favorable effects on skill, confidence, and satisfaction (Butt et al., 2018; Chang, 2022; Park & Yoon, 2023; Şendir et al., 2022; Yoon et al., 2024). Yet the available studies are constrained by immediate rather than delayed assessment, evaluation of a single catheterization type, reliance on self-report or single-rater scoring, and modest sample sizes. More fundamentally, the meta-analytic evidence is internally contradictory about which domain benefits most: some syntheses report larger effects on skills than knowledge (Hsieh et al., 2025), whereas others report the reverse (Lee et al., 2024; Liu et al., 2023). This inconsistency is what a constructive-alignment account would predict, because pooled estimates average studies whose scenarios and assessments align to differing degrees (Neher et al., 2025). What is missing is not another domain-level estimate but a within-trial test of the mechanism: whether a single VR intervention's advantage localizes to the specific steps its scenarios rehearse and the assessment samples, while sparing—or reversing on—unrehearsed components. Mirror-image findings sharpen the

question: a recent randomized trial of a knowledge-oriented VR session improved theoretical knowledge without changing OSCE performance (Tannemann et al., 2025), the inverse of the pattern expected from procedure-choreographed VR.

The study is anchored in the Cognitive-Affective Model of Immersive Learning (CAMIL; Makransky & Petersen, 2021), which posits that immersion supports learning through presence and agency while remaining subject to cognitive-load constraints: well-designed immersion may free working-memory resources for germane processing, whereas interface complexity or distraction may impose extraneous load (Hart & Staveland, 1988). Presence—the subjective sense of “being there”—can promote engagement and affective activation but is necessary rather than sufficient for learning (Schubert et al., 2001; Slater, 2018). Performance was therefore operationalized through a criterion-referenced Objective Structured Clinical Examination (OSCE), an established approach to competency-based performance assessment first introduced by Harden and colleagues (Harden et al., 1975; Boursicot et al., 2021). Performance was complemented by measures of knowledge retention, perceived workload, and presence, rather than by satisfaction alone.

Objectives. The primary objective was to compare the composite one-week OSCE score between VR and control arms. Secondary objectives were to compare procedure-specific OSCE performance, estimate inter-rater reliability, compare immediate and four-week knowledge (retention), and characterize satisfaction, self-confidence, perceived workload, and presence in the VR arm. Exploratory analyses additionally located the performance advantage at the checklist-item level and formally compared performance and knowledge effect sizes to test the alignment account.

Methods

Design and reporting

This was a cluster randomized controlled educational study, reported in accordance with the CONSORT statement, its extension for cluster randomized trials, and the simulation-research reporting extensions (Cheng et al., 2016); a participant-flow diagram is provided (Fig. 1). The unit of randomization was the intact teaching subgroup: before the sessions began, the 32 timetabled subgroups (13–14 students each) were allocated 1:1 to the arms (16 immersive VR; 16 low-fidelity control) using computer-generated random numbers (one number per subgroup, sorted in ascending order) by a team member not involved in the intervention, OSCE assessment, or analysis. Students remained in their pre-scheduled subgroups; for laboratory delivery, subgroups within the same arm attended sessions in their timetabled laboratory practice groups, which served as the data-recording unit. The teaching subgroup was the clustering unit in all cluster-robust analyses.

Sample and setting

The study was conducted within the *Clinical Simulation I* course of a single Bachelor of Nursing program at a Spanish university, during scheduled catheterization simulation sessions for second-year students. Data were collected across the 32 scheduled teaching subgroups. All teaching subgroups followed the same curriculum, objectives, procedural protocols, OSCE checklists, and assessment procedures. Eligible participants were second-year students who attended the sessions and consented to the use of anonymized educational data. Students were excluded from a given analysis if they did not attend the corresponding assessment, withdrew consent, or had incomplete data for that outcome.

Study procedure and intervention

The two arms differed only in the modality of the first training session; the second (one-week) seminar was delivered identically to both arms (Fig. 1).

Session 1 – training modality (4 hours). *VR arm:* training was delivered on a Meta Quest Pro head-mounted display in a dedicated VR laboratory with a technician present, following the INACSL cycle of orientation/prebriefing, scenario, and debriefing (INACSL Standards Committee, 2021). The session comprised (a) technician-led orientation to VR, its educational purpose, and equipment use; (b) a 5-minute primer introducing catheterization, which students had not previously studied; (c) an 8-minute immersive *nasogastric* scenario followed by structured debriefing (errors, correct actions, learner reflection); (d) a 14-minute *urinary* scenario covering both male and female catheterization, followed by debriefing; (e) a 30-minute break; and (f) a 5-minute *rectal* scenario followed by a final debriefing. The three immersive scenarios (8, 14, and 5 minutes) were embedded within approximately one hour of active VR practice. *Control arm:* students completed conventional low-fidelity simulation of the same three procedures using physical task-trainers and procedural guides, with approximately one hour of hands-on practice, across a 4-hour session.

Session 2 – one-week seminar (4 hours; identical for both arms). Catheterization theory was presented, followed by hands-on practice, the OSCE (the OSCE assessment section), and the immediate knowledge test (the knowledge assessment section). Because this seminar was standardized across arms, between-group differences in OSCE performance are attributable to the Session-1 training modality.

Four-week assessment. The knowledge test (the knowledge assessment section) was re-administered four weeks after Session 1 to assess retention.

VR-arm self-report questionnaires were administered at the end of Session 1. Each arm comprised two 4-hour sessions (eight hours of matched total contact time), and hands-on procedural practice time was one hour in each arm.

OSCE assessment

The OSCE was conducted at the end of the one-week seminar—one week after the initial training session—using three stations (urinary, nasogastric, rectal), each scored with the same standardized instrument. The instrument was developed by the teaching team from the course learning objectives, the clinical and simulation practice protocols used in the program, and the critical safety, technique, verification, and documentation steps of each catheterization procedure; its item content was additionally informed by the skill dimensions assessed in prior catheterization training studies (Butt et al., 2018; Park & Yoon, 2023; Şendir et al., 2022). The instrument followed the standard OSCE format of a structured station with an observable dichotomous checklist and a global rating scale (Harden et al., 1975; Khan et al., 2013; Boursicot et al., 2021). Checklist items and the global rating scale were reviewed by the teaching team for content relevance and clarity before use. The instrument comprised an 11-item dichotomous checklist (0 = not performed/incorrect; 1 = performed correctly) covering introduction and patient identification, hand hygiene and personal protective equipment, material preparation, patient positioning, maintenance of privacy and dignity, aseptic technique, correct insertion technique and direction, verification of placement, catheter fixation and patient comfort, safe disposal per protocol, and documentation. The procedure-specific percentage score was the checklist total divided by 11 and multiplied by 100. A Global Rating Scale (GRS) rated four dimensions—technical and procedural precision, safety and asepsis, communication and patient rapport, and professionalism and organization—on a 1–5 scale, with an overall rating computed as their mean. Each station was scored independently by the same two trained assessors, blinded to arm allocation, using anonymized participant codes; students were instructed not to disclose their training modality, and blinding integrity was verified. Final procedure-specific scores were the mean of the two assessors' checklist percentage scores; the composite OSCE score (mean of the three procedure scores) was the primary outcome, and the composite GRS was a secondary descriptor.

Knowledge assessment (immediate and four-week retention)

A 30-item single-best-answer multiple-choice test (10 items each for urinary, nasogastric, and rectal catheterization) was administered twice: immediately after the Session-2 seminar and again four weeks after Session 1 to assess retention. Items were scored 1 (correct) or 0 (incorrect); the total ranged 0–30 (subscores 0–10). Two parallel test versions (A and B) were used, distributed approximately evenly across arms at both administrations (61% of students received the same version twice); scoring was automated against a fixed answer key, and test version was included as a covariate in adjusted analyses. Internal consistency was estimated with the Kuder–Richardson formula 20 (KR-20). The test was administered in Spanish; both forms and the scoring key are provided as supplementary material.

VR-arm self-report measures

Satisfaction and self-confidence were measured before and after Session 1 with an adapted Student Satisfaction and Self-Confidence in Learning (SSSCL) scale (satisfaction 5–25; self-confidence 8–40).

Perceived workload used the Raw NASA-TLX (unweighted mean of six 0–10 dimensions); presence used an 8-item Igroup Presence Questionnaire (IPQ)-based score.

Data analysis

The sample comprised all eligible second-year students in the scheduled cohort (census sampling); no a priori sample-size calculation was performed. A minimum-detectable-effect calculation indicated that, with 222 versus 220 students, $\alpha = .05$ (two-sided) and 80% power, the minimum detectable between-group effect was Cohen's $d \approx 0.27$. The unit of analysis was the individual student. Continuous variables were summarized with means, standard deviations, medians, interquartile ranges, and ranges; categorical variables with frequencies and percentages. The primary analysis compared composite OSCE scores with Welch's independent-samples t-test, reporting mean differences, 95% confidence intervals, p-values, and Hedges' g . Because randomization was by intact teaching subgroup, an adjusted linear model regressed the composite OSCE score on arm, age, and sex (HC3 standard errors); inference for the arm effect additionally used cluster-robust standard errors with the teaching subgroup as the clustering unit (32 clusters) and wild cluster bootstrap p-values (Rademacher weights, 9,999 replications, null imposed), and subgroup-level intraclass correlations (ICC) and design effects were estimated with random-intercept models. Procedure-specific outcomes were analyzed analogously as secondary outcomes, with Mann–Whitney U sensitivity analyses for the composite OSCE and knowledge total. Knowledge retention was analyzed with within-arm change scores and an ANCOVA model regressing four-week scores on arm, immediate score, and test version (HC3 standard errors). In exploratory analyses, the between-arm difference was decomposed by checklist item (rater-averaged item scores collapsed to one value per student per item), with per-item differences tested using cluster-robust wild bootstrap (32 clusters) and Benjamini–Hochberg false-discovery-rate correction across the 11 items; the composite-performance and four-week knowledge effect sizes were formally compared with a z-test on their standardized estimates; because both were measured in the same students, the within-arm cross-domain correlation was examined to assess the independence assumption. Inter-rater reliability used two-way random-effects, absolute-agreement ICCs (single and average measures). VR-arm pre–post satisfaction and self-confidence used paired t-tests and Cohen's d_z on complete cases (all 13 SSSCL items at both administrations); scale reliabilities were reported (Cronbach's α ; KR-20 for knowledge). Under the cluster randomized design, between-group differences are interpreted as effects of the Session-1 training modality. Statistical analyses were performed using IBM SPSS Statistics version 28.0 (IBM Corp., Armonk, NY, USA). Microsoft Excel was used for data structuring, traceability, cleaning, and audit logs.

Ethical considerations

The study was approved by the institutional research ethics committee (ID: CE012606). Participation was voluntary, informed consent was obtained, and anonymized data were analyzed in accordance with applicable data-protection regulations.

Results

Participant flow and characteristics

Table 1
Baseline characteristics of participants by arm.

Variable	Immersive VR (n = 222)	Control (n = 220)	SMD
Age, years, mean ± SD	21.5 ± 4.6	20.9 ± 4.4	0.15
Age, median [IQR]	19 [19–22]	20 [19–20]	—
Female, n (%)	173 (77.9)	169 (76.8)	0.02
Male, n (%)	49 (22.1)	51 (23.1)	—
Prior clinical-simulation experience, basic level, n (%)	222 (100.0)	220 (100.0)	—

Of 443 consenting students, 442 had complete composite OSCE data (VR n = 222; control n = 220) and constituted the primary analysis set; all 442 completed both the immediate and the four-week knowledge tests (VR n = 222; control n = 220), and 201 VR-arm students had complete paired SSSCL data (NASA-TLX/IPQ n = 214). One control student was excluded for an incomplete OSCE record (Fig. 1). Mean age was 21.5 ± 4.6 years (VR) and 20.9 ± 4.4 years (control); women accounted for 77.9% (VR) and 76.8% (control). Most students in both arms reported prior clinical-simulation experience at a basic level (Table 1). Standardized mean differences were small (age 0.15; sex 0.02; Table 1).

Inter-rater reliability

Table 2
Inter-rater reliability of OSCE checklist and global rating scale (intraclass correlation coefficients).

Arm	Procedure	n	Checklist ICC (single)	Checklist ICC (average)	GRS ICC (single)	GRS ICC (average)
VR	Urinary	222	0.935	0.966	0.830	0.907
VR	Nasogastric	222	0.919	0.958	0.882	0.937
VR	Rectal	222	0.860	0.925	0.804	0.891
Control	Urinary	220	0.859	0.924	0.795	0.885
Control	Nasogastric	220	0.968	0.984	0.967	0.983
Control	Rectal	220	0.966	0.983	0.975	0.987

Inter-rater reliability was good to excellent for both the checklist and the global rating scale across both arms and all procedures. Average-measures checklist ICCs ranged from 0.925 to 0.966 in the VR arm

and from 0.924 to 0.984 in the control arm, and global-rating ICCs ranged from 0.891 to 0.937 in the VR arm and from 0.885 to 0.987 in the control arm (Table 2). These values supported using the mean of the two assessors' scores as the final procedure-specific score.

Table 2

Inter-rater reliability of OSCE checklist scores (two-way random effects, absolute agreement).

Arm	Procedure	n	Checklist ICC (single)	Checklist ICC (average)	GRS ICC (single)	GRS ICC (average)
VR	Urinary	222	0.935	0.966	0.830	0.907
VR	Nasogastric	222	0.919	0.958	0.882	0.937
VR	Rectal	222	0.860	0.925	0.804	0.891
Control	Urinary	220	0.859	0.924	0.795	0.885
Control	Nasogastric	220	0.968	0.984	0.967	0.983
Control	Rectal	220	0.966	0.983	0.975	0.987

One-week OSCE performance

Table 3

One-week OSCE performance by arm.

Outcome	VR (n = 222), M $\pm$ SD	Control (n = 220), M $\pm$ SD	Mean difference	95% CI	p	Hedges' g (95% CI)
Composite OSCE, 0–100	87.33 $\pm$ 5.05	70.70 $\pm$ 8.51	16.63	15.32–17.94	< 0.001	2.38 (2.14–2.62)
Urinary, 0–100	86.00 $\pm$ 9.38	69.21 $\pm$ 10.15	16.78	14.95–18.61	< 0.001	1.71 (1.49–1.93)
Nasogastric, 0–100	88.14 $\pm$ 8.25	72.29 $\pm$ 11.50	15.85	13.98–17.72	< 0.001	1.58 (1.37–1.79)
Rectal, 0–100	87.86 $\pm$ 8.20	70.60 $\pm$ 10.70	17.26	15.48–19.04	< 0.001	1.81 (1.59–2.03)
Composite global rating, 1–5	4.03 $\pm$ 0.25	3.32 $\pm$ 0.42	0.72	0.65–0.78	< 0.001	2.07 (1.84–2.30)

The composite OSCE score was higher in the VR arm (87.33 $\pm$ 5.05) than the control arm (70.70 $\pm$ 8.51), a mean difference of 16.63 points (95% CI 15.32–17.94; p < 0.001; Hedges' g = 2.38 [95% CI 2.14–2.62]; Table 3, Fig. 2A). A Mann–Whitney U sensitivity analysis of the composite score was concordant (p < 0.001), with a probability of superiority of 0.97 (i.e., a randomly selected VR student outscored a randomly selected control student 97% of the time). The adjusted model retained the effect after controlling for age and sex (B = 16.65; 95% CI 15.34–17.97, HC3; cluster-robust 95% CI 14.46–18.85). The teaching-subgroup ICC for the composite score was 0.15 (design effect $\approx$ 2.9; 32 clusters); cluster-

robust inference confirmed the arm effect ($t = 14.8$; wild cluster bootstrap $p < 0.001$). The VR arm scored higher on all procedures: urinary 86.00 ± 9.38 versus 69.21 ± 10.15 ($g = 1.71$), nasogastric 88.14 ± 8.25 versus 72.29 ± 11.50 ($g = 1.58$), and rectal 87.86 ± 8.20 versus 70.60 ± 10.70 ($g = 1.81$). Composite global-rating scores were also higher with VR (4.03 ± 0.25 versus 3.32 ± 0.42 ; $g = 2.07$).

Knowledge outcomes and four-week retention

Table 4
Key secondary outcomes: four-week knowledge retention and VR-arm self-report.

Outcome	Sample	VR / Pre	Control / Post	Difference / effect	Role
Knowledge total, immediate (0–30)	VR n = 222 / control n = 220	18.33 ± 3.19	13.53 ± 4.35	$4.80 (4.09–5.52)$; $g = 1.26 (1.06–1.46)$	Secondary
Knowledge total, four-week (0–30)	VR n = 222 / control n = 220	18.63 ± 3.24	14.60 ± 4.43	$4.03 (3.30–4.76)$; $g = 1.04 (0.84–1.24)$	Key secondary
Knowledge urinary, 0–10	VR / control	7.07 ± 1.94	4.12 ± 1.89	$2.94 (2.59–3.30)$; $g = 1.53 (1.32–1.74)$	Secondary
Knowledge nasogastric, 0–10	VR / control	5.28 ± 2.19	4.73 ± 2.22	$0.55 (0.14–0.96)$; $g = 0.25 (0.06–0.44)$	Secondary
Knowledge rectal, 0–10	VR / control	6.28 ± 1.71	5.75 ± 1.56	$0.53 (0.23–0.84)$; $g = 0.33 (0.14–0.52)$	Secondary
Satisfaction (SSSCL), 5–25	VR only, n = 201	22.50 ± 3.33	23.47 ± 2.25	$0.97 (0.43–1.50)$; $dz = 0.25 (0.11–0.39)$	Secondary
Self-confidence (SSSCL), 8–40	VR only, n = 201	33.90 ± 5.47	35.83 ± 4.34	$1.93 (1.01–2.85)$; $dz = 0.29 (0.15–0.43)$	Secondary
Raw NASA-TLX, 0–10	VR only, n = 214	—	2.65 ± 1.61	—	Exploratory
Presence (IPQ per-item), 1–7	VR only, n = 214	—	5.71 ± 1.15	—	Exploratory

Immediately after the Session-2 seminar, the knowledge total was higher in the VR arm (18.33 ± 3.19) than the control arm (13.53 ± 4.35), a difference of 4.80 points (95% CI 4.09–5.52; Hedges' $g = 1.26$ [95% CI 1.06–1.46]). This advantage was retained at four weeks: VR 18.63 ± 3.24 (median 18 [IQR 16–21]) versus control 14.60 ± 4.43 (median 14 [IQR 11–18]), difference 4.03 (95% CI 3.30–4.76; $p < 0.001$; $g = 1.04$ [95% CI 0.84–1.24]; Table 4, Fig. 2B), robust to Mann–Whitney U ($p < 0.001$; probability of superiority 0.75). Within-arm change from the immediate to the four-week test was small in both arms (VR $+ 0.30 \pm 3.86$; control $+ 1.07 \pm 5.02$), indicating retention of the acquired knowledge, and the ANCOVA model adjusting for immediate score and test version confirmed the four-week advantage ($B = 3.23$; 95% CI

2.70–3.77; $p < 0.001$; cluster-robust wild bootstrap $p < 0.001$; teaching-subgroup ICC for the four-week total = 0.22). The advantage was concentrated in urinary catheterization (7.07 ± 1.94 versus 4.12 ± 1.89 ; $g = 1.53$), with small differences for nasogastric (5.28 ± 2.19 versus 4.73 ± 2.22 ; $g = 0.25$) and rectal (6.28 ± 1.71 versus 5.75 ± 1.56 ; $g = 0.33$). Internal consistency of the knowledge test was moderate overall (KR-20 = 0.67 pooled; 0.44 in the VR arm and 0.68 in the control arm) but heterogeneous across versions and subscales (version A 0.34; version B 0.56); the total is therefore best interpreted as an aggregate index of declarative knowledge rather than a strictly unidimensional scale.

Table 4
Key secondary outcomes: four-week knowledge and VR-arm self-report.

Outcome	Sample	VR / Pre	Control / Post	Difference / effect	Role
Knowledge total, immediate (0–30)	VR n = 222 / control n = 220	18.33 ± 3.19	13.53 ± 4.35	4.80 (4.09–5.52); $g = 1.26$ (1.06–1.46)	Secondary
Knowledge total, four-week (0–30)	VR n = 222 / control n = 220	18.63 ± 3.24	14.60 ± 4.43	4.03 (3.30–4.76); $g = 1.04$ (0.84–1.24)	Key secondary
Knowledge urinary, 0–10	VR / control	7.07 ± 1.94	4.12 ± 1.89	2.94 (2.59–3.30); $g = 1.53$ (1.32–1.74)	Secondary
Knowledge nasogastric, 0–10	VR / control	5.28 ± 2.19	4.73 ± 2.22	0.55 (0.14–0.96); $g = 0.25$ (0.06–0.44)	Secondary
Knowledge rectal, 0–10	VR / control	6.28 ± 1.71	5.75 ± 1.56	0.53 (0.23–0.84); $g = 0.33$ (0.14–0.52)	Secondary
Satisfaction (SSSCL), 5–25	VR only, n = 201	22.50 ± 3.33	23.47 ± 2.25	0.97 (0.43–1.50); $dz = 0.25$ (0.11–0.39)	Secondary
Self-confidence (SSSCL), 8–40	VR only, n = 201	33.90 ± 5.47	35.83 ± 4.34	1.93 (1.01–2.85); $dz = 0.29$ (0.15–0.43)	Secondary
Raw NASA-TLX, 0–10	VR only, n = 214	—	2.65 ± 1.61	—	Exploratory
Presence (IPQ per-item), 1–7	VR only, n = 214	—	5.71 ± 1.15	—	Exploratory

Note. Knowledge: Welch's t-tests; for knowledge rows, VR/Pre and Control/Post columns show VR and control group means, respectively. ANCOVA of the four-week total on arm, immediate score, and test version: $B = 3.23$ (95% CI 2.70–3.77; cluster-robust wild bootstrap $p < 0.001$; teaching-subgroup ICC = 0.22). Knowledge KR-20 = 0.67 pooled (0.44 VR; 0.68 control). Reliabilities: satisfaction $\alpha = 0.94$; self-confidence $\alpha = 0.95$; presence $\alpha = 0.92$.

VR-arm self-reported outcomes

Among 201 VR-arm students with complete paired SSSCL data, satisfaction increased from 22.50 ± 3.33 to 23.47 ± 2.25 ($\Delta = 0.97$; 95% CI 0.43–1.50; $p < 0.001$; $dz = 0.25$) and self-confidence from 33.90 ± 5.47 to 35.83 ± 4.34 ($\Delta = 1.93$; 95% CI 1.01–2.85; $p < 0.001$; $dz = 0.29$). The Raw NASA-TLX score was 2.65 ± 1.61 (low workload), and the mean per-item IPQ score was 5.71 ± 1.15 of 7 (high presence). Scale reliabilities were high (satisfaction $\alpha = 0.94$; self-confidence $\alpha = 0.95$; presence $\alpha = 0.92$).

Dissociation and item-level localization of the VR effect (exploratory)

The VR advantage localized to specific checklist components rather than elevating performance uniformly, providing a within-trial test of the alignment account that domain-level effect sizes cannot supply. The composite performance effect also exceeded the four-week knowledge effect ($g = 2.38$ versus 1.04 ; $\Delta g = 1.34$)—an asymmetry consistent with, but weaker than, the item-level evidence below. Because the two outcomes were measured in the same students, we compared the standardized effects with a z-test whose standard error would be conservative under positive cross-domain correlation; empirically, however, individual performance and knowledge scores were essentially uncorrelated within the VR arm ($r = -0.001$), so the assumption of independence was met and the difference was large ($z = 8.35$; $p < 0.001$). This near-zero individual-level correlation is itself consistent with the dissociation: students who performed well procedurally were not those who scored highest on declarative knowledge. The item-level advantage was selective rather than uniform, ranging from -0.11 to $+0.39$ across the 11 checklist steps (Fig. 3). All 11 differences were statistically significant after Benjamini–Hochberg false-discovery-rate correction (adjusted $p < 0.014$), but their magnitudes were highly non-uniform (SD across items = 0.13). The largest advantages arose on documentation ($+0.39$; VR 0.73 versus control 0.34 ; adjusted $p < 0.001$), patient positioning ($+0.29$; adjusted $p < 0.001$), placement verification ($+0.26$; adjusted $p < 0.001$), aseptic technique ($+0.25$; adjusted $p < 0.001$), and safe disposal ($+0.21$; adjusted $p < 0.001$)—steps explicitly and repeatedly choreographed in the VR scenarios—whereas the remaining steps showed small advantages ($+0.10$ to $+0.12$). Critically, in the opposite direction, control students significantly outperformed the VR arm on material preparation (0.84 versus 0.73 ; -0.11 ; cluster-robust $t = -3.2$; adjusted $p = 0.003$), the one checklist step requiring physical handling of real equipment that an immersive environment cannot rehearse. This reversal—a significant control-arm advantage on a single, theoretically predicted component embedded within a large overall VR advantage—is the clearest evidence that the VR effect tracked what the scenarios rehearsed rather than reflecting a global elevation of ability or a generic response bias. Knowledge effects, in turn, tracked scenario dose: the procedure with the longest immersive scenario (urinary, 14 minutes) showed the largest knowledge advantage ($g = 1.53$), while the shorter nasogastric (8 minutes) and rectal (5 minutes) scenarios showed small advantages ($g = 0.25$ and 0.33 , respectively). Together, these patterns indicate that the VR effect followed what the scenarios rehearsed rather than reflecting a uniform elevation of ability.

Discussion

In this cluster randomized controlled study, immersive VR produced a very large advantage in blinded, double-rated OSCE performance ($g = 2.38$) alongside a substantially smaller—though retained—advantage in declarative knowledge ($g = 1.04$ at four weeks), and exploratory analyses located the performance advantage selectively in the checklist steps the VR scenarios explicitly rehearsed, with the single physical-manipulation step favoring the control arm. The central contribution of this trial is therefore not that immersive VR ‘works’, but the first adequately powered, blinded, within-trial demonstration that its effects localize as constructive alignment predicts—concentrated where scenario, objectives, and assessment coincide, and reversing on the one component (material preparation) the immersive environment cannot rehearse. This helps reconcile a contradictory meta-analytic literature, in which VR appears to favor skills in some syntheses and knowledge in others (Hsieh et al., 2025; Lee et al., 2024; Liu et al., 2023): the domain that benefits is the one the scenario rehearses and the assessment samples. The interpretive question thus shifts from whether immersive VR is effective to what its effects are effects of. Because the one-week seminar was delivered identically to both arms, the observed differences are attributable to the Session-1 training modality. Key design features—cluster randomization, delayed rather than immediate assessment, two independent blinded assessors, estimated inter-rater reliability, and cluster-robust analysis—strengthen inference relative to studies relying on immediate assessment or self-report alone, and the findings align in direction with contemporary syntheses reporting benefits of immersive VR for nursing skills and competency (Cho & Kim, 2024; Hsieh et al., 2025; Liu et al., 2023; Park et al., 2024; Vogelsang et al., 2024) and with catheterization-specific studies (Chang, 2022; Park & Yoon, 2023; Yoon et al., 2024).

The magnitude of the composite OSCE effect ($g = 2.38$) exceeds pooled meta-analytic estimates for nursing competency ($g \approx 0.88$; Cho & Kim, 2024) and warrants cautious interpretation. The most plausible contributors are close alignment between the VR scenarios, procedural objectives, and OSCE checklists—such that the assessment closely mirrored the content the VR arm rehearsed—together with ceiling effects, most evident for the nasogastric and rectal stations, and chance imbalance between the randomized teaching subgroups. Consistent with an assessment-alignment mechanism, the knowledge advantage was considerably smaller ($g = 1.04$ at four weeks) and concentrated in urinary catheterization, indicating that the benefit was greater for construct-aligned procedural performance than for declarative knowledge and may be procedure- and scenario-dependent. This dissociation echoes the broader pattern that immersive VR effects on knowledge are smaller and less consistent than effects on skill (Hsieh et al., 2025; Woon et al., 2021), and it underscores the value of interpreting large performance effects against domain-appropriate benchmarks rather than in isolation.

This account gains force from mirror-image findings. A recent randomized trial that used a knowledge-oriented VR session before an identical guideline-based seminar improved theoretical knowledge without changing OSCE performance (Tannemann et al., 2025)—the exact inverse of the present pattern, in which procedure-choreographed VR moved checklist performance far more than knowledge. Read together, the two trials indicate that immersive VR effects follow what the scenario rehearses, with direct consequences for measurement: assessments aligned with the rehearsed choreography will register large effects, while unaligned domains may register little or none (Neher et al., 2025).

Affective and cognitive-load indicators were coherent with the CAMIL account (Makransky & Petersen, 2021): high presence and low perceived workload suggest that the immersive interface imposed limited extraneous load and was usable, while the small pre–post gains in satisfaction and self-confidence ($d_z = 0.25–0.29$) indicate that the principal signal in this study is objective performance rather than self-reported change. Clustering by teaching subgroup (composite ICC = 0.15) did not overturn the between-group difference under cluster-robust wild bootstrap inference, mitigating—though not eliminating—concerns arising from cluster-level randomization.

These findings should be situated within an evidence base in which enthusiasm for immersion has at times outpaced demonstration of transfer to the clinical setting; superior simulated performance does not, by itself, establish transfer to patient care, and the comparative benefit of VR over hands-on simulated practice for psychomotor skill acquisition remains an area of active investigation (Rourke, 2020). The present results are therefore best read as evidence that immersive VR, when scenarios are tightly aligned to competency criteria, can strengthen early procedural performance under standardized assessment, pending evidence of durable retention and clinical transfer.

Implications for simulation-based nursing education

Immersive VR appears most valuable as a preparatory or adjunctive activity that primes students' procedural schemas before supervised physical practice, particularly for procedures demanding clear sequencing, safety checks, and repeated exposure to a standardized exemplar. The results do not support substituting VR for hands-on practice; rather, VR may complement low-fidelity simulation within a blended pathway. Programs adopting VR should align scenarios explicitly with competency checklists, pair immersive rehearsal with deliberate physical practice—particularly for material preparation and equipment handling, the step on which conventional practice outperformed VR here—and design OSCE blueprints that sample beyond the rehearsed choreography to guard against alignment-inflated scores.

Limitations

Several limitations apply. First, randomization was at the level of 32 intact teaching subgroups rather than individuals, and participants could not be blinded to their training modality; although baseline characteristics were similar and cluster-robust analyses supported the findings, a modest number of clusters offers weaker protection against chance imbalance than individual randomization, and because the first knowledge assessment followed training, chance baseline differences in knowledge cannot be fully excluded. Second, although total contact time (two 4-hour sessions) and hands-on practice time (one hour) were matched across arms, the two Session-1 modalities differed in nature, so the contrast reflects two training modalities rather than a controlled manipulation of a single active ingredient. Third, very high VR scores on some stations indicate ceiling effects, which compress variance and inflate standardized effects; medians and non-parametric analyses are therefore reported alongside parametric estimates. Inter-rater reliability was correspondingly very high on the lower-variance control-arm stations, as expected when scores cluster near a boundary. Fourth, the knowledge test showed heterogeneous internal consistency across versions and subscales (KR-20 0.34–0.68), and one urinary item on

indications for indwelling catheterization showed below-chance accuracy in the control arm on both test forms, with response patterns consistent with a persistent shared misconception about appropriate indications (e.g., treating convenience- or incontinence-related use as acceptable) rather than an item artifact; between-group comparisons should therefore be read as aggregate differences in correct responses. A sensitivity analysis excluding this item left conclusions unchanged (four-week difference 3.21 points on the 29-item total; 95% CI 2.48–3.94; $g = 0.82$). The item-level decomposition and the formal effect-size comparison were exploratory rather than prespecified and should be read as hypothesis-generating. In addition, because the knowledge test's internal consistency was lower than that of the double-rated OSCE, differential measurement reliability could attenuate the knowledge effect and thereby inflate the apparent between-domain dissociation; the item-level selectivity within the OSCE—including the reversed material-preparation step—is not susceptible to this artifact and therefore constitutes the stronger evidence for the alignment account. The OSCE instrument's content was developed and reviewed by the teaching team but was not subjected to a separate formal validation study; its use is supported by that content development and by the inter-rater reliability reported here (single-measures ICC 0.80–0.98; average-measures 0.89–0.99 across checklist and global-rating scores, all stations). Fifth, self-report measures were administered only in the VR arm, precluding between-arm comparison. Finally, the OSCE captured short-term performance one week after training; medium- and long-term retention, physical-skill transfer, and clinical transfer were not assessed, and the individual date of the four-week test was not recorded, so the exact window cannot be verified per participant.

Conclusions

Immersive VR produced a large but selective advantage in blinded OSCE performance and a smaller, retained knowledge advantage in undergraduate nursing catheterization training. The advantage concentrated in the checklist steps the VR scenarios rehearsed and, for knowledge, in the procedure with the longest scenario, while the single physical-manipulation step favored conventional practice. These findings support immersive VR as a powerful preparatory complement to hands-on practice—and, equally, they caution that large VR performance effects are partly alignment-dependent and should be interpreted against assessments designed to sample beyond the rehearsed choreography. Individually randomized trials with matched practice dose, longer follow-up, physical-skill transfer outcomes, and deliberately misaligned as well as aligned assessments are warranted.

Declarations

Ethics approval and consent to participate. The study was approved by the institutional research ethics committee (ID: CE012606); participation was voluntary and written informed consent was obtained from all participants.

Consent for publication. Not applicable.

Authors' contributions. María del Carmen Agüera-Bolea: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. Pablo Fernández-Molina: Methodology, Investigation, Data curation, Writing – review & editing. Felipe Jesús Portillo-Castaño: Investigation, Data curation, Resources, Writing – review & editing. Eloina Valero-Merlos: Investigation, Resources, Writing – review & editing. Paula García-Ibáñez: Formal analysis, Methodology, Writing – review & editing. Sergio Navarro-Sánchez: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing.

Acknowledgements. Not applicable.

Funding. This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests. The authors declare no conflicts of interest.

Availability of data and materials. The de-identified dataset supporting the conclusions of this article, together with the study instruments, is available in the Open Science Framework repository (<https://osf.io/6egdq/>). The final protocol and statistical analysis plan are archived as a retrospective registration on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/T5JSW>).

Registration. The trial was retrospectively registered; the project record is archived on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/T5JSW>).

Generative AI. The authors declare that no generative artificial intelligence (AI) tools were used in the preparation, writing, or editing of this manuscript.

References

1. Boursicot K, Kemp S, Wilkinson T, Findyartini A, Canning C, Cilliers F, Fuller R. (2021). Performance assessment: Consensus statement and recommendations from the 2020 Ottawa Conference. *Medical Teacher*, 43(1), 58–67. <https://doi.org/10.1080/0142159X.2020.1830052>
2. Butt AL, Kardong-Edgren S, Ellertson A. Using game-based virtual reality with haptics for skill acquisition. *Clin Simul Nurs*. 2018;16:25–32. <https://doi.org/10.1016/j.ecns.2017.09.010>.
3. Chang C-L. Effect of immersive virtual reality on post-baccalaureate nursing students' in-dwelling urinary catheter skill and learning satisfaction. *Healthcare*. 2022;10(8):1473. <https://doi.org/10.3390/healthcare10081473>.
4. Chen F-Q, Leng Y-F, Ge J-F, Wang D-W, Li C, Chen B, Sun Z-L. Effectiveness of virtual reality in nursing education: Meta-analysis. *J Med Internet Res*. 2020;22(9):e18290. <https://doi.org/10.2196/18290>.
5. Cheng A, Kessler D, Mackinnon R, Chang TP, Nadkarni VM, Hunt EA, Duval-Arnould J, Lin Y, Cook DA, Pusic M, Hui J, Moher D, Egger M, Auerbach M. Reporting guidelines for health care simulation research: Extensions to the CONSORT and STROBE statements. *Simul Healthc*. 2016;11(4):238–48. <https://doi.org/10.1097/SIH.0000000000000150>.

6. Cho M-K, Kim MY. Enhancing nursing competency through virtual reality simulation among nursing students: A systematic review and meta-analysis. *Front Med*. 2024;11:1351300. <https://doi.org/10.3389/fmed.2024.1351300>.
7. Efendi D, Apriliyasari RW, Massie JGEP, Wong CL, Natalia R, Utomo B, Sunarya CE, Apriyanti E, Chen K-H. The effect of virtual reality on cognitive, affective, and psychomotor outcomes in nursing staffs: Systematic review and meta-analysis. *BMC Nurs*. 2023;22:170. <https://doi.org/10.1186/s12912-023-01312-x>.
8. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *BMJ*. 1975;1(5955):447–51. <https://doi.org/10.1136/bmj.1.5955.447>.
9. Hart SG, Staveland LE. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati, editors, *Human mental workload* (pp. 139–183). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
10. Hsieh J-Y, Lin P-C, Sun W-N, Lin T-R, Kuo C-C, Hsu H-T. Effectiveness of immersive virtual reality in nursing education for nursing students and nursing staffs: A systematic review and meta-analysis. *Nurse Educ Today*. 2025;151:106725. <https://doi.org/10.1016/j.nedt.2025.106725>.
11. INACSL Standards Committee. Healthcare Simulation Standards of Best Practice™. *Clin Simul Nurs*. 2021;58:1–4. <https://doi.org/10.1016/j.ecns.2021.08.018>.
12. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide 81. Part I: An historical and theoretical perspective. *Med Teach*. 2013;35(9):e1437–46. <https://doi.org/10.3109/0142159X.2013.818634>.
13. Kyaw BM, Saxena N, Posadzki P, Vseteckova J, Nikolaou CK, George PP, Divakar U, Masiello I, Kononowicz AA, Zary N, Car T, L. Virtual reality for health professions education: Systematic review and meta-analysis by the Digital Health Education Collaboration. *J Med Internet Res*. 2019;21(1):e12959. <https://doi.org/10.2196/12959>.
14. Lee M, Kim S, Go Y, Jeong H, Lee Y, et al. The effectiveness of nursing education using immersive virtual reality or augmented reality: Systematic review and meta-analysis. *J Korean Acad Soc Nurs Educ*. 2024;30(2):109–23. <https://doi.org/10.5977/jkasne.2024.30.2.109>.
15. Liu K, Zhang W, Li W, Wang T, Zheng Y. Effectiveness of virtual reality in nursing education: A systematic review and meta-analysis. *BMC Med Educ*. 2023;23:710. <https://doi.org/10.1186/s12909-023-04662-x>.
16. Makransky G, Petersen GB. The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychol Rev*. 2021;33:937–58. <https://doi.org/10.1007/s10648-020-09586-2>.
17. Neher AN, Bühlmann F, Müller M, Berendonk C, Sauter TC, Birrenbach T. Virtual reality for assessment in undergraduate nursing and medical education – a systematic review. *BMC Med Educ*. 2025;25(1):292. <https://doi.org/10.1186/s12909-025-06867-8>.

18. Park S, Shin HJ, Kwak H, Lee HJ. Effects of immersive technology-based education for undergraduate nursing students: Systematic review and meta-analysis using the GRADE approach. *J Med Internet Res*. 2024;26:e57566. <https://doi.org/10.2196/57566>.
19. Park S, Yoon HG. Effect of virtual-reality simulation of indwelling catheterization on nursing students' skills, confidence, and satisfaction. *Clin Simul Nurs*. 2023;80:46–54. <https://doi.org/10.1016/j.ecns.2023.05.001>.
20. Rourke S. How does virtual reality simulation compare to simulated practice in the acquisition of clinical psychomotor skills for pre-registration student nurses? A systematic review. *Int J Nurs Stud*. 2020;102:103466. <https://doi.org/10.1016/j.ijnurstu.2019.103466>.
21. Schubert T, Friedmann F, Regenbrecht H. The experience of presence: Factor analytic insights. *Presence: Teleoperators Virtual Environ*. 2001;10(3):266–81. <https://doi.org/10.1162/105474601300343603>.
22. Şendir M, Kızıl H, Inangil D, Kabuk A, Türkoğlu İ. Effectiveness of haptic technology in teaching urinary catheterization skill: A randomized controlled study. *Teach Learn Nurs*. 2022;17(1):42–8. <https://doi.org/10.1016/j.teln.2021.08.001>.
23. Slater M. Immersion and the illusion of presence in virtual reality. *Br J Psychol*. 2018;109(3):431–3. <https://doi.org/10.1111/bjop.12305>.
24. Tannemann N, Tsarenko O, Herbstreit F, Gestmann M, Brenner T, Szalai C. Enhancing theoretical BLS knowledge with virtual reality: A randomized controlled trial in medical students. *Resusc Plus Article*. 2025;101169. <https://doi.org/10.1016/j.resplu.2025.101169>.
25. Vogelsang L, Wright S, Risling T, de Padua A, Leidl D, Wilson J, Thompson D. Exploring the use of immersive virtual reality in nursing education: A scoping review. *Clin Simul Nurs*. 2024;97:101648. <https://doi.org/10.1016/j.ecns.2024.101648>.
26. Woon APN, Mok WQ, Chieng YJS, Zhang HM, Ramos P, Mustadi HB, Lau Y. Effectiveness of virtual reality training in improving knowledge among nursing students: A systematic review, meta-analysis and meta-regression. *Nurse Educ Today*. 2021;98:104655. <https://doi.org/10.1016/j.nedt.2020.104655>.
27. Yoon H, Lee E, Kim C-J, Shin Y. Virtual reality simulation-based clinical procedure skills training for nursing college students: A quasi-experimental study. *Healthcare*. 2024;12(11):1109. <https://doi.org/10.3390/healthcare12111109>.

Figures

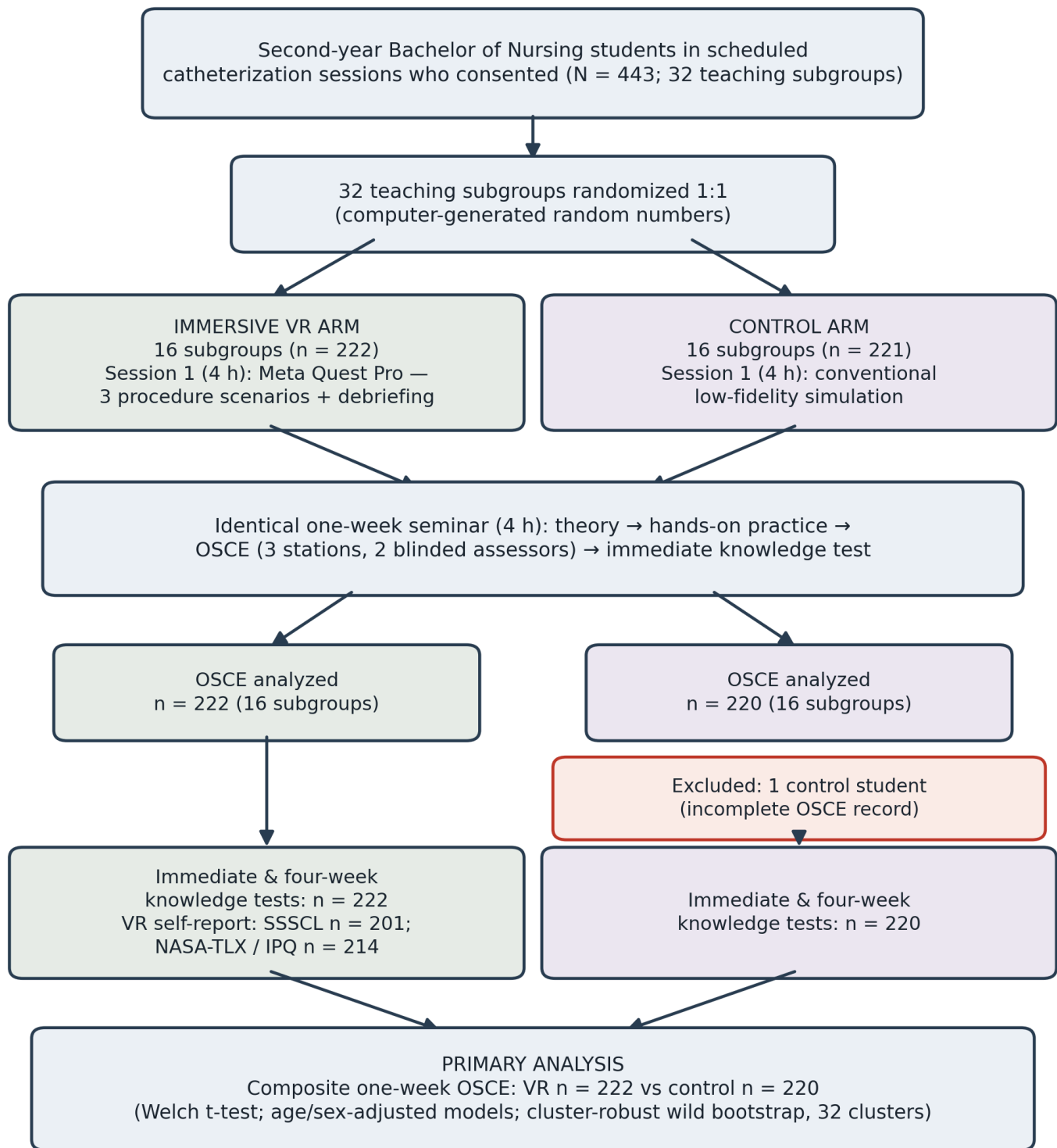


Figure 1

Participant flow (CONSORT cluster/simulation-research extension). Cluster randomization of 32 intact teaching subgroups (1:1; 16 per arm); the one-week seminar was identical across arms, which differed only in the Session-1 training modality.

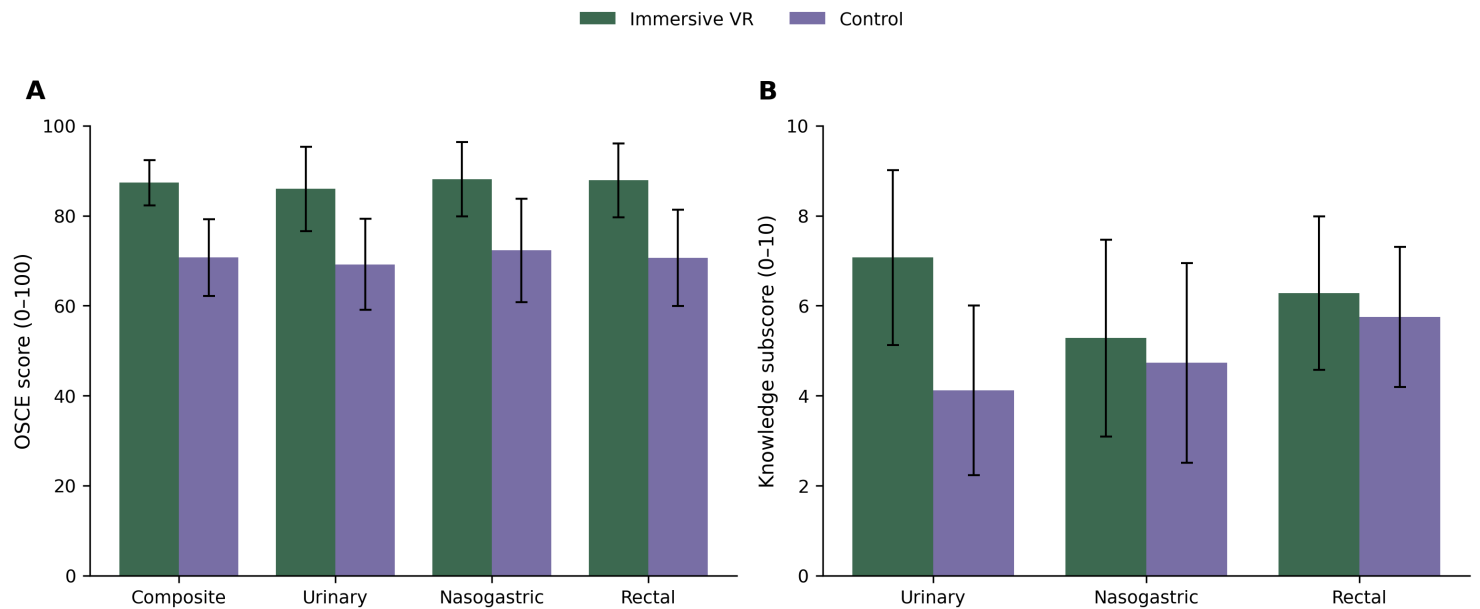


Figure 2

Performance and knowledge outcomes by arm (mean ± SD; Hedges' g). (A) One-week OSCE scores (0–100) for the composite and each procedure. (B) Four-week knowledge subscores (0–10) by procedure.

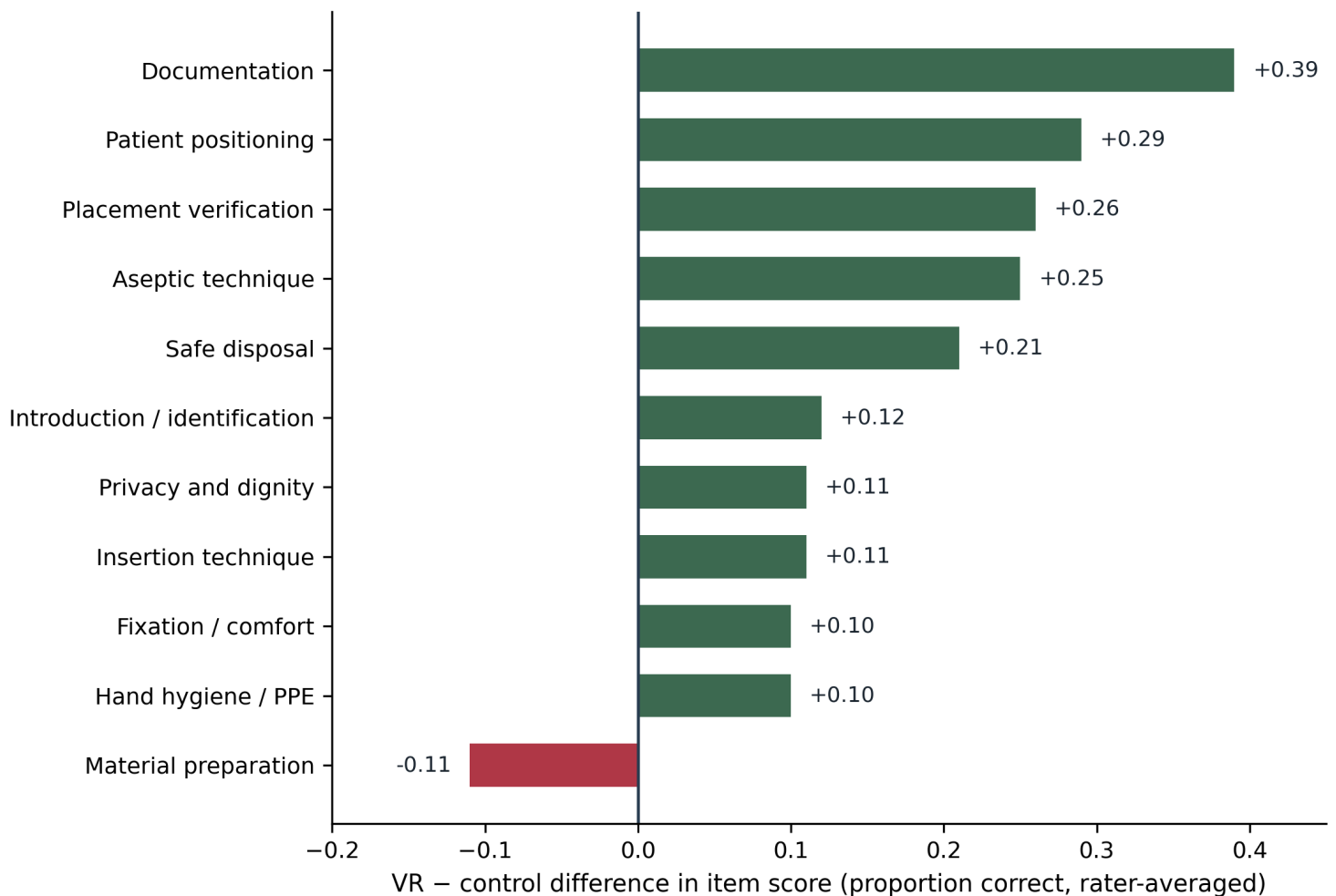


Figure 3

Item-level localization of the VR effect (exploratory). VR – control difference in rater-averaged checklist item scores (proportion correct), pooled across the three stations and sorted by magnitude; all differences were significant after false-discovery-rate correction, and the single negative value indicates a significant control-arm advantage on material preparation (adjusted $p = 0.003$).

Supplementary Files

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

- [SupplementInstructorscoringguide.pdf](#)
- [SupplementKnowledgetestFormsABstudent.pdf](#)
- [SupplementPlantilladocentetestconocimientos.pdf](#)
- [SupplementTestconocimientosFormasABestudiante.pdf](#)
- [Supplementaryinstruments.pdf](#)