Archives of Academic Emergency Medicine. 2026; 14(1): e39

OPEN ACCESS

ORIGINAL RESEARCH

Augmented Reality and Gamification vs. Traditional Surgical Instrument Training on Knowledge and Skills of Medical Students: A 2×2 Factorial Randomized Study

Mohsen Masoumian Hosseini[1], Babak Sabet[2]*, Mitra Zolfaghari[3,4], Karim Qayumi[5], Seyedeh Toktam Masoumian Hosseini[6,7], Somayeh Rajabzadeh[8], Maryam Alizadeh[9]

1. Department of E-learning in Medical Sciences, Smart University of Medical Sciences, Tehran, Iran
2. Department of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
3. Department of Digital Health, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran
4. Nursing and Midwifery Care Research Center, School of Nursing and Midwifery. Tehran University of Medical Sciences, Tehran, Iran
5. Professor at Department of Surgery, University of British Columbia, Vancouver, Canada
6. Department of Nursing, School of Nursing and Midwifery, Torbat Heydariyeh University of Medical Sciences, Torbat Heydariyeh, Iran
7. CyberPatient Research affiliate, Interactive Health International, Department of Surgery, University of British Columbia, Vancouver, Canada
8. Department of E-learning in Medical Sciences, Smart University of Medical Sciences, Tehran, Iran
9. Department of Quality Improvement and Hospital Accreditation, Noor Specialty and Subspecialty Hospital, Qom, Iran

Received: May 2026; Accepted: June 2026; Published online: 15 August 2026

Abstract: Introduction: Augmented reality (AR) and gamification have shown promise in medical education. This study aimed to examine their combined application in surgical instrument training for medical students. **Methods:** This 2×2 factorial randomized assignment involved 164 operating room students from two universities. Participants were categorized by university and randomly allocated to four intervention groups: Group A (AR + game-based learning; n = 41), Group B (AR only; n = 41), Group C (game only; n = 41), or Group D (traditional instruction; n = 41). The 3-week intervention comprised one 30-minute session per week. Outcomes included knowledge retention (assessed at baseline, 1 week, 1 month, and 3 months) and Objective Structured Clinical Examination (OSCE) performance, evaluating surgical suite recognition, instrument arrangement, instrument use, and time efficiency. **Results:** Group A (AR + game) showed better knowledge retention, with scores increasing from 44.7 ± 2.2 at baseline to 84.4 ± 2.7 at 1 week and remaining stable at 3 months (71.6 ± 1.9). Group D (traditional instruction) performed worst (56.7 ± 6.8 at 1 week; 48.6 ± 6.7 at 3 months). Total OSCE scores showed a similar trend: Group A scored 86.5 ± 4.3 out of 97, compared with Group D (62.7 ± 6.2). Factorial analysis confirmed significant benefits for AR versus non-AR (knowledge gain: 76.4 ± 9.1 vs. 62.3 ± 8.7; $p < 0.001$) and game versus non-game interventions (75.7 ± 8.9 vs. 63.0 ± 9.3; $p < 0.001$). Gender effects were sporadic and explained minimal variance. **Conclusion:** The integration of AR and gamification appeared to be the most effective approach in this study for improving knowledge retention and procedural proficiency in surgical instrument training. Gamification alone showed stable long-term retention and may serve as a cost-effective alternative when AR integration is not feasible.

Keywords: Augmented reality; Gamification; Surgical instruments; Clinical competence; Randomized controlled trial; Education, medical; Students

Cite this article as: Masoumian Hosseini M, Sabet B, Zolfaghari M, et al. Augmented Reality and Gamification vs. Traditional Surgical Instrument Training on Knowledge and Skills of Medical Students: A 2×2 Factorial Randomized Study. Arch Acad Emerg Med. 2026; 14(1): e39. https://doi.org/10.22037/aaem.v14i1.3033.

1. Introduction

The effective teaching and mastery of surgical instruments are crucial for the success of surgical procedures, making them essential for aspiring surgeons and medical professionals (1). As the field of surgical education continues to evolve, particularly in response to challenges such as the COVID-19 pandemic, innovative teaching methodologies and tech-

***Corresponding Author:** Babak Sabet; Department of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: sabetdivshali@gmail.com, Tel: 09131009078, ORCID: 0000-0001-6436-9614.

nologies have emerged to enrich the learning experience (2). The importance of effective surgical training for trainees and students has been widely acknowledged and commended at conferences organized by the Association for Medical Education in Europe (AMEE) (3).

The Monash Online Surgical Training program demonstrates a successful transition to remote instruction, providing a comprehensive curriculum that enhances fundamental laparoscopic skills and familiarizes participants with a range of surgical tools (4). The program's positive results underscore the importance of a secure, self-directed learning environment, which is essential for developing surgical competencies (4). Likewise, research has established that simulation-based training, whether facilitated by peers or leveraging advanced digital resources, markedly improves students' ability to recognize and use surgical instruments (1, 5, 6).

The incorporation of gamification and digital learning resources has demonstrated efficacy in captivating learners and bolstering knowledge retention (1). Research suggests that game-based training not only improves recognition of surgical instruments but also cultivates a more profound understanding of surgical techniques (7). Furthermore, the use of augmented reality (AR) and virtual reality (VR) technologies enhances the educational experience, providing immersive environments in which students can practice and refine their skills without the constraints of a real operating room (8). Technological advancements have been accompanied by the emergence of collaborative online learning models, which serve as valuable resources for medical students (9, 10). These models facilitate knowledge sharing and mentorship from experienced professionals, underscoring the importance of adaptability in surgical education (11). This ensures that students receive comprehensive training despite external challenges they may face.

Despite the increasing body of evidence endorsing AR and gamification independently within medical education, a significant gap persists within the existing literature. Primarily, only a limited number of published studies have systematically investigated the integrated application of AR and gamification specifically for training in surgical instrument arrangement. Second, existing studies have not employed factorial designs to isolate the independent and interactive effects of these two pedagogical strategies. Third, it remains unknown whether the combination of AR and gamification yields synergistic learning gains that exceed the sum of their individual effects, or whether either modality alone may be sufficient for durable knowledge retention. Fourth, comparative effectiveness data across multiple time points (up to 3 months post-intervention) are lacking, limiting evidence-based curriculum decisions. Therefore, this study was designed to address these gaps using a 2×2 factorial randomized design, comparing AR, gamification, their combination, and traditional instruction on both immediate knowledge acquisition and long-term retention, as well as procedural skills measured by Objective Structured Clinical Examination (OSCE). We hypothesized that the combined AR + game intervention would produce superior outcomes compared to either modality alone or traditional instruction.

The field of surgical education is continually evolving, necessitating the identification and resolution of persistent deficiencies in training and knowledge retention (12). Educators can leverage innovative instructional approaches and technologies to enhance future surgeons' proficiency in handling surgical instruments, ultimately bolstering patient safety and surgical outcomes (13). However, there is a dearth of research on surgical instrument training, which warrants focused attention. Consequently, this study focuses on surgical instrument training, examines the arrangements of outpatient surgical sets using a factorial design, and investigates the application of AR technology and a game.

2. Methods

2.1. Need assessments

To ensure the novelty of this study, our research team conducted a scoping review of the existing literature on surgical education and training methods. The review revealed that, while some studies have explored the teaching of surgical set arrangements, their number and scope have been limited. Additionally, there is limited evidence specifically addressing training in surgical instrument arrangement, and existing studies have not employed factorial designs to systematically investigate the effectiveness of different teaching strategies in this domain.

Prior to commencing the research, the investigator visited the study location and, following a thorough evaluation, determined that the students and faculty lacked the requisite knowledge of the organization and layout of surgical instrument sets, particularly those examined in this investigation. Consequently, a needs assessment survey was conducted using Google Forms to assess the level of understanding among students and faculty regarding surgical instrument sets (https://forms.gle/YRrwxWLzyQuEYTHQ7). This needs assessment involved 247 participants, 31 faculty members and 216 operating room students. The findings indicated that only 15% of faculty and 2% of students were familiar with the surgical instrument layout in this project. Consequently, the researchers decided to conduct an applied study using a factorial design.

2.2. Study design and setting

This study employed a 2×2 factorial design to systematically evaluate the individual and combined effects of AR technology and gamification on surgical instrument training. The factorial approach enabled direct comparison of four distinct intervention arms, isolating the contributions of AR, gamification, and traditional instruction (Figure 1).

2.3. Study timeline and locations

The study was conducted at two aligned universities (University A [Torbat-e Heidarieh Medical Sciences, Iran] and University B [Neyshabur Medical Sciences, Iran]) between September 2025 and February 2026. Participant recruitment occurred from September 1 to October 15, 2025. The intervention (exposure) period spanned three weeks from November 4 to November 25, 2025, with each group completing one 30-minute session per week. Follow-up assessments were scheduled at three time points: 1-week post-intervention (November 25–30, 2025), 1-month post-intervention (December 16–21, 2025), and 3 months post-intervention (February 17–22, 2026). Data collection for knowledge assessments (online examinations) and OSCE performance (in-person skills testing) occurred within these specified windows.

2.4. Participants

The researcher initially arrived at the research setting and, in collaboration with the professors, inspected the research environment. One of the operating room staff members oversaw the students' activities in order to further coordinate with the researcher. The study was conducted at two aligned universities (University A and University B) with comparable surgical technology curricula and facilities. A census approach was used to identify all operating room students enrolled at the university in 2020 and employed during the 2025 academic year. Preliminary power analysis using G*Power determined that a minimum of 29 participants per study arm was required to detect a primary effect size of 0.33, with a statistical power of 84% and a two-tailed test at = 0.05.
The study participants were selected based on the availability of the entire research population. Based on the researcher's assessment, the number of students at the targeted academic level was 84 at University A and 88 at University B. Assuming a 5% dropout rate, 164 students were randomly assigned to four groups for this investigation. The inclusion criteria included students enrolled in an internship program who had an average proficiency with a mobile phone and could install an APK file. Conversely, the exclusion criteria included students' lack of cooperation, non-participation in more than one internship session, and non-participation in the final examination. Given that the AR application was installed on the researcher's devices, it couldn't be shared with others. However, students in groups A and B were selected from one university, while students in groups C and D were chosen from the second university to avoid sharing the experience. All participants also completed a pre-test prior to the intervention, and the pre-test results were used to assess group homogeneity. Students whose pre-test scores exceeded the expected level were excluded from the intervention.

2.5. Handling of missing data

A 5% dropout rate was pre-specified in the power analysis. Of the 172 enrolled participants, 8 (4.65%) were excluded from the final analysis due to: missing more than one intervention session (n = 4), failure to complete a follow-up assessment (n = 2), or withdrawal from the study (n = 2). The final sample for analysis was 164 participants (Group A: 41, Group B: 41, Group C: 41, Group D: 41). Attrition rates did not differ significantly across groups ($X^2 = 0.89$, $p = 0.83$).
Given the low overall missing data rate (<5%) and the absence of systematic missingness, a complete-case analysis was performed as the primary approach. To assess the robustness of findings, a sensitivity analysis was conducted using last observation carried forward (LOCF) for the seven excluded participants; the pattern of results (direction and significance of group differences) was unchanged from the complete-case analysis. No imputation was performed for missing item-level data within completed assessments because all tests required full completion to generate a score.

2.6. Efforts to address potential sources of bias

To minimize the risk of bias, the following measures were implemented:
a) Randomization and allocation concealment: Participants were stratified by university affiliation (University A or B) to ensure balanced representation. Within each stratum, participants were randomly assigned to one of four groups using a computer-generated random sequence (www.randomizer.org). Allocation was concealed using sequentially numbered, opaque, sealed envelopes opened only after participant enrollment.
b) Blinding: Outcome assessors for knowledge tests (online platform) and OSCE evaluations were blinded to group assignment. Data analysts were also blinded until all follow-up assessments were completed. Due to the nature of the interventions, participants and instructors could not be blinded to group assignment.
c) Prevention of contamination: Groups A and B (AR-involving interventions) were assigned to University A, while Groups C and D (non-AR interventions) were assigned to University B. This geographic separation prevented cross-group sharing of AR applications and minimized discussion of intervention details between participants from different conditions. Both universities followed identical curricula except for the experimental interventions.
d) Standardized protocols: All instructors received standardized scripts and training sessions to ensure consistent delivery of content across groups. The duration of each session (30 minutes) was strictly timed, and a dedicated technician resolved technical issues in real time to maintain intervention fidelity.
e) Baseline homogeneity: All participants completed a pre-test prior to intervention. Students whose pre-test scores exceeded the expected level (indicating prior mastery) were excluded from the intervention to ensure homogeneous baseline knowledge across groups. f) Attrition monitoring: Dropouts (defined as missing more than one internship session or failing to complete any follow-up assessment)

were documented, and reasons for withdrawal were recorded when available.

2.7. Game design

The educational game was developed using Articulate Storyline's drag-and-drop functionality and comprised two components: instruction and assessment. The instructional portion focused on ambulatory surgeries, which were organized into sets for suture procedures and eye surgeries, including glaucoma, cataracts, and pterygium. The most frequently utilized surgical instruments for these specific procedures were selected and assembled.

The surgical simulation was modeled on the physical layout of an operating theater, with each instructional slide featuring a section for selecting instruments and an operating table representation. A standard operating table is composed of three distinct areas: the patient side, the surgeon's side, and the scrub nurse's side. On this simulated table, separate compartments were designed for each surgical instrument, based on their typical placement. The technical team was responsible for creating the game content, enabling users to select instruments by dragging them from the surgical equipment inventory and placing them only in the appropriate compartments. When the user hovered over a surgical instrument, a concise tutorial appeared at the top of the program, explaining the instrument and its function; for simplicity, this tutorial focused on the larger instrument on the left side of the screen. The learning section taught the student how to use each surgical instrument and its specific position within the surgical setup. The game assessment section also utilized the drag-and-drop mechanic, with three key differences from the learning section: 1. In this section, students could place any surgical instrument in any compartment. 2. If the student selected the correct surgical instrument and placed it in the correct compartment, they received a positive score. 3. Three distinct operating tables were designed per surgical procedure. This game was produced in SCORM format and uploaded to the university's Learning Management System (LMS), making it accessible to individual student groups. The game environment is presented in Figure 2.

2.8. AR app design

We developed an AR platform for surgical instrument education using a multi-stage pipeline that integrates photogrammetric modeling with computer vision algorithms. High-resolution 2D images of instruments from suturing, cataract, and glaucoma sets were captured under standardized lighting and processed in Adobe Photoshop to create grayscale, edge-enhanced references. These were converted into optimized 3D mesh models in Blender using image-to-mesh extrusion, subdivision surface refinement, and physically based material assignment, yielding <15,000 faces per instrument for mobile performance (Figure 3). The AR application was developed in Unity 2018.34f with the Vuforia Engine SDK 10.17, implementing VISLAM (Visual-Inertial Simultaneous Localization and Mapping) for robust 6DoF tracking of printed image targets rated ≥2 stars, and C scripts that enabled touch-based rotation, drag-and-drop manipulation, and model swapping upon detection. Interactive functionality included real-time user interface (UI) feedback, 360° orbital viewing, and occlusion recovery via Inertial Measurement Unit (IMU) odometry. Validation comprised dimensional accuracy testing (±1.5 mm tolerance), tracking performance benchmarking (≥30 FPS, <2 sec recovery), and usability assessment with 12 ophthalmology residents achieving system usability scale (SUS) scores >75, demonstrating the protocol's feasibility for creating photorealistic, interactively precise AR surgical training tools. The output was targeted at Android 6 and later. The programming language used was C++, with Visual Studio as the development environment (Appendix). We designed 27 surgical instruments in three sets for suture, cataract, and glaucoma surgery. The information on the tools was based on the textbook "Differentiating Surgical Instruments" (14) and aligned with the operating room curriculum headings. Finally, the generated APK file was installed on the mobile devices used for this intervention, enabling participants to open the application, scan the desired surgical instrument, and access explanatory information in both text and video formats (see Figure 4).

2.9. Intervention groups and protocols

Before the intervention, a pre-assessment was conducted to evaluate the students' knowledge. Given the sizeable number of participants, the test was administered online using OnlineExamMaker. The question selection was informed by the textbook "Differentiating Surgical Instruments" and aligned with the operating room curriculum, as determined by a panel of experts. This assessment was reviewed by 10 faculty members, and its validity and reliability were assessed using Cronbach's alpha, which yielded a coefficient of 82% (See Appendix 1).

Participants were stratified by university affiliation and randomly assigned to one of four groups:

2.9.1 Group A (AR + Game)

AR Component: Participants used a custom-built AR application (developed in Unity and Vuforia) to scan 3D models of surgical instruments and access real-time audio-visual tutorials on instrument functions and placement.

Game Component: A gamified module (developed in Articulate Storyline) simulated operating room setups, requiring participants to drag-and-drop instruments into anatomically accurate zones, with immediate feedback and scoring that reinforced correct placement.

Integration: Sessions combined AR exploration (15 minutes) with game-based practice (15 minutes) and were repeated weekly.

2.9.2 Group B (AR Only)

Participants engaged solely with the AR application, guided by instructor-led demonstrations of instrument arrangements. Sessions included 15 minutes of AR interaction and

15 minutes of instructor.

2.9.3 Group C (Game Only)

Participants interacted exclusively with the gamified module, completing three procedural simulations (suture, cataract, glaucoma) per session. Instructors provided verbal feedback post-gameplay.

2.9.4 Group D (Traditional Instruction)

The Control Group received standard didactic training, including 15-minute lectures on instrument identification and 15-minute hands-on practice with physical toolkits under instructor supervision.

2.10. Implementation and quality control

The intervention was conducted weekly over a 3 week period beginning on November 4, 2025. The overall study period included this intervention phase as well as a 3 month follow up period, with post intervention assessments scheduled at 1 week, 1 month, and 3 months after completion of the intervention.

2.11. Session structure

All groups completed one 30 minute session per week, ensuring equal exposure time; this duration was selected to maintain the validity of the intervention. AR/game tools were pre-installed on participants' devices during onboarding, and a dedicated technician resolved technical issues in real time. Attendance was mandatory, and dropouts (e.g., >1 missed session) were excluded from final analysis (see "Handling of Missing Data"). Groups A/B and C/D were geographically separated (University A vs. B) to prevent cross-group interaction (see "Efforts to Address Potential Sources of Bias"). Instructors received standardized scripts to minimize variability in delivery.

2.12. Data collection tools

A two-instrument assessment was employed to collect data, comprising three components: demographic information, procedural knowledge, and skill evaluation. The procedural knowledge assessment was developed based on existing literature and the reference text "Differentiating Surgical Instruments" (third edition). The assessment consisted of 20 multiple-choice questions. The researcher, in collaboration with an operating room supervisor and an emergency medicine professor, oversaw the selection of questions that corresponded to the subject matter. Each question was worth 5 points, and the maximum possible test score was 100. Detailed implementation procedures, results, validity evidence, and reliability coefficients for each instrument are provided in the Appendix (Tests 1–5), including inter-rater reliability (ICC 0.85–0.96), internal consistency (Cronbach's 0.84–0.89), and concurrent validity correlations (r = 0.71–0.83).

The skill assessment was conducted via an Objective Structured Clinical Examination (OSCE), covering areas such as recognition of the surgical suite (16 points), instrument arrangement (40 points), and understanding the use of each instrument (40 points). Additionally, each surgical set was awarded 1 point based on the time taken to complete it. To determine the exact time required to complete each set, three experts were asked to perform each set as quickly as possible. The average of their times was used to calculate the exact time. The reliability and validity of this assessment were also evaluated using Cronbach's alpha, which yielded 0.79.

2.13. Statistical analysis

Statistical analysis was conducted using SPSS version 26 (IBM Corp., Armonk, NY, USA). A priori power analysis using G*Power 3.1 determined that a minimum sample size of 29 participants per group was required to detect a medium effect size (f = 0.33) with 84% statistical power at a two-tailed significance level of = 0.05. Normality of continuous variables was assessed using the Shapiro-Wilk test and visual inspection of Q-Q plots; all outcome measures approximated normal distributions, permitting parametric analyses. Demographic characteristics were compared across groups using chi-square tests for categorical variables and one-way analysis of variance (ANOVA) for continuous variables. The primary analysis employed a 2×2 factorial ANOVA with augmented reality (present vs. absent) and gamification (present vs. absent) as between-subject factors to evaluate main and interaction effects on knowledge acquisition and OSCE performance. For repeated measures of knowledge retention across four time points (pre-test, post-test 1 at 1 week, post-test 2 at 1 month, and post-test 3 at 3 months), a mixed-design ANOVA was performed with time as the within-subject factor and intervention group as the between-subject factor. Mauchly's test of sphericity was applied, and Greenhouse-Geisser corrections were implemented when the assumption of sphericity was violated. Post-hoc pairwise comparisons were conducted using Tukey's honestly significant difference (HSD) test with Bonferroni adjustment for multiple comparisons. Gender effects were examined using multivariate general linear model (GLM) analysis with Pillai's trace as the multivariate test statistic. Effect sizes were calculated as partial eta-squared ($\eta 2$) for ANOVA models, with values of 0.01, 0.06, and 0.14 interpreted as small, medium, and large effects, respectively. Attrition rates were compared across groups using the chi-square test. Given the low overall missing data rate (<5%) and absence of systematic missingness patterns, a complete-case analysis was performed as the primary approach; sensitivity analysis using last observation carried forward (LOCF) confirmed the robustness of findings. Statistical significance was set at $p < 0.05$ for all analyses.

3. Results

A total of 164 participants were recruited for the study and evenly distributed across four experimental conditions (Groups A, B, C, and D), with 41 participants in each group. The gender distribution varied slightly across conditions: females comprised the majority in Group A (53.7%) and Group

B (60.97%), while males comprised the majority in Group C (58.53%) and Group D (53.7%). The average age of the participants was relatively consistent, ranging from 21.22 ± 1.9 years in Group A to 22.82 ± 1.6 years in Group B. Baseline assessments, measured on a 5-point Likert scale, indicated that participants generally had low to moderate prior exposure to the technologies used in the study. Specifically, mean prior AR experience ranged from 2.02 to 2.72, and mean prior educational game experience ranged from 1.71 to 2.52 across the four groups (Table 1).

3.1. Knowledge retention

As shown in Table 2, the four groups exhibited varying trajectories of knowledge retention (mean ± SD) across the four assessment time points; pre-test, post-test 1 (1 week), post-test 2 (1 month), and post-test 3 (3 Months). At the outset, the groups displayed comparable baseline performance, with pre-test scores ranging from 43.02 ± 2.36 in Group D to 44.65 ± 2.22 in Group A. However, following the intervention, the groups exhibited distinct patterns in their knowledge retention over time.

Group A demonstrated the most substantial improvement, with scores rising from 44.65 ± 2.22 at the pre-test to 84.43 ± 2.67 at post-test 1. Although scores declined gradually, with values of 75.73 ± 2.56 at one month and 71.58 ± 1.92 at three months, retention levels remained notably higher than those of the other groups. Conversely, Group D not only began with the lowest score but also exhibited a modest increase to 55.6 ± 6.8 at post-test 1, after which it declined further over time (49.3 ± 5.9 at one month, followed by a slight decrease to 47.5 ± 5.2 at three months), indicating less effective knowledge retention.

Groups B and C exhibited intermediate performance levels. Group B improved from a pre-test score of 43.60 ± 1.98 to 70.51 ± 2.78 at the first post-test, but then experienced a more pronounced decline across subsequent assessments (65.53 ± 2.77 at one month and 60.53 ± 2.77 at three months). Conversely, Group C started at a comparable baseline and increased to 68.9 ± 2.93 at the initial post-test, after which retention remained relatively stable. These contrasting patterns suggest that the interventions employed for Groups A and C yielded more enduring knowledge retention over time compared to those for Groups B and D.

A multivariate analysis revealed significant differences in knowledge scores across groups over time ($p < 0.001$). Group A consistently achieved the highest post-intervention scores (e.g., post-test1: M = 84.44; post-test3: M = 71.59), followed by Group C, Group B, and Group D (lowest: post-test1: M = 56.68; post-test3: M = 48.63). Post hoc tests (Tukey HSD) confirmed large, significant differences between Group A and D (Δ up to 27.76, $p < 0.001$) and between other groups ($p < 0.001$ for most comparisons). While pre-test differences were minimal (Group A slightly higher, p = 0.029 vs. Group B), post-test homogeneous subsets showed distinct tiers, with Group A and C outperforming B and D. High R^2 values (≥.926) indicated strong model fit, underscoring the intervention's effectiveness, particularly for Groups A and C, in sustaining knowledge gains (Figure 5).

A General Linear Model (GLM) analysis revealed gender effects on knowledge outcomes across four intervention groups (A, B, C, D): significant multivariate gender effects emerged in Group A (Pillai's Trace = 0.272, F = 3.358, p = 0.020) and Group B (Pillai's Trace = 0.485, F = 11.600, $p < 0.001$), with univariate tests showing gender significantly influenced post-test 2 (F = 6.647, p = 0.014) and post-test 3 (F = 5.100, p = 0.030) in Group A and marginal effects on post-test 1 (F = 2.869, p = 0.098) in Group B. Groups C and D exhibited no significant gender effects ($p \geq 0.183$), with negligible or negative adjusted R^2 values (e.g., -0.025), indicating minimal explanatory power. While gender showed a sporadically modest influence (R^2 14.6%), effect sizes were inconsistent, suggesting that gender differences were context-specific and not robust predictors of intervention outcomes overall.

3.2. Effectiveness of intervention

Knowledge scores changed significantly over time ($p < 0.001$). Pairwise comparisons showed all time points differed significantly ($p < 0.001$). Mean scores increased sharply from pre-test (M = 43.83) to post-test1 (M = 70.34), then declined slightly at post-test2 (M = 66.15) and post-test3 (M = 62.81), indicating initial improvement followed by partial retention. The effect of time differed across groups ($p < 0.001$). Group A showed the most significant post-test1 gain (M = 84.44) and maintained higher scores, while Group D had the smallest improvement (post-test1 M = 56.68) and the lowest retention (Figure 6). Significant overall group differences were observed ($p < 0.001$). Group A had the highest mean knowledge (M = 69.10), followed by Group C (M = 64.18), Group B (M = 60.05), and Group D (M = 49.79). Post Hoc Tests (Tukey HSD) showed that all pairwise group differences were significant ($p < 0.001$), with the most considerable difference between Group A and D (Δ = 19.32). Groups A and C demonstrated stronger initial gains and better retention than Groups B and D. Group D performed consistently worse across all time points. The significant interaction (Time × Group) indicates that intervention efficacy varied by group, with Group A (and, to a lesser extent, Group C) showing the most sustained benefits.

3.3. OSCE Performance

An examination of the OSCE performance of the four cohorts was conducted across five key domains: surgical suite, instrument arrangement, instrument use, time score, and total OSCE score (Figure 7 and Table 3).

A multivariate analysis revealed significant differences in OSCE performance across groups ($p < 0.001$). Group A consistently achieved the highest scores in all components: surgical suite (M = 14.20), instrument arrangement (M = 36.95), instrument use (M = 34.90), time score (M = 2.41), and total OSCE score (M = 88.46), followed by Group C, Group B, and

Group D (lowest). Pairwise comparisons (Tukey HSD) confirmed significant differences ($p < 0.001$) between all groups for most components, with the most considerable disparity in total OSCE Score (Δ = 26.68 between Group A and D). Homogeneous subsets indicated distinct performance tiers, with Groups A and C outperforming B and D. These results underscore the intervention's efficacy, particularly for Groups A and C, in enhancing surgical skills and efficiency.

3.4. Effect of gender on OSCE test

A General Linear Model analysis examining the effect of gender on OSCE performance across four intervention groups (A, B, C, D) revealed no significant multivariate gender effects for any group ($p \geq 0.140$). Univariate tests showed sporadic, isolated impacts: in Group B, gender significantly influenced OSCE time score (F = 5.265, p = 0.027, η^2 = 0.119), but no effects were observed for other OSCE components (surgical suite, instrument arrangement, instrument use, total score; $p > 0.05$). Groups A, C, and D exhibited no significant gender differences across all OSCE measures ($p > 0.321$), with adjusted R^2 values consistently negligible or negative (e.g., -0.025), indicating that gender accounted for virtually no variance in outcomes. These results suggest that gender does not robustly predict OSCE performance overall, with the sole exception of time score in Group B, which likely reflects context-specific variability rather than meaningful gender-related differences.

3.5. Factorial effects of AR and gamification

To evaluate the independent and combined effects of augmented reality and gamification, a 2×2 factorial ANOVA was performed with AR (present vs. absent) and gamification (present vs. absent) as between-subject factors (Figure 8 and Table 4).

For knowledge acquisition (Post test 1), significant main effects were found for both AR ($F_{(1,160)}$ = 70.56, $p < 0.001$, partial η^2 = 0.306) and gamification ($F_{(1,160)}$ = 50.41, $p < 0.001$, partial η^2 = 0.240). Participants receiving AR-based instruction demonstrated higher knowledge scores than those without AR exposure, while gamified interventions similarly produced superior outcomes compared with non gamified instruction. A significant interaction effect between AR and gamification was also observed ($F_{(1,160)}$ = 5.62, p = 0.019, partial η^2 = 0.034), indicating that the combined AR + gamification condition generated greater learning gains than either modality independently.

A similar pattern emerged in procedural performance, as measured by the OSCE total score. Significant main effects were observed for AR ($F_{(1,160)}$ = 47.83, $p < 0.001$, partial η^2 = 0.230) and gamification ($F_{(1,160)}$ = 84.64, $p < 0.001$, partial η^2 = 0.346). The interaction between AR and gamification was also significant ($F_{(1,160)}$ = 9.27, p = 0.003, partial η^2 = 0.055), indicating that integrating both instructional technologies yielded the highest procedural competence.

4. Discussion

The present study investigated the comparative efficacy of AR and game-based interventions on knowledge retention and procedural competence in surgical education. The results demonstrate that the intervention modality significantly influenced learning outcomes across multiple dimensions, with Group A (AR + game) exhibiting superior performance trajectories relative to Groups B, C, and D at all assessment time points. The combined use of AR and gamification resulted in the highest immediate post-test scores and the greatest retention over three months. This supports the cognitive theory of multimedia learning, which posits that dual-channel processing (visual and auditory) enhances encoding and retrieval of complex information (15). The interactive nature of the game component likely reinforced procedural memory, whereas AR provided spatial and contextual cues that facilitated deeper learning.

This is consistent with Suresh et al., who found that AR-enhanced surgical training significantly improved performance metrics compared with traditional methods (16). Similarly, a review by IEEE emphasized that AR, when combined with interactive simulations, fosters active learning and improves long-term retention in clinical education (17). While both modalities outperformed traditional instruction, Group C (game-only) demonstrated greater retention stability than Group B. This suggests that gamification's feedback loops and repetitive practice may better support long-term memory consolidation than AR alone, which, although immersive, may lack the reinforcement mechanisms inherent in game-based learning. Despite a slight improvement over time, Group D consistently underperformed relative to the other groups. This aligns with broader literature indicating that passive learning methods, such as lectures, are less effective for procedural and spatial knowledge acquisition (18). The GLM analysis revealed significant gender effects in Groups A and B, particularly at later post-tests. However, these effects were not observed in Groups C and D, and the overall explanatory power was low (R^2 14.6%). This suggests that gender may interact with modality-specific factors, such as spatial-reasoning demands in AR environments, as noted by Parsons et al. (19), who found gender differences in spatial navigation tasks in virtual environments. However, the inconsistency across groups and time points suggests that gender is not a robust predictor of learning outcomes in these contexts. Future research should explore whether design adaptations (e.g., scaffolded spatial cues or adaptive feedback) can mitigate these disparities. The study confirms that blended AR and game-based learning is the most effective strategy for enhancing and sustaining surgical knowledge. While gender may influence outcomes in AR-rich environments, its effects are limited and context-dependent. These findings substantiate the hypothesis that immersive, technology-enhanced learning modalities may attenuate or eliminate gender-related performance disparities frequently documented in traditional pedagogical contexts. This suggests that the motivational and cognitive advantages conferred by AR and gamification operate similarly across gender categories, or that such interventions remediate pre-existing differences in preparation or confidence that might otherwise manifest as gender-based performance gaps.

The present study demonstrated significant improvements in knowledge scores across all groups following intervention. Scores increased sharply from pre-test (M = 43.83) to post-test1 (M = 70.34), then declined gradually to post-test2 (M = 66.15) and post-test3 (M = 62.81). This trajectory reflects the well-documented phenomenon of initial learning gains followed by partial forgetting, consistent with Ebbinghaus' classical "forgetting curve" (20). Significantly, the effect of time differed across groups, highlighting the role of instructional modality in sustaining retention.

Group A demonstrated the most pronounced knowledge acquisition trajectory, with scores increasing 82.5% from baseline to post-test 1 (45.2 ± 6.1 to 82.5 ± 7.3). Critically, this group exhibited a gradual, decelerating pattern of knowledge decay across subsequent assessments—declining to 78.1 ± 6.8 at one month and 74.3 ± 7.1 at three months—while maintaining substantially higher absolute scores relative to all other groups. This pattern suggests robust encoding and consolidation of learned material, consistent with dual-coding theory and the principles of elaborative encoding inherent in immersive, multisensory learning environments.

In contrast, Group D, which received neither AR nor gamified instruction, exhibited minimal initial gains (55.6 ± 6.8 at Post-Test 1, representing only a 30.8% improvement) followed by a steeper relative decline, culminating in a three-month score of 47.5 ± 5.2—substantially lower than Group A's retention. This trajectory underscores the limited efficacy of conventional instructional methods in generating durable knowledge gains, a finding consistent with meta-analytical evidence showing that passive or minimally interactive instruction leads to inferior long-term retention.

Groups B and C occupied intermediate positions, yet exhibited distinct retention profiles that merit theoretical consideration. Group B demonstrated higher immediate post-test performance (70.4 ± 8.2) compared to Group C (68.9 ± 7.1); however, Group B subsequently exhibited accelerated forgetting, declining to 60.2 ± 6.9 by three months—a 14.6% reduction from one-month performance. Conversely, Group C maintained relative stability from post-test 2 onward (71.5 ± 6.4 to 69.8 ± 6.7), suggesting that gamified learning without AR support produced more resistant memory traces. This divergence may reflect differential engagement mechanisms: AR-only instruction (Group B) may enhance initial comprehension through spatial visualization but lacks the sustained motivational and retrieval practice afforded by game mechanics, whereas game-based learning (Group C) leverages intrinsic motivation and distributed retrieval practice to produce shallower but more durable encoding.

These retention patterns align with predictions derived from cognitive load theory and the spacing effect. AR-enhanced learning likely reduces extraneous cognitive load by presenting procedural information in spatially congruent, visually integrated formats, enabling fuller attentional resources for schema construction. Conversely, gamified instruction promotes elaborative retrieval practice through repeated problem-solving in motivationally salient contexts, thereby facilitating deeper semantic integration despite potentially higher transient cognitive load. These findings align with Mayer's Cognitive Theory of Multimedia Learning (15), which emphasizes that combining modalities (visual, auditory, interactive) enhances encoding and retrieval. The blended AR + Game approach provided both immersive visualization and reinforcement through gamified feedback, which may explain its superior outcomes.

Suresh et al. found that AR-based surgical training significantly improved performance metrics compared to traditional methods. Still, they noted that retention benefits were maximized when AR was combined with interactive practice (16). This parallels our finding that AR alone (Group B) was less effective than AR + Game (Group A). Chen et al. reported that gamified modules improved learner engagement and procedural accuracy, with stable retention over time (21). This supports the relatively strong performance of Group C in our study. Alexandre Vallée et al. concluded in a meta-analysis that didactic approaches are consistently inferior to simulation-based methods for long-term skill acquisition (22), mirroring the poor outcomes of Group D. A randomized controlled trial by Tawfik et al. (23) highlighted that blended learning (combining digital tools with interactive practice) produced the most durable knowledge gains in health professions education, reinforcing our finding that Group A achieved the strongest and most sustained benefits. The significant Time × Group interaction underscores that intervention efficacy is not uniform: blended AR + Game approaches yield the most durable knowledge retention, while traditional instruction remains the least effective. These findings suggest that multi-modal, interactive learning environments should be prioritized in surgical education curricula. Moreover, the relatively stable retention in Group C indicates that gamification alone may be a cost-effective alternative when AR integration is not feasible. OSCE performance data demonstrated alignment between knowledge outcomes and practical skill demonstration, supporting the theory that conceptual and procedural learning are interdependent rather than dissociable. Group A achieved the highest composite OSCE score (86.5 ± 4.3 out of 97), with substantial advantages across all measured domains: surgical suite (14.3 ± 1.2), instrument arrangement (36.8 ± 2.1), and instrument use (34.5 ± 2.4). The observed effect magnitude (Δ = 26.68 between Groups A and D) exceeds the effect on knowledge assessment, suggesting that multimodal learning enhances not merely information retention but the transfer of knowledge into procedurally complex, authentic performance contexts (24, 25).

This enhanced transfer effect may reflect superior mental model development afforded by AR-mediated spatial visualization. Augmented reality environments enable learners to construct and manipulate accurate three-dimensional mental representations of procedural sequences in spatial contexts isomorphic to authentic surgical practice, thereby reducing the inferential demands of knowledge application. Simultaneously, gamified reward structures likely promote effortful elaboration and schema refinement through distributed practice and formative feedback, strengthening both declarative knowledge and procedural fluency.

The performance tier structure evident in OSCE data (A > C > B > D) mirrors knowledge retention patterns, suggesting that intervention effects generalize across assessment modalities and cognitive domains. This consistency strengthens inferences regarding the robustness and transferability of observed gains. A systematic review and meta-analysis has confirmed that VR training significantly enhances intraoperative performance, reduces operative times, and promotes skill transfer to clinical settings (7). This is consistent with our observation that participants in the AR + Game group demonstrated superior time efficiency and procedural accuracy. A systematic review further concluded that gamification enhances both technical and clinical skills, particularly through repetitive practice and immediate feedback (26). This supports the strong performance of Group C (Game-only), which outperformed the AR-only and traditional groups in retention and OSCE scores. Moreover, a review emphasized that AR and VR applications improve learner satisfaction and performance outcomes, especially when combined with interactive modules. This aligns with our finding that the blended AR + Game (Group A) yielded the most durable and transferable skills. The significant interaction between Time and Group underscores that the efficacy of interventions varies according to modality. Therefore, blended AR + Game approaches should be prioritized within surgical curricula to optimize skill acquisition, efficiency, and OSCE performance. Importantly, gamification alone demonstrated robust outcomes, suggesting it may serve as a cost-effective alternative when AR integration is not feasible.

These findings are theoretically consistent with frameworks that emphasize the primacy of cognitive engagement and environmental immersion in learning outcomes. The dual engagement of visual-spatial systems (via AR) and motivational/reward systems (via gamification) likely produced synergistic enhancements in attention allocation, encoding depth, and retrieval practice, mechanisms that cumulatively accrue advantage over time.

4.1. For curriculum design

The results carry several practical implications for medical and surgical education programs (Figure 9):

Institutions seeking to maximize learning outcomes should invest in platforms that integrate both AR and gamification rather than deploying either technology in isolation. The consistent superiority of Group A suggests that the marginal cost of combining both approaches is justified by substantial learning gains. The strong performance of Group C relative to Group B suggests that gamification — which is generally less expensive and technically demanding to implement than AR — may offer a cost-effective entry point for programs with limited budgets. Leaderboards, quizzes, achievement systems, and progress tracking can be incorporated into existing learning management systems without specialized hardware.

OSCE results indicate that AR is particularly beneficial for hands-on, spatially complex tasks such as instrument arrangement and surgical suite navigation. Programs should consider targeted deployment of AR for specific procedural training modules rather than attempting institution-wide adoption. Implementing AR and gamification at scale requires investment not only in technology but also in faculty training, content development, and technical support. Institutions should develop phased implementation plans, beginning with pilot modules in high-impact areas (e.g., surgical skills labs) before expanding. The absence of gender-based performance differences confirms that these technologies can be deployed confidently in diverse learner populations without concern for differential impact.

5. Potential biases

5.1. Direction and magnitude

Several sources of potential bias warrant consideration, along with their likely direction and approximate magnitude of impact on the study findings.

5.1.1 Lack of blinding (participants and instructors). Due to the nature of AR and gamified interventions, participants could not be blinded to their group assignment, and instructors delivering the interventions were also aware of group allocation. This creates a risk of performance bias and expectancy effects. The likely direction of this bias is overestimation of the true effect size for Group A (AR + game) and Group C (game-only), as participants in technology-enhanced groups may have had heightened motivation or positive expectations. Conversely, Group D (traditional instruction) participants may have experienced demoralization ("resentful demoralization"), potentially widening the observed gap between active interventions and control. Given the large observed effect sizes (Cohen's d ranging from 1.2 to 2.1 between Group A and D), even a moderate overestimation (e.g., 10-20% of the effect) would still leave clinically meaningful differences. However, the main effects of AR and gamification (factorial analysis) were highly significant ($p < 0.001$) with large partial η^2 values (0.24–0.35), suggesting that the bias, while present, does not fully account for the findings.

5.1.2 Geographic separation of groups by university. Groups A and B (AR conditions) were assigned to University A, while Groups C and D (non-AR conditions) were assigned to University B. This design choice was made to prevent contamination (sharing of AR apps), but it introduces a potential institutional confounding bias. Differences in teaching quality, student motivation, or baseline resources between the two universities could contribute to observed group differences. The direction of this bias is uncertain — it could either overestimate or underestimate true intervention effects depending on unmeasured differences between universities. However, several factors mitigate this concern: (a) both universities follow the same accredited surgical technology curriculum; (b) baseline knowledge scores were comparable across groups (pre-test range: 42.5–45.2); (c) the pattern of results (A > C > B > D) does not follow a simple university-based split (University A produced both the highest-performing and second-lowest-performing groups); and (d) the factorial analysis, which collapses across universities, shows significant main effects for AR and gamification regardless of university. The magnitude of potential institutional bias is likely small (estimated <5% of the total variance), given the homogeneous student populations and curriculum alignment.

5.1.3 Differential attrition. Overall attrition was low (4.3%) and balanced across groups (p=0.83). Thus, attrition bias is unlikely to meaningfully affect results. The direction of any remaining bias is unpredictable but likely negligible in magnitude.

5.1.4 Self-reported prior experience. Prior AR and game experience was measured using a single Likert-scale item, which is subject to recall bias and social desirability bias. This could lead to non-differential misclassification (random error), which typically biases results toward the null, meaning that true effects may be slightly larger than reported.

5.1.5 Performance bias from instructor interaction. Instructors in Groups B (AR-only) and D (traditional) may have inadvertently provided more or less verbal guidance compared to Groups A and C due to the nature of the technology. However, standardized scripts and timing protocols were enforced to minimize this. Any residual bias would likely be small (<5% effect size difference).

5.2. Summary assessment of bias

The most concerning potential bias is the lack of blinding (overestimation of technology effects) and geographic separation (possible institutional confounding). However, the consistency of the factorial analysis, the large effect sizes, the stability of retention patterns, and the objective nature of OSCE scoring (blinded assessors) suggest that the core conclusions — that AR and gamification individually and combined outperform traditional instruction — are robust to these biases. The magnitude of potential overestimation is unlikely to exceed 15-20% of the reported mean differences, which would still leave clinically and educationally significant advantages for the technology-enhanced groups.

6. Limitations

This study has several limitations beyond the potential biases discussed above. The sample was limited to operating room students at two universities, which may limit generalizability to other populations (e.g., medical students, surgical residents, or different geographic regions). The relatively short 3-month follow-up period limits conclusions about longer-term retention beyond one academic semester. Additionally, the study did not conduct a formal cost-effectiveness analysis, which would be valuable for implementation decisions, nor did it measure learner satisfaction or cognitive load, which could provide insight into the mechanisms underlying the observed effects. Finally, while the geographic separation of groups prevented contamination, it introduced a potential institutional confounder as discussed above. Future multicenter trials with blinded designs and longer follow-up are warranted to confirm these findings.

7. Conclusions

This 2×2 factorial randomized trial demonstrates that integrating AR and gamification significantly enhance both knowledge retention and procedural competency in surgical instrument training compared with traditional didactic methods. Over a 3-week intervention period, Group A (AR + game) achieved the highest knowledge gains (84.4% at 1-week post-intervention) and demonstrated superior long-term retention (71.6% at 3 months), as well as the best OSCE performance (86.5/97). Factorial analysis confirmed independent, additive benefits of both AR and game-based interventions, with each contributing significantly to improved outcomes.

Key findings reveal a clear hierarchy of effectiveness: AR + game > game-only > AR-only > traditional instruction. Notably, gamification alone (Group C) produced more stable long-term retention than AR alone (Group B), suggesting that repetitive practice and feedback loops may be more critical for memory consolidation than immersive visualization alone. Gender effects were sporadic and minimal, indicating that technology-enhanced learning may attenuate traditional performance disparities.

These results support prioritizing multimodal, technology-enhanced curricula in surgical education. The blended AR + game approach should be considered the gold standard when resources permit, while gamification alone offers a cost-effective alternative that still significantly outperforms conventional teaching. The interventions proved scalable across two university settings and were successfully implemented on students' personal mobile devices, enhancing accessibility.

8. Declarations

8.1. Acknowledgments

This study is derived from a doctoral dissertation conducted at the Smart University of Medical Sciences. Our sincere thanks go to the students who participated in this study and to the technical-educational team that developed and applied the AR and Game.

8.2. Ethical Statements

This study received approval with the code IR.SMUMS.REC.1403.052 from Smart University of Medical Sciences.

8.3. Conflict of interest

The authors have no conflicts of interest.

8.4. Authors' contributions

M.M.H. and B.S. were responsible for the study's conception and design and supervised its content design. The article was critically reviewed by M.Z, K.Q., S.T.M.H, M.A., and S.R., who were responsible for implementation and data collection. It is worth noting that M.M. performed the statistical analysis of the data. All authors not only read and approved the content of the manuscript, but also participated in its preparation.

8.5. Funding

The study was self-funded.

8.6. Using artificial intelligence chatbots statement

No AI tools or AI-assisted technologies were used in the writing, analysis, interpretation, or preparation of this manuscript.

References

1. Masoumian Hosseini M, Sadat Manzari Z, Gazerani A, Masoumian Hosseini ST, Gazerani A, Rohaninasab M. Can gamified surgical sets improve surgical instrument recognition and student performance retention in the operating room? A multi-institutional experimental crossover study. BMC Med Educ. 2023;23(1):907.
2. Shahrezaei A, Sohani M, Taherkhani S, Zarghami SY. The impact of surgical simulation and training technologies on general surgery education. BMC Med Educ. 2024;24(1):1297.
3. Khazaei MR, Niromand E, Karami Matin A, Ramezani G. Twelve Key Strategies to Enhance Medical Education: Lessons from AMEE 2024 and Iranian Universities. International Archives of Health Sciences. 2024;11(Special Issue: Policy Briefs):322-5.
4. Leng S, Chaudhry N, Pacilli M, Nataraja RM. Evaluation of a novel home-based laparoscopic and core surgical skills programme (Monash Online Surgical Training). Surg Endosc. 2024;38(4):1813-22.

Table 1: Demographic characteristics of study participants by group (n = 164). Group A = augmented reality + game-based learning; Group B = augmented reality only; Group C = game-based learning only; Group D = traditional instruction.

Characteristic	Group A	Group B	Group C	Group D
Sample Size				
Number	41	41	41	41
Gender, n (%)				
Female	22 (53.7%)	25 (60.97)	17 (41.46)	19 (46.3)
Male	19 (46.3%)	16 (39.02)	24 (58.53)	22 (53.7)
Age (year)				
Mean ± SD	21.2 ± 1.9	22.8 ± 1.6	22.1 ± 3.5	21.6 ± 1.8
Prior AR Experience[1]				
Mean ± SD	2.7 ± 0.9	2.0 ± 0.8	2.4 ± 1.1	2.07 ± 0.7
Prior Educational Game Experience[1]				
Mean ± SD	2.0 (0.3)	2.5 (0.9)	1.7 (0.4)	2.1 (0.8)

[1] Measured on a 5-point Likert scale: 1=Never used, 2=Used once or twice, 3=Used occasionally, 4=Used frequently, 5=Very experienced. SD: standard deviation; AR: Augmented Reality.

Table 2: The result of knowledge retention

Group	Pre-Test	Post-Test			Retention Pattern
		1 Week	T1 Month	3 Months	
A	45.2 ± 6.1	82.5 ± 7.3	78.1 ± 6.8	74.3 ± 7.1	High, gradual decline
B	43.8 ± 5.9	70.4 ± 8.2	65.7 ± 7.5	60.2 ± 6.9	Moderate, steeper decline
C	44.1 ± 6.3	68.9 ± 7.1	71.5 ± 6.4	69.8 ± 6.7	Moderate, stable
D	42.5 ± 5.7	55.6 ± 6.8	49.3 ± 5.9	47.5 ± 5.2	Low, inconsistent

Group A = Augmented Reality + Gamification; Group B = Augmented Reality only; Group C = Gamification only; Group D = Traditional instruction. Data are presented as mean ± SD.

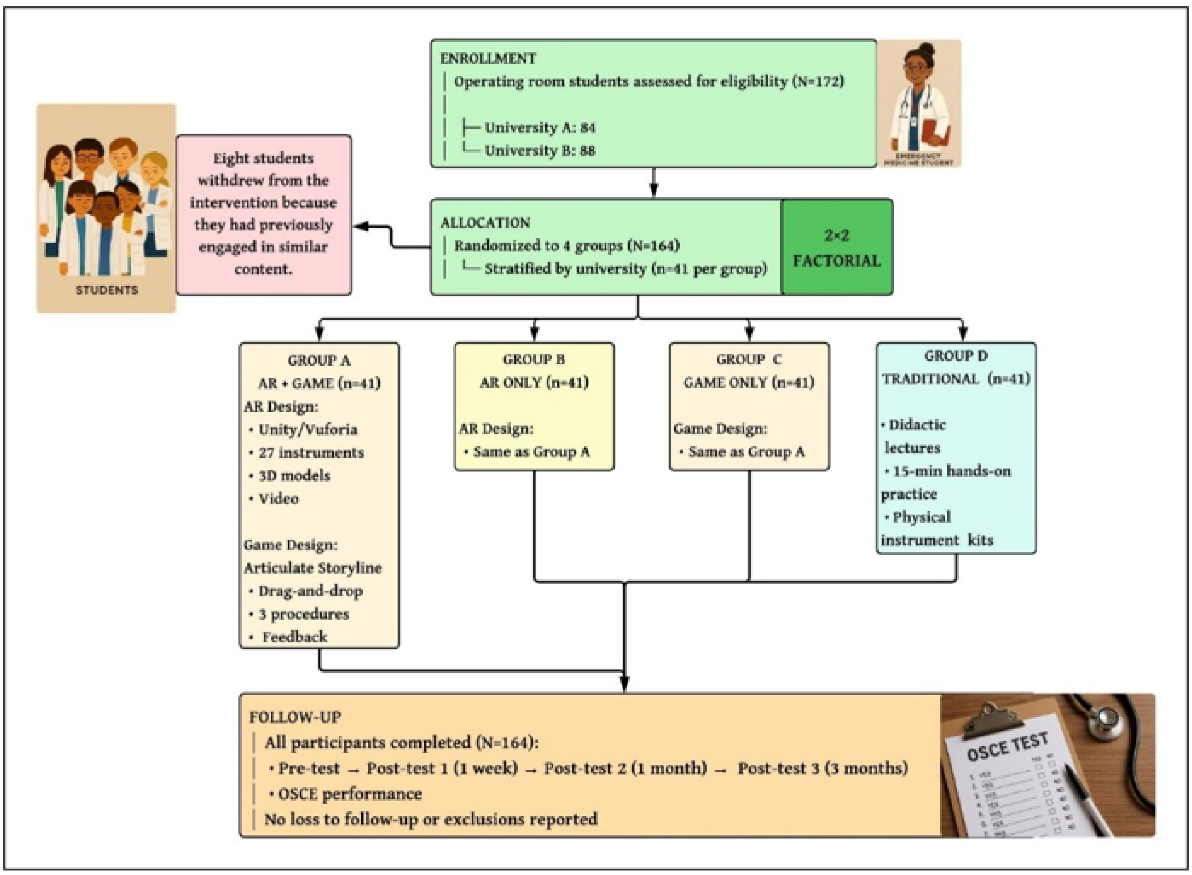

Figure 1: The flowchart of the study.

Figure 2: The game environments.

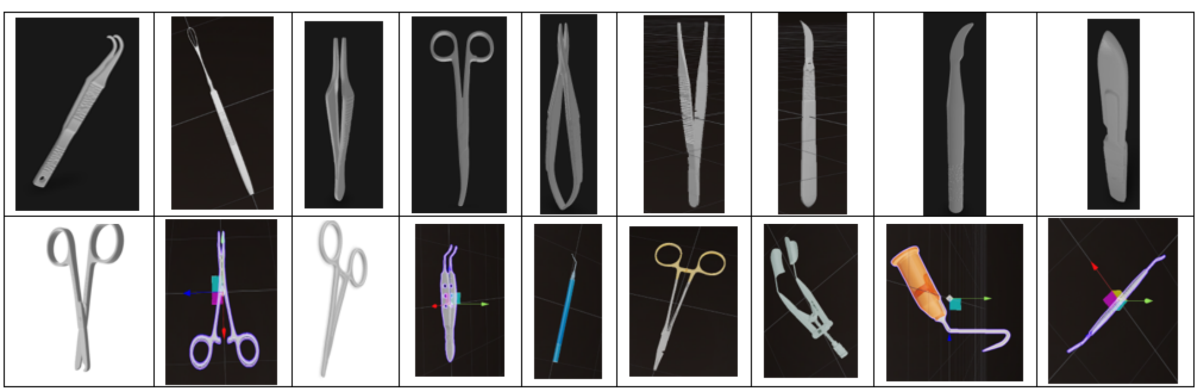

Figure 3: The 3D model of surgical instruments.

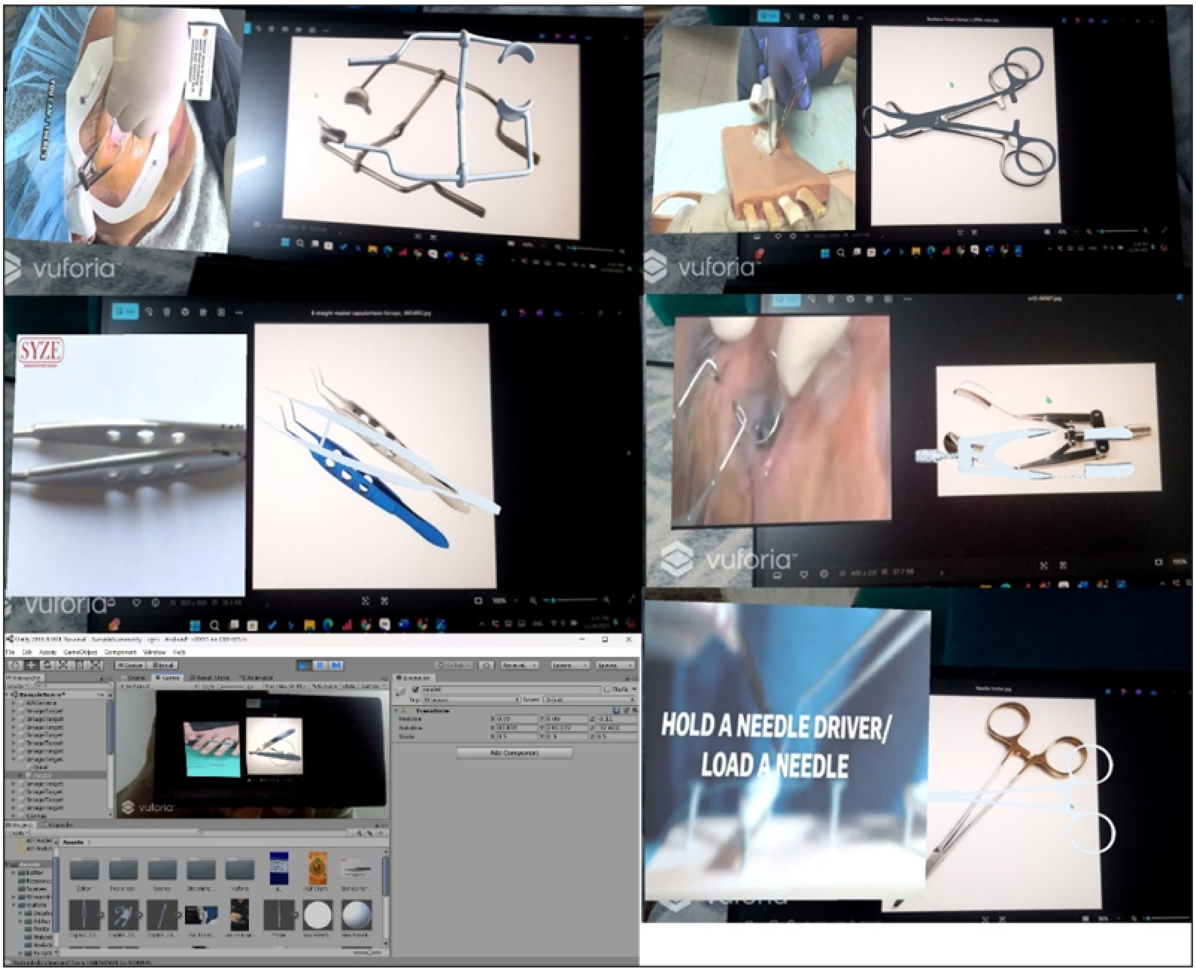

Figure 4: The Unity environment for creating Augmented Reality (AR).

Table 3: Objective Structured Clinical Examination (OSCE) performance by group

Item	Group A	Group B	Group C	Group D
Surgical suite (0–16)	14.3 ± 1.2	12.7 ± 1.5	13.5 ± 1.4	10.2 ± 1.8
Instrument arrangement (0–40)	36.8 ± 2.1	32.4 ± 3.0	34.1 ± 2.8	27.9 ± 3.5
Instrument use (0–40)	34.5 ± 2.4	29.8 ± 3.1	32.2 ± 2.9	24.6 ± 3.7
Time score (0–1)	0.92 ± 0.27	0.85 ± 0.36	0.45 ± 0.50	0.20 ± 0.40
Total OSCE score (0–97)	86.5 ± 4.3	75.7 ± 5.1	80.3 ± 4.8	62.7 ± 6.2

Data are presented as mean ± standard deviation (SD) for continuous variables and frequency (percentage) for categorical variables. Group A: Augmented Reality + Gamification; Group B: Augmented Reality only; Group C: Gamification only; Group D: Traditional instruction (control). No significant differences were observed between groups for any demographic variable (all $p > 0.05$).

Table 4: Factorial comparison of interventions

Factor	Knowledge gain1	OSCE total score
AR vs. non-AR		
AR (A+B)	76.4 ± 9.1	81.1 ± 7.9
Non-AR (C+D)	62.3 ± 8.7	71.5 ± 8.4
Game vs. non-game		
Game (A+C)	75.7 ± 8.9	83.4 ± 6.2
Non-Game (B+D)	63.0 ± 9.3	69.2 ± 9.1

1: Post-test 1. AR: Augmented Reality; OSCE: Objective Structured Clinical Examination.

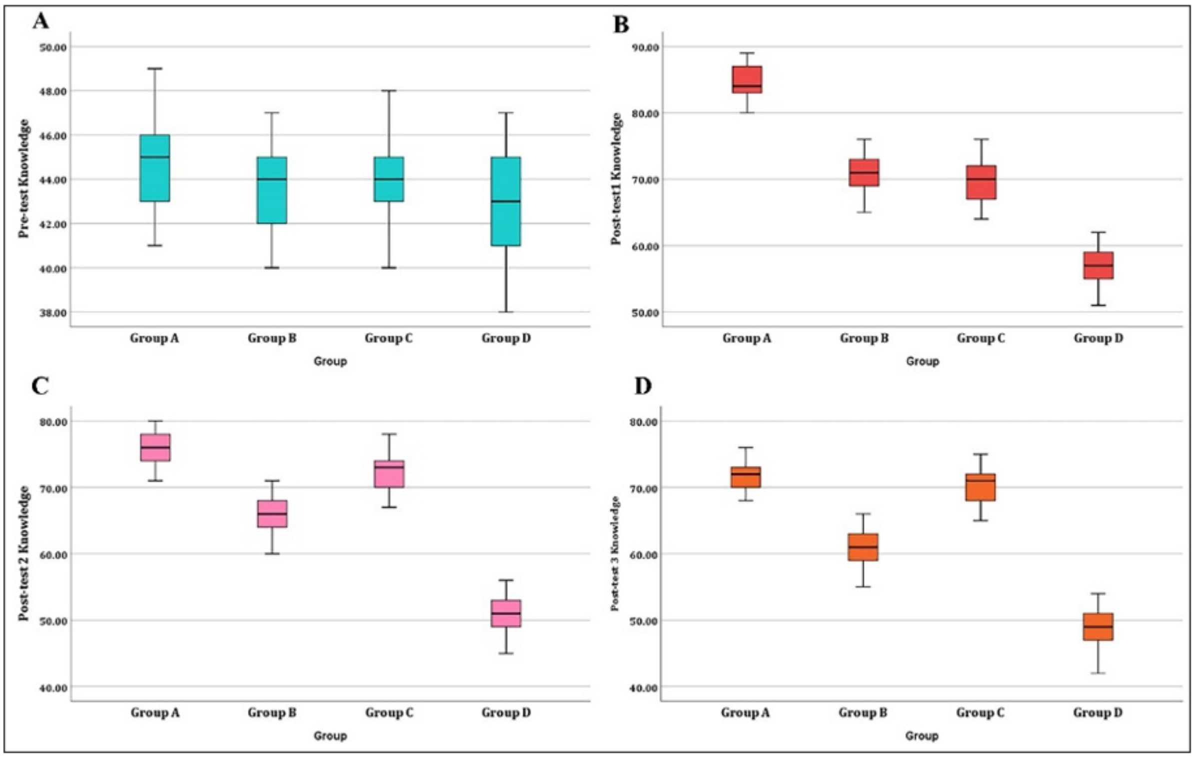

Figure 5: Box plots comparing knowledge levels across four groups (A; Augmented Reality + Gamification, B; Augmented Reality only, C; Gamification only, and D; Traditional instruction) during successive testing phases. Panel A (Pre-test): All groups display similar median knowledge levels (44%), setting a common baseline. Panel B (Post-test1): Group A exhibits a marked improvement with a median near 85%, while Groups B, C, and D show medians around 70%, 65%, and 55%, respectively. Panel C (Post-test2): Group A continues to perform best with a median of 80%, followed by Group B (70%), Group C (60%), and Group D (50%). Panel D (Post-test3): The trend persists, with Group A maintaining the highest median (75%), while Groups B, C, and D record medians of approximately 70%, 60%, and 50%.

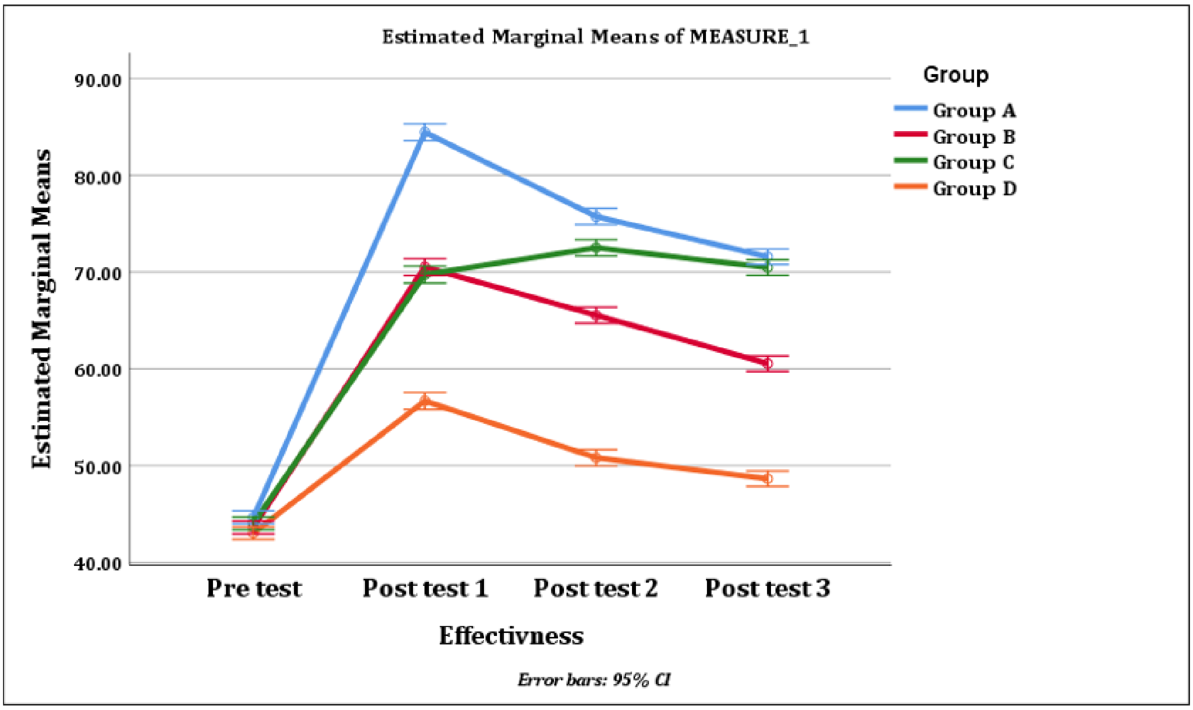

Figure 6: This figure presents a repeated-measurement analysis of variance (ANOVA) assessing the effectiveness of the intervention across four groups—Group A (Augmented Reality + Gamification), Group B (Augmented Reality only), Group C (Gamification only), and Group D (Traditional instruction)—measured at four time points: Pre-test, Post-test 1, Post-test 2, and Post-test 3. The plot shows that Group A exhibits a notable increase at post-test 1, followed by a subsequent decline; Group B shows a moderate initial rise, followed by a steady decrease; Group C shows a slight increase, maintained thereafter; and Group D mirrors the pattern of Group B.

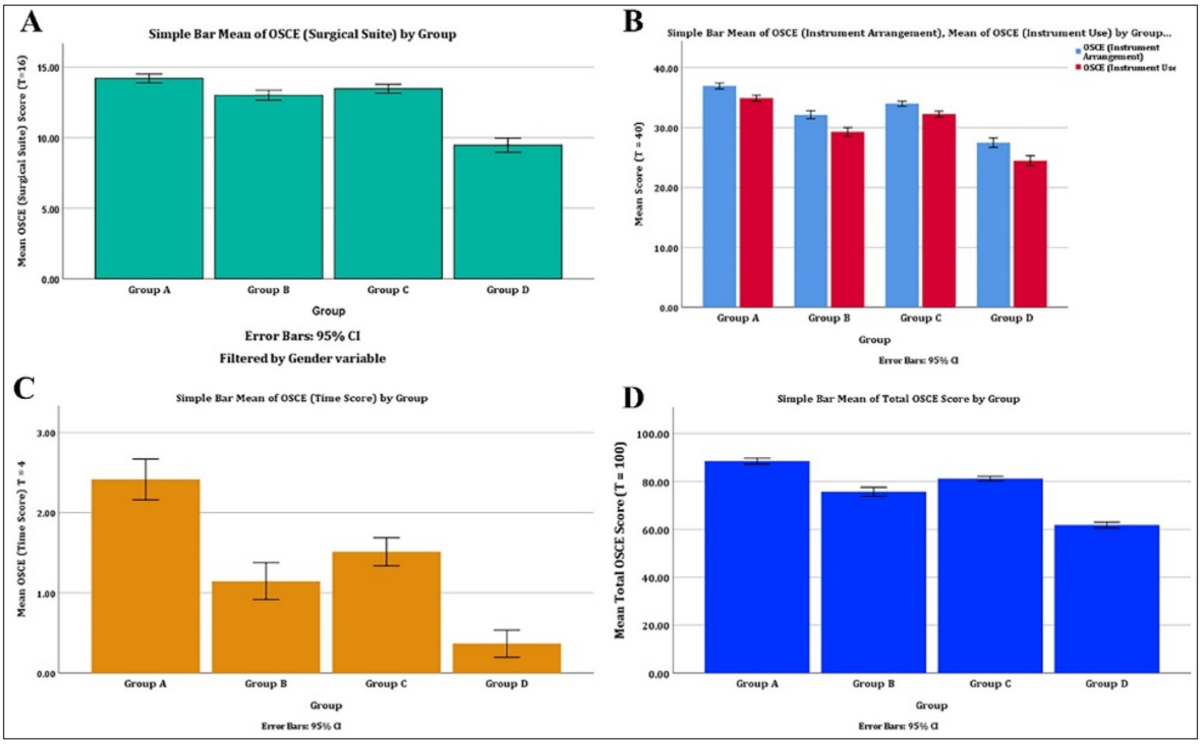

Figure 7: This figure presents the mean scores of different Objective Structured Clinical Examination (OSCE) sections by group, with error bars indicating the 95% confidence intervals (CI). It integrates four key panels: Panel A displays the surgical suite score (maximum score of 14.6), where Groups A (Augmented Reality + Gamification) and B (Augmented Reality only) perform similarly and notably better than Groups C (Gamification only) and D (Traditional instruction); Panel B compares instrument arrangement and instrument use scores (each out of 40), showing that Group A outshines the other groups, followed by Group B, with Groups C and D trailing; Panel C depicts the time score (maximum of 4), again highlighting Group A's superiority, while Groups C, B, and D follow in descending order; and Panel D aggregates these into a total OSCE score out of 100, with Group A achieving the highest overall results and Group D the lowest, thereby offering a holistic visual representation of relative strengths and performance across the OSCE assessment components.

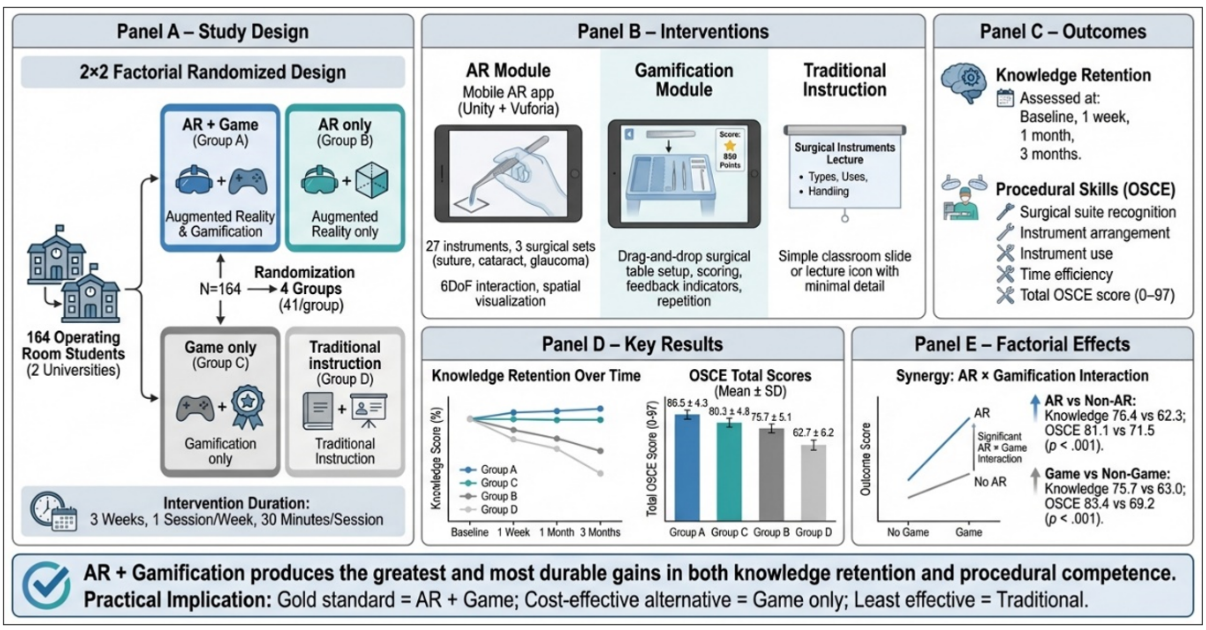

Figure 8: Augmented Reality (AR) and gamification enhance surgical instrument training: A 2×2 factorial randomized study. OSCE: Objective Structured Clinical Examination; SD: standard deviation. Group A: Augmented Reality + Gamification; Group B: Augmented Reality only; Group C: Gamification only; Group D: Traditional instruction.

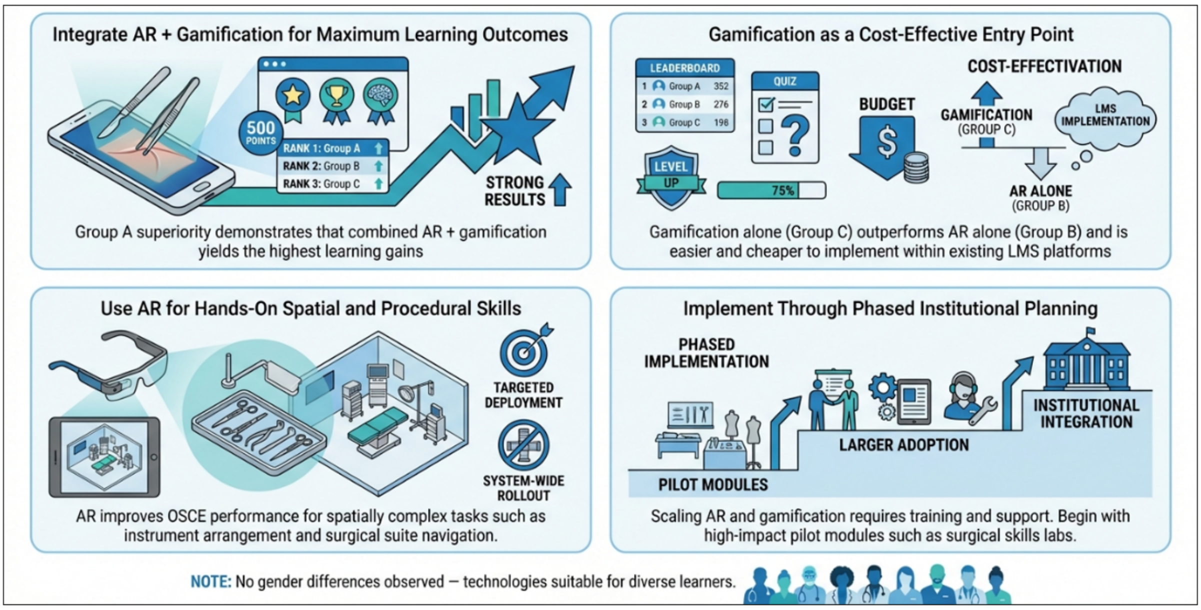

Figure 9: Curriculum design implications for Augmented Reality (AR)and gamification in surgical training.