2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
Paper ID #51770

Quantifying the Impact of Immersive VR on Learning Retention: Evidence from Flight simulator training

Ms. Madison Lynn Burlett, University of South Carolina
Dr. Mostafa Mobli, University of South Carolina

Quantifying the Impact of Immersive VR on Learning Retention: Evidence from Flight simulator training.

Abstract
Immersive virtual reality (VR) delivered through head-mounted displays (HMDs) has emerged as a promising instructional tool in engineering education, offering spatially rich, simulation-based environments for learning complex procedural and conceptual content [5]. Prior work has demonstrated mixed effects of immersive VR on learning outcomes, with comparatively limited quantitative evidence addressing its impact on knowledge retention in technically demanding domains [3,5]. This study examines whether HMD-based VR flight simulation improves both immediate learning and one-week retention relative to non-immersive instruction.

Undergraduate engineering students ($n = 69$) were randomly assigned to a VR or non-VR training condition using identical X-Plane flight simulation software. Both groups completed a standardized 15-minute training module covering aerodynamic fundamentals and flight procedures. Learning outcomes were assessed using a 37-item multiple-choice instrument administered immediately after training and again one week later. Independent-samples *t*-tests and a mixed-design ANOVA were used to evaluate group differences and retention effects over time.

Results showed that students trained with VR achieved higher mean scores on both the immediate and delayed assessments, with small but consistent effect sizes favoring the VR condition (immediate post-test: $d = 0.37$; delayed post-test: $d = 0.34$). While both groups demonstrated improved performance over time ($p < 0.01$), VR-trained students exhibited a smaller performance decline between assessments, indicating more stable retention of key aerodynamic and procedural concepts.

These findings provide quantitative evidence that immersive VR can support durable learning outcomes in flight simulation–based engineering instruction. Given its relatively short implementation time and compatibility with existing simulation infrastructure, HMD-based VR represents a scalable complement to traditional training approaches across aerospace and engineering curricula.

Introduction and Background
Immersive virtual reality (VR) delivered through head-mounted displays (HMDs) has gained increasing attention as an instructional technology due to its ability to integrate perceptual realism, spatial interaction, and embodied practice within simulated learning environments. In engineering education, this capability is particularly relevant because many foundational concepts, such as aerodynamic forces, and procedural competencies, such as control coordination and instrument scanning, are inherently spatial and action oriented. Padilla Perez and Keleş emphasize that the educational value of VR lies not in technological novelty but in its capacity to support embodied learning by allowing students to physically interact with simulated systems and processes, thereby strengthening connections between abstract theory and operational

experience [1]. Their work demonstrates significant pre–post learning gains and positions immersive VR as a means to expand access to complex training environments that would otherwise be cost- or safety-prohibitive.

Despite these promising affordances, prior research cautions that immersion alone does not guarantee improved learning outcomes. In a large-scale systematic review, report that immersive VR studies in education have produced mixed results, with learning gains highly dependent on instructional design, assessment alignment, and learner cognitive load [2]. Similarly, it was found that while immersive VR environments can increase engagement and emotional involvement, they may also introduce extraneous cognitive demands that reduce conceptual retention when instructional scaffolding is insufficient [3].

This tension between engagement and learning effectiveness has motivated more rigorous, quantitatively grounded investigations of immersive VR. One study examined immersive VR field trips across multiple studies and reported learning gains in certain contexts, while also noting that evidence for long-term retention remains limited and inconsistent across domains [4]. Complementing this work, Makransky et al. conducted a meta-analysis demonstrating that immersive VR yields small-to-moderate learning effects overall, but that these effects vary substantially depending on task complexity and instructional design [5].

Recent work in engineering and technical training provides a foundation for such investigations. Lin et al. employed a controlled experimental design comparing physical, desktop, and immersive VR assembly training, reporting superior post-test performance in the immersive condition and linking learning outcomes to specific dimensions of presence [6]. In an industrial training context, another study evaluated immersive VR for automotive assembly and found comparable learning and transfer outcomes to traditional training methods, while emphasizing VR's role as a complementary instructional tool rather than a replacement [7]. Earlier conceptual work further anticipated the educational promise of immersive interfaces, arguing that experiential and situated learning environments could fundamentally reshape how complex technical knowledge is acquired and retained [8].

Motivated by the need for quantitative evidence on durable learning outcomes in immersive technical training, the present study investigates whether head-mounted display–based virtual reality enhances knowledge retention in flight simulation–based engineering education. Building on prior work that has demonstrated mixed learning effects of immersive environments, this study isolates retention as the primary outcome measure by comparing immediate and delayed post-training assessments across VR and non-VR conditions. By holding instructional content constant and systematically examining performance decay over time, this work aims to clarify whether immersive VR provides measurable advantages for long-term conceptual and procedural learning.

Beyond performance outcomes, this investigation also considers the broader instructional implications of immersive simulation technologies. Traditional flight simulation laboratories are often constrained by cost, hardware infrastructure, and scheduling limitations, potentially restricting access to high-fidelity experiential learning environments. Standalone head-mounted display systems offer a comparatively portable and scalable alternative, suggesting potential for broader institutional adoption across programs with varying resource levels. By evaluating both the effectiveness and scalability of immersive VR, this study contributes empirical guidance for educators considering integration of immersive technologies within diverse engineering education contexts.

Methodology

The study employed a 2 × 2 mixed (between–within) factorial design, a structure commonly used in repeated-measures experimental research where one factor is manipulated between participants, and another is measured within participants over time [9]. The effect of immersive virtual reality on both immediate learning and longer-term knowledge retention in flight simulation–based engineering instruction can be examined. Instructional modality (VR versus non-VR) served as the between-subjects factor, while time of assessment (immediate post-training versus one-week delayed post-training) served as the within-subjects factor. This 2 × 2 mixed design enabled direct comparison of performance across instructional conditions while also capturing changes in learning over time within the same participants, an approach well suited to retention-focused educational research.

Participants were undergraduate engineering students enrolled in a required laboratory course. A total of $n = 79$ students initially consented to participate and were randomly assigned to either the VR training condition or the non-VR training condition. All participants received identical instructional content, differing only in the mode of delivery. Following data collection, $n = 10$ participants were excluded due to failure to attend the delayed testing session or because they fell outside the approved age range, resulting in a final analytic sample of $n = 69$ students. This attrition was anticipated and documented as part of the study's exclusion criteria.

Training sessions were conducted using X-Plane flight simulation software in both conditions to ensure equivalence of aircraft model, environment, and control response. In the immersive condition, students used a Meta Quest head-mounted display, providing a first-person virtual cockpit experience. In the non-immersive condition, the same simulation software was displayed on a standard desktop monitor, with identical control inputs. This design choice isolated immersion level as the primary independent variable while controlling for software functionality and instructional content.

Fig. 1: Experimental setup for the immersive VR flight training condition.

Each training session lasted approximately 15 minutes and followed a standardized, instructor-guided protocol. The training sequence introduced students to basic aircraft controls and instruments, followed by a structured flight exercise that included takeoff, climb, level flight, and stall recognition. Throughout the session, students were prompted to observe relationships between control inputs, aircraft attitude, airspeed, and aerodynamic forces such as lift and drag. A detailed training script was used to ensure consistency across sessions, minimizing variability in instructional delivery.

Learning outcomes were measured using a 37-item multiple-choice assessment administered immediately after training and again one week later. The assessment was intentionally designed to include a range of question types, including conceptual questions (e.g., lift, drag, angle of attack), procedural questions (e.g., takeoff and stall recovery steps), and instrument interpretation questions. Several concepts were assessed through paired or reworded questions, allowing the same underlying knowledge to be evaluated across multiple contexts. This approach reduced the likelihood that correct responses reflected memorization or test familiarity and instead emphasized conceptual consistency and transfer of understanding.

Using the identical assessment at both time points allowed retention to be isolated as change in performance over time, left unaffected by differences in test difficulty or construct coverage. Participants received no refresher instruction between assessments, ensuring that delayed performance reflected retained knowledge rather than re-exposure.

Statistical analyses included independent-samples *t*-tests to compare VR and non-VR performance at each testing interval and a mixed-design analysis of variance (ANOVA) to evaluate the interaction between instructional modality and time. Together, these analyses allowed for examination of both immediate learning effects and differential retention across instructional conditions.

Scoring and Statistical Analysis

Student performance was quantified using an answer key for the 37-item multiple-choice instrument. For student i on item q, correctness was coded as a binary indicator y_iq, where y_iq = 1 if the student's response matched the answer key and y_iq = 0 otherwise.

Total test score (percentage correct) for student i at time t was computed as:

$$S_{it} = \left(\Sigma_{q=1}^{37} y_{iqt} / 37 \right) \times 100$$

For descriptive reporting, scores were grouped into 10-point bins (e.g., 0-9.99, 10-19.99, …, 90-100). For group g and bin b, the percentage of students within each score bin was computed as:

$$P_{bg} = \left(n_{bg} / N_g \right) \times 100$$

To evaluate learning and retention, inferential analyses included independent-samples t-tests comparing VR and non-VR performance at each testing interval, paired t-tests evaluating within-group change over time, and a mixed-design analysis of variance (ANOVA) to test the Condition × Time effects. Conceptually, the mixed model can be expressed as:

$$S_{it} = \mu + \alpha_{Cond(i)} + \beta_t + (\alpha\beta)_{Cond(i),t} + u_i + \varepsilon_{it}$$

Retention was operationalized as the within-student difference between the delayed and immediate tests:

$$R_i = S_{i2} - S_{i1}$$

To quantify practical magnitude of between-group differences, Cohen's d was computed using the pooled standard deviation:

$$d = \left(\bar{X} - \bar{Y} \right) / s_p, \quad s_p = \sqrt{\left(\left((n_X-1)s_X^2 + (n_Y-1)s_Y^2\right) / (n_X + n_Y - 2) \right)}$$

Because individual item responses are binary, an item-level logistic regression was used to model the probability of a correct response as a function of VR condition, test time, their interaction, and concept category:

$$\text{logit}(\Pr(y_{iq} = 1)) = \beta_0 + \beta_1(VR) + \beta_2(Post) + \beta_3(VR\times Post) + \gamma_k(Concept)$$

Odds ratios (OR) were reported as exponentiated coefficients, with 95% confidence intervals computed from the coefficient standard errors:

$$OR_j = e^{\beta j}$$

$$CI_{95\%}(OR_j) = \left(e^{\beta j-1.96\cdot SE(\beta j)},\ e^{\beta j+1.96\cdot SE(\beta j)} \right)$$

Statistical significance was evaluated using $\alpha = 0.05$. A result was considered statistically significant when $p < 0.05$. For the mixed ANOVA and *t*-tests, p-values were derived from the corresponding test statistics under their respective sampling distributions. For logistic regression models, statistical significance was evaluated using Wald tests for individual regression coefficients.

Results and Discussion

Overall performance trends indicated that students in the immersive VR condition demonstrated stronger outcomes across both immediate and delayed assessments, particularly at the upper end of the score distribution. As shown in Fig 2. Below, median scores on the immediate post-test were identical between groups, VR participants achieved higher upper quartiles and higher maximum scores, suggesting that immersive training may disproportionately benefit higher-performing learners. On the delayed assessment, the VR group exhibited a higher median score (63.51) compared to the non-VR group (62.16), as well as a higher 75th percentile. These distributional differences indicate that VR participants not only retained more knowledge on average but also maintained stronger overall performance profiles over time.

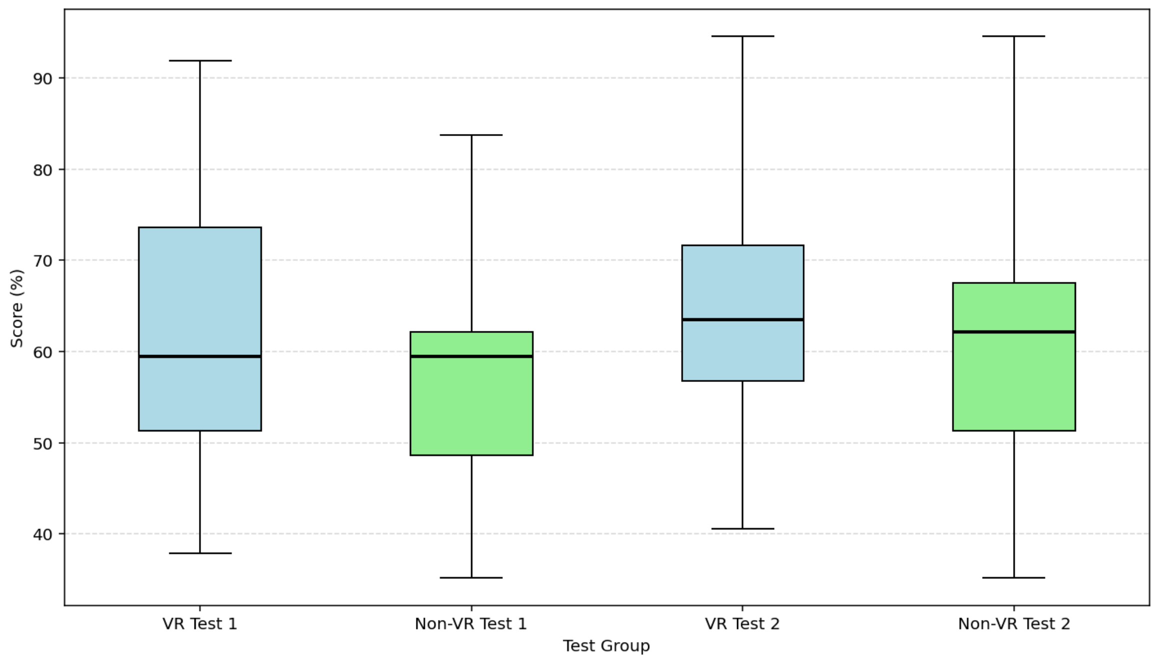

Fig. 2: Total test score distribution organized by test type.

To evaluate whether these observed improvements were statistically meaningful, a mixed-design ANOVA was conducted with instructional condition (VR vs. non-VR) as the between-subjects factor and time (immediate vs. delayed test) as the within-subjects factor. Results summarized in Table 4 revealed a statistically significant main effect of time ($p = 0.0014$), indicating that students in both conditions improved from the immediate to the one-week assessment. This suggests that learning continued for both instructional modalities. However, the main effect of condition was not statistically significant ($p = 0.1356$), nor was the interaction between condition and time ($p = 0.7134$). These findings indicate that, while VR students tended to outperform non-VR students descriptively, the rate of improvement over time did not differ significantly between groups when analyzed at the aggregate score level.

Logistic regression analysis was conducted to examine item-level performance, with odds ratios (ORs) used to characterize the direction and relative strength of instructional effects as summarized in Table 1. On the immediate post-test, students in the immersive VR condition demonstrated marginally higher odds of answering questions correctly compared to the non-VR group (OR = 1.23), indicating a consistent directional advantage associated with VR-based training, though this effect did not reach statistical significance ($p = 0.133$). Across instructional conditions, students were significantly more likely to answer questions correctly on the delayed assessment (OR = 1.14, $p = 0.011$), suggesting improved performance over time.

	OR	2.5%	97.5%	P-Value
VR	1.234	0.938	1.625	0.133
Post Test	1.136	1.029	1.254	0.011
VR:Post Test	0.979	0.853	1.123	0.7629

Table 1: Logistic regression statistical analysis results.

Although the interaction between instructional modality and time was not statistically significant (OR = 0.98, $p = 0.763$), odds ratios remained close to unity, indicating similar learning trajectories across conditions. Importantly, VR students maintained comparable or slightly higher odds of correctness across both testing intervals, supporting the interpretation that immersive training yields stable performance advantages without introducing negative learning effects. Statistical significance was evaluated using a conventional $\alpha = 0.05$ threshold.
To better understand the practical magnitude of these differences, Cohen's *d* effect sizes were calculated, as shown in Table 2. Small but consistent effects were observed for both the immediate ($d = 0.368$) and delayed ($d = 0.335$) assessments, with an overall effect size of $d = 0.353$. While classified as "small," these effects indicate a stable performance advantage for VR-trained students that persisted over time, aligning with prior work suggesting that immersive environments may yield incremental but durable learning benefits rather than dramatic gains.

	Cohen's d	Effect Size
Test 1 (VR vs. NonVR)	0.368	Small
Test 2 (VR vs. NonVR)	0.335	Small
Overall (VR vs. NonVR)	0.353	Small

Table 2: Cohen's d statistical analysis results.

Further analysis by concept category organized in Table 3, revealed that VR participants demonstrated higher probabilities of correct responses across all categories on both assessments. Although none of the category-specific VR × time interactions were statistically significant, the VR group consistently maintained higher accuracy, particularly for flight procedures, which rely heavily on spatial reasoning and embodied interaction. This pattern supports the interpretation that immersive VR may be especially well suited for procedural and action-oriented learning, even when overall retention gains are modest.

Concept Group	VR Test 1	VR Test 2	Non-VR Test 1	Non-VR Test 2
Aerodynamic Concepts	66.59%	69.71%	65.07%	65.07%
Flight Instruments and controls	58.85%	59.11%	53.60%	56.53%
Flight Procedures	61.72%	65.62%	53.15%	57.88%

Table 3: Average probability of answering correctly, organized by question type.

By examining the results for statistical method, results indicate that immersive VR-based flight training yields small but consistent quantitative advantages in learning and retention, particularly evident in score distributions and effect sizes rather than large interaction effects. While VR did not significantly alter the rate of learning over time, the persistence of higher performance levels supports its value as a pedagogical tool for simulation-based engineering education.

Conclusion and Future Studies

This study investigated the effects of head-mounted display (HMD)–based virtual reality on immediate learning and one-week knowledge retention in flight simulation–based engineering instruction. Using a mixed factorial design, the results indicate that students trained with immersive VR demonstrated consistently higher performance than their non-VR counterparts across both testing intervals. Although large interaction effects were not observed, the presence of statistically significant temporal effects and small but stable effect sizes suggests that immersive VR supports durable learning gains, particularly for procedural and spatially grounded concepts central to flight mechanics.

By focusing explicitly on retention rather than engagement or perceived realism, this work extends prior VR education research that has predominantly emphasized affective outcomes. The use of repeated, conceptually aligned assessment items allowed learning to be evaluated in terms of conceptual consistency rather than short-term recall, strengthening the interpretation that immersive environments can facilitate sustained understanding in technically complex domains. These findings align with prior evidence that embodied interaction and spatial fidelity contribute to cognitive integration and knowledge persistence in engineering learning contexts [1].

Several limitations should be acknowledged. The study was conducted at a single institution with a relatively short retention window, which may constrain generalizability. In addition, the standardized training protocol and assessment instrument, while necessary for experimental control, limit insight into how instructional design variations or alternative assessment formats might influence outcomes.

Future research should extend this work in three directions. First, longer retention intervals and multi-phase assessments would clarify the durability of immersive learning beyond one week. Second, comparative instructional design studies could examine how variations in feedback, guided discovery, or adaptive simulation elements influence retention and transfer. Third, expanding the participant pool to include students from diverse engineering disciplines, institutional types, and demographic backgrounds would provide greater insight into how immersive VR functions across varied educational contexts. Because standalone HMD systems are comparatively portable and require less physical infrastructure than traditional fixed-base simulators, immersive VR may offer scalable pathways to broaden experiential access in programs with differing resource constraints. Investigating these structural and population-level impacts represents an important next step in understanding the full educational implications of immersive simulation technologies.

The findings suggest that immersive VR, when thoughtfully implemented and rigorously evaluated, represents a viable and scalable complement to traditional simulation-based instruction in engineering education. Continued empirical work that refines instructional strategies, broadens participant representation, and evaluates implementation across varied institutional settings will be essential to fully realizing the pedagogical and accessibility potential of immersive technologies.

Bibliography:

[1] R. Padilla Perez and Ö. Keleş, "Immersive Virtual Reality Environments for Embodied Learning of Engineering Students," *Proc. ASEE Annual Conf.*, 2020.

[2] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education," *Computers & Education*, vol. 147, 2020.

[3] J. Parong and R. E. Mayer, "Learning science in immersive virtual reality," *Journal of Educational Psychology*, vol. 110, no. 6, pp. 785–797, 2018.

[4] D. M. Markowitz, R. Laha, B. P. Perone, R. D. Pea, and J. N. Bailenson, "Immersive virtual reality field trips facilitate learning about climate change," *Frontiers in Psychology*, vol. 9, 2018.

[5] G. Makransky, G. B. Petersen, and R. E. Mayer, "Immersive virtual reality and learning: A meta-analysis," *Educational Psychology Review*, vol. 31, pp. 1–26, 2019.

[6] W. Lin, L. Chen, W. Xiong, K. Ran, and A. Fan, "Measuring the sense of presence and learning efficacy in immersive virtual assembly training," *Int. J. Mechanical Engineering Education*, 2025.

[7] S. Schwarz, G. Regal, M. Kempf, and R. Schatz, "Learning success in immersive virtual reality training environments: Practical evidence from automotive assembly," *Proc. NordiCHI*, 2020.

[8] C. Dede, "Immersive interfaces for engagement and learning," *Science*, vol. 323, no. 5910, pp. 66–69, 2009.

[9] Laerd Statistics, "Mixed ANOVA using SPSS Statistics," Laerd Statistics. [Online]. Available: https://statistics.laerd.com/spss-tutorials/mixed-anova-using-spss-statistics.php.