www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

ENHANCING SAFETY-CRITICAL PROCEDURAL TRAINING FOR BATTERY ENERGY STORAGE SYSTEMS USING IMMERSIVE VIRTUAL REALITY

SUBMITTED: November 2025
PUBLISHED: April 2026
EDITOR: Robert Amor
DOI: 10.36680/j.itcon.2026.023

Xiaohui Wang, Graduate Research Assistant (corresponding author)
Dept. of Architectural Engineering, The Pennsylvania State University
https://orcid.org/0009-0005-2880-6916
xxw56@psu.edu

John I. Messner, Ph.D., Charles and Elinor Matts Professor
Dept. of Architectural Engineering, The Pennsylvania State University
https://orcid.org/0000-0002-7957-1628
jim101@psu.edu

SUMMARY: *Safety training in high-risk industries often employs immersive technologies to raise risk awareness, yet this does not always translate into evidence-based effective safe practice. Battery Energy Storage Systems (BESS) are increasingly deployed across construction and energy infrastructure, but training personnel to operate and maintain them safely remains challenging due to high risks. Immersive Virtual Reality (VR) offers a promising solution by enabling realistic, hands-on practice of hazardous procedures in a controlled environment. This study addresses whether VR's immersive engagement translates into improved procedural learning outcomes for a safety-critical BESS soft-shutdown procedure. This is done by evaluating an immersive VR training module against conventional video-based instruction in a randomized experiment with 60 engineering students. Results show that VR training significantly improved spatial understanding of equipment layout (27% higher scores) and procedural comprehension (28% higher), while VR participants completed productive tasks 45% faster on average. Participants also rated VR as more engaging with greater learning confidence, though additional instructor guidance was often needed for navigation and interaction. However, a confidence-competence mismatch was observed with perceived abilities exceeding objective performance in some domains. This motivates embedded assessment and feedback. While both training methods supported similar recall of step sequences, neither adequately conveyed the underlying safety rationale behind the procedure. These findings suggest that while VR enhances certain learning outcomes through interactive, risk-free practice, it is not a standalone solution for full procedural mastery. Practical implications include design guidelines for VR training (e.g., explicit rationale cues and embedded assessments) and recommendations for integrating VR with construction safety programs to strengthen workforce readiness for emerging technologies in the built environment.*

KEYWORDS: *virtual reality training, construction safety training, battery energy storage systems (BESS), procedural learning, cognitive learning outcomes, psychomotor performance, near transfer.*

REFERENCE: *Wang, X., & Messner, J. I. (2026). Enhancing safety-critical procedural training for battery energy storage systems using immersive virtual reality. Journal of Information Technology in Construction (ITcon), 31, 507–537. https://doi.org/10.36680/j.itcon.2026.023*

1. INTRODUCTION

Battery energy storage systems (BESS) have become critical infrastructure for modern electrical grids, buffering fluctuations in supply and demand as renewable energy adoption accelerates (Prakash et al., 2022). As deployments expand across utility-scale installations, buildings, and infrastructure projects, the need for workers who can safely operate and maintain these systems has increased (Taibi et al., 2020). Technically, BESS integrate high-energy direct current (DC) battery strings with power conversion equipment that interfaces with alternating-current (AC) loads and the grid, requiring personnel to interpret system state across subsystems and perform correct isolation actions at the appropriate devices (Idaho National Laboratory, 2024). Because incorrect de-energization sequencing can expose workers to severe electrical hazards (e.g., shock and arc flash/arc blast), safety guidance emphasizes strict shutdown and verification protocols, including industry standards for stationary storage installations and workplace electrical safety (Exeter Associates, 2022; Jeevarajan et al., 2022; National Fire Protection Association, 2026).

However, providing hands-on training for BESS maintenance is challenging. Actual energy storage systems are expensive (ranging from $122–409 per kWh of capacity (Viswanathan et al., 2022)) and often operate at lethal voltages above 600 V DC (Jeevarajan et al., 2022), making direct hands-on training impractical and unsafe. Traditional classroom instruction (diagrams, slide decks, or one-way videos) can convey declarative knowledge, but it often under-supports the procedural fluency and state-based cue recognition needed to execute a strict shutdown sequence under field constraints (Bavishi et al., 2022). One way to interpret this persistent shortfall is through Rasmussen (1997)'s risk-management perspective: when training does not adequately instantiate the constraints and feedback of real work, practice can drift from work-as-imagined toward work-as-done, increasing the likelihood of unsafe deviations. Addressing that drift is consistent with a proactive, prevention-oriented safety strategy rather than waiting for incidents to reveal training gaps (Brady & Naikar, 2022).

Immersive virtual reality (VR) offers a promising solution. By simulating BESS environments in an interactive 3D virtual space, VR enables risk-free practice where mistakes have no real consequences (Cochrane et al., 2022). VR enables unlimited practice without equipment wear and can embed context-specific feedback (e.g., indicators and meter behavior) that supports learning-by-doing (Daling & Schlittmeier, 2024). VR has the potential to enable the visualization of normally hidden system states and internal processes (e.g., electrical readings, component status), enhancing trainees' understanding of each step (Hajirasouli & Banihashemi, 2022). This enables trainees to experience rare or dangerous situations virtually and learn to respond safely.

Despite promising affordances, the evidence for construction-safety VR training is heterogeneous: Meta-analytic syntheses of construction safety VR report positive average effects overall, but they also show that effects differ across outcome categories (e.g., experience/perception vs skills/behaviors) and are moderated by study and trainee context (Scorgie et al., 2024). Broader VR safety-training reviews further note inconsistency in outcome definitions, assessment instruments, and reporting practices across domains (Man et al., 2024). As a result, "VR is engaging" does not by itself establish that VR improves the specific procedural competencies that prevent harm in high-hazard tasks, particularly where safe performance depends on correctly interpreting system-state cues and executing strict isolation sequences.

In BESS maintenance, unsafe sequencing or misinterpretation of system-state indicators can elevate exposure to electrical hazards and thermal events, and hands-on training is constrained by cost and safety risk. Against this backdrop, a clear gap emerges at the intersection of construction-safety training research and emerging energy infrastructure: while many VR safety studies emphasize presence, engagement, or generalized safety awareness, fewer evaluate curriculum-aligned, stepwise, safety-critical procedural training using objective psychomotor outcomes (accuracy, time-to-safe-state, critical-error rates) and explicit measures of system-state comprehension (Jeevarajan et al., 2022). The present study addresses this gap by experimentally comparing an immersive VR module against conventional video instruction for an eight-step BESS soft-shutdown procedure (N=60), using a measurement suite that triangulates cognitive understanding (spatial action mapping, completion-indicator and rationale comprehension), simulated execution performance, and learner experience. Cognitive learning outcomes are central to safety-critical procedural work because correct action depends on correctly identifying the right component, interpreting system-state cues, and understanding verification evidence, not merely recalling an ordered list of steps. Measuring spatial action mapping and indicator/rationale comprehension therefore provides an early, safety-relevant signal of whether trainees can execute (and appropriately adapt) a shutdown sequence under realistic constraints.

First, we present a curriculum-aligned immersive VR module for an eight-step BESS soft-shutdown procedure and a content-matched video tutorial for comparison. Second, we evaluate training effects using a multi-domain outcome set spanning cognitive learning (procedural understanding) and psychomotor learning (unguided simulated execution) alongside learner experience. Third, we translate findings into design implications for BESS-relevant safety training, including the need for explicit rationale/verification cues and embedded calibration checks to reduce confidence–competence mismatches.

2. RELATED WORK

2.1 Virtual reality in construction safety training

Virtual reality (VR) is increasingly used for occupational safety and safety-relevant training because it can simulate dangerous or difficult-to-access scenarios in controlled conditions that support repeated practice without real-world consequences (Scorgie et al., 2024; Stefan et al., 2023). From an educational standpoint, however, "training effectiveness" should be conceptualized as a multi-dimensional set of learning outcomes rather than a single score, including cognitive, skill-based (psychomotor), and affective domains (Kraiger et al., 1993). In procedural tasks, cognitive learning includes declarative and conceptual knowledge (e.g., what components are, where they are, what system states mean, and why steps must occur in a given order), whereas psychomotor learning reflects proceduralized "knowing how" (e.g., executing actions accurately and fluently under realistic constraints) (Anderson, 1982; Kraiger et al., 1993). Because durable retention and real-world transfer depend on both the generalization of learned capabilities to the job context and the maintenance of those capabilities over time, studies of VR safety training benefit from explicitly measuring both cognitive and psychomotor outcomes rather than relying solely on perceived engagement or satisfaction (Baldwin & Ford, 1988).

Construction-specific evidence confirms the general promise of VR while also showing that reported benefits depend on what investigators measure. A systematic review and meta-analysis focused on construction safety training and education found VR significantly more effective than traditional methods, reporting higher effectiveness for behavior, skill, and experience measurement categories (Man et al., 2024). This pattern underscores that construction VR studies frequently report strong gains on experience-level outcomes (e.g., perceived immersion/engagement), while educational value and safety relevance depend on whether VR also strengthens cognitive understanding and psychomotor performance that can generalize to the field (Man et al., 2024). Construction safety training reviews that compare traditional methods with computer-aided technologies likewise emphasize outcomes such as knowledge acquisition and behavior change, while noting that educational evidence quality and comparability can be uneven across studies and outcome definitions (Gao et al., 2019).

Within construction and closely related trades, VR safety training research has frequently targeted hazard recognition, risk perception, and work-at-heights contexts, with many evaluations emphasizing immediate post-training gains and learner engagement (Babalola et al., 2023; Gao et al., 2019). Notably, some foundational construction studies have included delayed tests (e.g., one-month follow-up) when comparing immersive VR with classroom safety training, demonstrating that retention can be measured in construction contexts even though it is not routinely included (Sacks et al., 2013). In parallel, immersive storytelling and 360-degree panorama approaches have been used to improve visualization and engagement for fall-hazard learning in trade contexts, highlighting the educational logic for emphasizing cognitive learning about hazards and conditions, not only experience ratings (Eiris et al., 2020). At the same time, work-at-heights reviews caution that many virtual safety training programs are not consistently designed or evaluated as multidimensional across knowledge, skills, and attitudes, and that outcome criteria are often not assessed holistically across evaluation levels, an issue that complicates interpretation when studies over-weight reaction/experience outcomes (Rey-Becerra et al., 2021).

Taken together, the construction safety VR literature establishes positive average effectiveness, but it also motivates a targeted gap for safety-critical, sequence-dependent procedures: comparatively limited evidence exists for curriculum-aligned VR training that simultaneously evaluates (i) cognitive learning outcomes needed for safe procedural reasoning (e.g., spatial localization of action points and interpretation of completion indicators) and (ii) psychomotor learning outcomes that reflect competent execution quality (accuracy, sequence integrity, and efficiency toward a verified safe end-state) (Kraiger et al., 1993; Man et al., 2024). This need for stronger educational-outcome specification is reinforced by broader VR safety-relevant training reviews showing that evaluation measures disproportionately focus on learning and reaction, with relatively few studies assessing

behavior/results-level outcomes (Stefan et al., 2023). Recent work similarly emphasizes combining subjective feedback with objective behavioral evidence (e.g., spatial tracking, heatmaps, and real-world hazard simulation frameworks) to improve evaluation rigor and comparability in virtual construction safety training (Getuli et al., 2024; Speiser & Teizer, 2024). In response, the present study evaluates VR-based BESS soft-shutdown training using both cognitive outcomes (e.g., spatial localization and comprehension of procedural indicators/logic) and psychomotor outcomes (e.g., execution accuracy, sequencing integrity, and efficiency/time-to-safe-state), while also collecting user-experience measures as adoption-relevant reaction indicators (Kraiger et al., 1993). Because transfer includes both generalization to job contexts and maintenance over time, strong evidence ultimately requires delayed retention testing and transfer-oriented assessment beyond the training medium (Baldwin & Ford, 1988; Sacks et al., 2013). These outcomes are not directly evaluated in the present study and should be addressed in future work using delayed post-tests and, where feasible, physical mock-up or field-based performance assessments.

2.2 VR in renewable energy and battery storage training

Although VR safety training has been extensively studied in construction and other high-risk sectors, its integration into renewable-energy workforce curricula remains limited, especially curriculum-aligned BESS procedural training. . Examples include a virtual environment for teaching photovoltaic system installation (Gonzalez Lopez et al., 2019) and an immersive solar design training with eye-tracking to study engagement (AlQallaf et al., 2024). Comprehensive integration of VR into standard curricula for emerging energy technologies like BESS remains rare. An industry-supported curriculum for BESS technicians, the Battery Energy Storage and Microgrid Training and Certification (ESAMTAC) program developed with support from the U.S. National Science Foundation (2015), offered an ideal foundation on which to build a VR training module. The shutdown procedure and training objectives were designed to prepare workers for safe assembly, commissioning, maintenance, repair, retrofitting, and decommissioning of energy storage and microgrid systems. Because these nationally adopted materials are already used by programs across the country to prepare electricians for the grid-interactive battery industry, they enabled us to create a VR simulation closely aligned with real-world BESS procedures and to evaluate its effectiveness against conventional instructional methods. Integrating VR into this existing safety curriculum addresses a clear gap in renewable energy education and ensures the training content is immediately relevant to industry needs.

We also drew on recent VR research in related electrical safety domains. Wang and Messner (2020) developed VR scenarios for energy storage systems and found immersive VR outperformed video in knowledge gain and engagement and focus group feedback emphasized the importance of features like immersion, ease of control, and the value of hazard scenarios and potential AR integration. Wang et al. (2024) further examined VR design features by comparing an immersive battery assembly module with equivalent hands-on exercises. The finding revealed that trainees were less efficient and accurate in VR, but targeted design improvements (e.g., better feedback and interface cues) could narrow this gap. These findings highlight the need for realism, interactivity, and guidance to achieve hands-on equivalence. Building on these lessons, our BESS module incorporated stepwise scaffolds, immediate feedback, and realistic device behavior within the ESAMTAC framework to test whether an improved, context-specific VR design can better support safety critical skills than traditional methods.

2.3 Pedagogical framework for VR learning

Our VR module design drew on established learning theories. We use multimedia learning, embodied interaction, and cognitive load mechanisms to generate outcome-specific expectations: multimedia alignment is expected to support cognitive comprehension of step logic and indicators, embodied interaction to support spatial action mapping and procedural fluency, and cognitive load to explain potential trade-offs when interface demands compete with sequence learning (Anderson, 1982; Kraiger et al., 1993). Mayer's Cognitive Theory of Multimedia Learning suggests that learning is deeper when multiple sensory channels are engaged and when words and pictures are aligned in time and space (Mayer, 2002, 2009). Traditional safety instruction often relies on 2D diagrams and verbal descriptions, which struggle to convey 3D spatial relationships or the timing of complex tasks. In contrast, an immersive 3D simulation presents information in context, following the spatial contiguity principle. For example, the simulation can align an indicator state change with the trainee's action at the moment it occurs, which should support encoding of both *what happened* (cognition) and *how to do it* (psychomotor enactment) (Mayer, 2002, 2009).

Embodied cognition theory posits that cognitive processes are deeply linked to sensorimotor experience, so physically performing a task (even virtually) reinforces understanding and memory through muscle memory and spatial awareness (Wilson, 2002). In our VR module, learners use hand controllers to mimic real actions (opening disconnect switches, turning knobs, reading meters), engaging the same motor pathways and decision processes as the actual task. This learning-by-doing approach helps integrate conceptual knowledge with practical skill, as grounding abstract instructions in concrete actions strengthens the associations needed to perform the procedure correctly in the field.

We also applied Cognitive Load Theory, which stresses balancing content complexity with limited working memory capacity (Sweller, 1994). A well-designed VR simulation minimizes extraneous load through intuitive interfaces, clear visual cues, and immediate feedback. For instance, highlighting the correct component for the next step or visualizing an action's outcome helps novices focus on relevant information rather than struggling with the interface. Conversely, if interface complexity or concurrent cues introduce high extraneous load, learners may allocate attention to navigation/control rather than to step rationale and sequencing, limiting cognitive learning even when engagement is high. We emphasized usability and provided guidance within the VR training to keep cognitive demands manageable. We also measured user perceptions of workload and difficulty as part of our evaluation to determine whether the immersive experience imposed any undue cognitive strain.

Procedural mastery requires not only 'what to do' but 'why/when to verify,' suggesting the need for rationale cues and conceptual-state explanations within VR, beyond action prompts. By grounding our VR training approach in these pedagogical frameworks, including multimedia learning, embodied cognition, and cognitive load management, we aimed to harness VR's advantages for engagement and interactivity while mitigating potential downsides. This theoretical foundation was intended to maximize learning outcomes and safety performance, ensuring that the immersive simulation not only captivates learners but also effectively builds the procedural knowledge and skills they will need in the field.

3. RESEARCH METHODS

This study employed a between-subjects experimental design to compare the effectiveness of VR training versus a non-interactive video tutorial for learning a real-world safety-critical procedure. Sixty engineering students were randomly assigned (1:1) to VR or video conditions. Immediately after training, participants completed (i) a cognitive assessment (spatial understanding, procedural knowledge, and step-sequence recall), (ii) a user-experience survey, and (iii) a hands-on performance test of the procedure in VR. Analyses focused on group differences using multivariate and univariate models (MANCOVA/ANCOVA) with prespecified covariates. The study protocol was approved by the university IRB, and all participants provided informed consent.

3.1 Task scenario and authenticity

The target task was a standardized soft shutdown of a lithium-ion battery energy storage system (BESS), an eight-step procedure that includes actions such as opening disconnects, powering down inverters, and isolating battery modules. Correct sequence is mandatory to prevent hazardous exposures (e.g., electric shock, arc flash) and to ensure controlled de-energization in line with industry practice. The procedure content and ordering were drawn from nationally adopted training materials (the ESAMTAC program) and reflected current field practice for grid-interactive BESS operations.

BESS is a system-integrated assembly that combines battery enclosures with power conversion and supervisory control/communication functions. System-level safety and certification frameworks treat ESS/BESS as integrated systems with protection, control, and communication behaviors that must operate coherently under charging/discharging and fault conditions. In operational settings, maintenance-oriented shutdown is therefore not only "turning equipment off," but a sequence of isolation and verification actions that depend on correctly interpreting system-state cues (controller status, inverter states, indicator lights, and meter behavior). This systems perspective motivated both the eight-step shutdown scenario design and the measurement strategy focusing on component localization, action mapping, and indicator-based verification (see Table 1).

The eight-step soft shutdown is framed as hazardous-energy control for maintenance: it specifies sequential actions for shutting down and isolating equipment and includes explicit verification that isolation and de-energization are effective. OSHA guidance emphasizes that energy-control procedures must include sequential steps to shut down

and isolate equipment and that employees must verify isolation/de-energization before servicing begins, including via monitoring instruments such as voltmeters where appropriate. In BESS contexts, sequence is safety-relevant because installations can include multiple interacting subsystems and energy paths. The same "off" action may not guarantee isolation unless the correct isolation points are actuated and verified through system-state evidence.

Table 1: BESS Soft Shutdown Eight-Step Procedure and Design.

Step #	Action	Component	Indicator / completion evidence (in simulation)	Safety rationale
1	Turn off inverter via site controller	Site controller HMI; inverter/PCS	HMI indicates inverter OFF; status indicator changes	Reduce load before opening disconnects; avoid opening under load (arc risk)
2	Switch off AC disconnect	AC disconnect switch/breaker	Disconnect handle state change; indicator/meter response	Isolate AC side to prevent backfeed; safer switching sequence
3	Switch off DC disconnect	DC disconnect switch	DC disconnect state change; DC meter response	Isolate DC energy path into PCS/bus; reduce shock/arc hazard
4	Open contactors (battery isolation)	Battery enclosure contactors (via control interface)	Contactor status shows OPEN/isolation	Electrically isolate batteries from system bus
5	Turn off power to door-lock mechanism	Door-lock power/interlock circuit	Lock power indicator changes; interlock status updates	Enforce safe access sequencing / interlock logic
6	Open battery enclosure door	Battery enclosure	Door state OPEN	Physical access only after de-energization/isolation stages
7	Restore power to lock mechanism	Door-lock circuit	Circuit re-energized; status updates	Return system to controlled configuration
8	Open breaker in battery enclosure	Enclosure main breaker	Breaker state OPEN	Final physical isolation / verified safe state

BESS deployments introduce safety challenges that span electrical hazards and, in some scenarios, fire and emissions concerns. Public agencies emphasize the need for strong safety planning and incident preparedness at BESS sites. At the same time, the rapid growth and diversity of ESS/BESS installations create practical constraints on hands-on training access. We therefore view the VR module as a risk-reduced environment for repeated practice of shutdown isolation and verification cues. Our performance test is intentionally framed as near-transfer within a simulated BESS architecture, and future work should evaluate delayed retention and transfer to non-VR assessments and site-specific field procedures.

The target procedure and assessment constructs were aligned with ESAMTAC, a workforce-oriented program explicitly aimed at preparing workers for safe and effective work across the lifecycle of energy storage and microgrid systems, including maintenance and decommissioning. This alignment helps ensure that the eight-step shutdown scenario reflects real training priorities in the BESS domain (safe isolation, verification evidence, and correct component interaction) rather than generic VR usability goals (See Figure1 for the step interface and indicator cues). Both conditions covered the same steps, components, and verification cues. modalities differed only in interactivity/embodiment.

3.2 Development of training modules

We developed parallel training modules for the BESS soft shutdown procedure in two formats: an interactive VR simulation and an instructional video. The VR module was created using Unity3D and deployed on an HTC Vive Pro head-mounted display, providing a first-person, immersive simulation of a battery storage system and its control interface. The virtual BESS environment was modeled to full scale based on an actual facility (a lithium-ion battery storage system previously located at the Navy Yard in Philadelphia, PA), emphasizing ecological validity and procedural realism. Trainees could move freely around the virtual equipment, exploring space without physical risk. Step-by-step guidance labels were placed next to relevant components to support wayfinding and prompt required actions. To scaffold learning, each step was introduced by a brief audio narration explaining its purpose. Learners then interacted with the equipment (operating switches, pushing buttons, reading meter displays) and observed immediate system feedback (e.g., indicator lights changing, meter values dropping as components

powered down) confirming successful completion. This design simulated an embodied interaction, leveraging spatial cues and motor practice to strengthen "muscle memory." Figure 2 outlines the VR module development process.

Figure 1: Examples of simulated soft shutdown steps and participants interacting.

The comparison condition was a video tutorial of the same procedure. The video was produced by screen-capturing a desktop-rendered version of the VR scenario, mirroring the design and content of the immersive system. An investigator executed the shutdown step by step while the session was screen-recorded, with synchronized narration explaining each action's purpose and the expected system responses (e.g., indicator lights, meter readings). A checklist of steps appeared along the left margin, and the active step was highlighted in real time as it was executed (see Figure 3 for sample screenshots of the VR and video training). The procedural content and sequence matched the VR module exactly, with the only systematic difference being the delivery medium (immersive, interactive VR vs. non-interactive video).

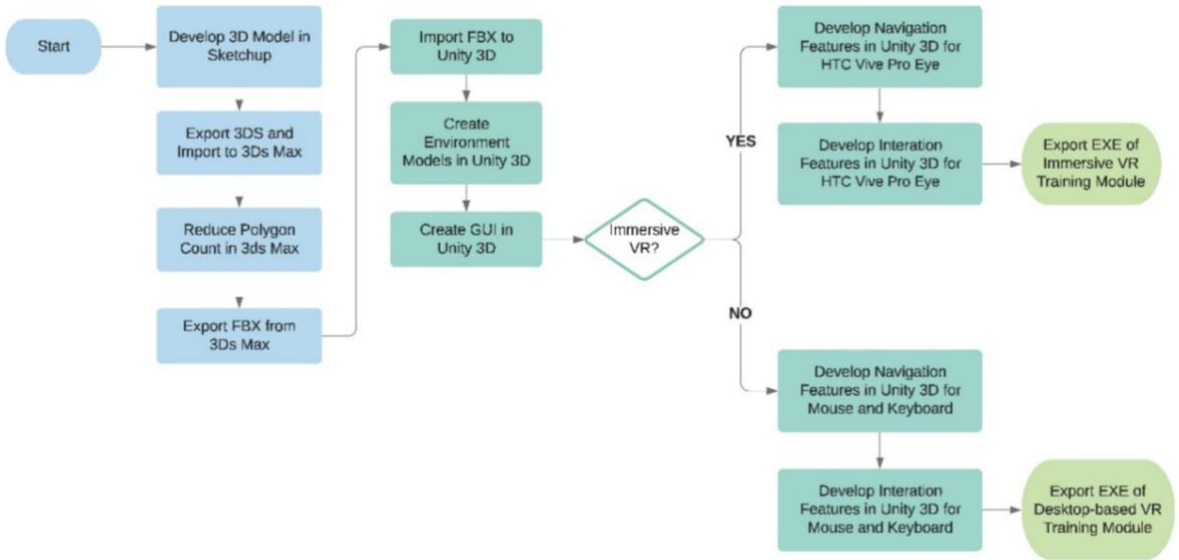

Figure 2: VR training module development process.

Figure 3: Example of VR (top) and video-based training modules (bottom).

During development, BESS subject-matter experts and safety-training instructors iteratively reviewed storyboards and early builds of both the immersive VR module and video training materials. Their feedback ensured that required knowledge, skills, and procedural logic were faithfully represented as performed in the field, and that the terminology, visuals, and instructions were clear and industry-relevant.

3.3 Hypotheses

Based on the theoretical frameworks and empirical evidence reviewed, this study examines how immersive VR training affects both learning outcomes and user experience in procedural skill acquisition. The independent variable is the training modality, and the dependent variables spanned multiple learning outcome measures and user experience metrics. We conceptualize effective learning of the BESS soft shutdown procedure as multi-component procedural competence. First, trainees must form a shutdown-relevant spatial action map (accurately localizing each required action point/interface within the equipment layout) to avoid acting on incorrect components. Second, trainees must understand procedural logic and verification cues (i.e., what system-state evidence confirms safe completion of each step and why specific precedence relations exist). Third, trainees must recall the required step order and precedence constraints. Fourth, trainees must demonstrate behavioral competence via criterion-referenced execution accuracy, including step correctness, sequence integrity, verification, and time-to-safe-state while reaching a defined "safe shutdown" state. In addition, we capture user experience and perceived learning/confidence as reaction-level outcomes and theoretically motivated mediators that influence engagement, self-regulation, and adoption of VR training, while also serving as diagnostics for usability, workload, and adverse effects that can impact performance.

Learning Outcomes Hypotheses:

- **H1a:** VR training will enhance spatial understanding of the equipment layout compared to video-based training, as measured by post-training assessments.
- **H1b:** VR training will improve procedural knowledge comprehension (understanding of steps and their logic) compared to video-based training, as measured by posttraining assessments.
- **H1c:** VR training will improve sequential retention of procedural steps compared to video-based training, as measured by posttraining assessments.
- **H1d:** VR training will enhance procedural skill execution compared to video-based training, as measured by an independent handson performance test in VR.

User Experience Hypotheses:

- **H2a:** VR training will yield more positive overall user experience ratings compared to video training, as measured by a posttraining user experience survey (e.g., higher engagement and satisfaction).
- **H2b:** VR training will lead to higher perceived learning effectiveness and confidence compared to video training, as measured by a posttraining user experience survey.

Figure 4 illustrates the mapping of each hypothesis to the corresponding evaluation components. H1d is interpreted as near transfer to an unguided simulation, not validated field transfer.

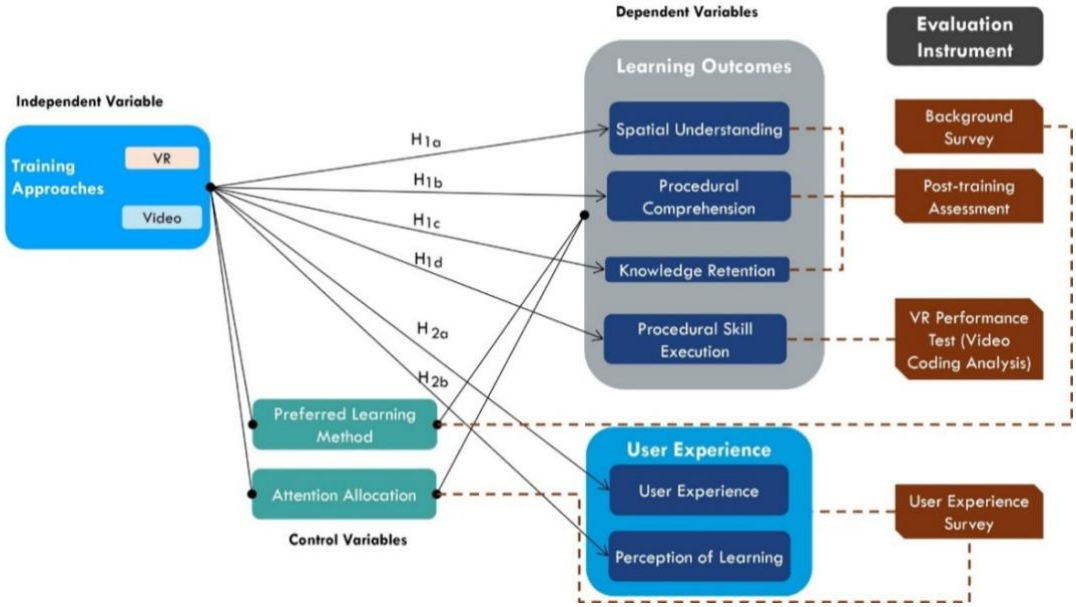

Figure 4: Overall mapping of hypotheses.

3.4 Participants

We recruited sixty undergraduate volunteers from the Department of Architectural Engineering at The Pennsylvania State University. Participants were randomly assigned to either the VR training condition (n=30) or the video-based training condition (n=30). The sample included 27 female (45%) and 33 male (55%) participants, ages 18–25, with varied familiarity with VR and gaming technology. There were no significant demographic differences between groups in terms of age, gender distribution, or prior VR experience (see Table 2 for baseline comparisons).

3.5 Experimental procedure

The evaluation study was conducted in the Immersive Construction (ICon) Lab at The Pennsylvania State University. The lab was equipped with an HTC Vive Pro Eye VR headset, high-performance computing workstations, and a three-screen projection system for video-based instruction. Each participant was assigned a

unique ID code to ensure anonymity of data. All sessions were conducted one-on-one to maximize participant focus. Figure 5 outlines the overall experimental protocol.

Prior to their lab visit, participants watched a brief introductory video about BESS technology to establish baseline knowledge of the system. This video provided context on battery storage systems and safety considerations but did *not* cover the specific shutdown procedure used in training. Upon arriving at the lab, participants provided informed consent in accordance with the Institutional Review Board protocol and viewed a short orientation briefing to familiarize them with the study setup.

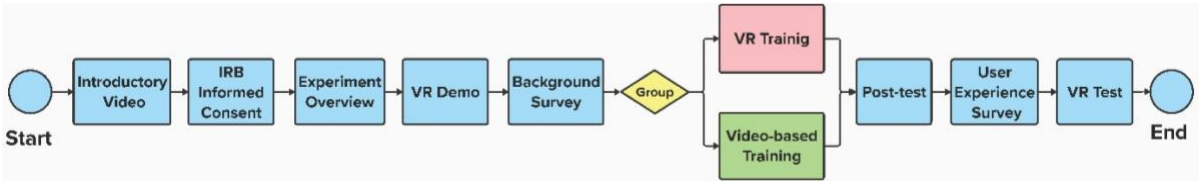

Figure 5: Experimental process.

To minimize interface-related confusion, all participants (including those in the video condition) were fitted with the VR headset and completed a 5–10 minute VR orientation. During this orientation, they learned to use the VR controls (e.g., teleport navigation, hand controller operation) and practiced until they felt comfortable with the equipment. After this VR orientation, each participant completed a background survey capturing individual difference variables (e.g., prior VR exposure, gaming experience) that might influence learning outcomes.

Participants then proceeded with their assigned training condition. Those in the VR condition completed the immersive training module at their own self-guided pace, following on-screen prompts and audio narration embedded in the simulation. The VR training took approximately 10–15 minutes (mean ≈ 14 min). An experimenter was present to answer questions or provide assistance if participants encountered difficulties. Participants in the video condition watched the instructional video on an 8'×6' projection screen in a classroom setting. An experimenter was present to monitor attention and ensure the video played without interruption. The video lasted approximately 8 minutes and could not be paused or rewound, replicating a realistic one-time tutorial scenario (see Figure 6). Although there is time difference in two treatment, the experiment compares two realistic training packages, not only two media (Clark, 1983). The duration difference is inherent nature of two training packages, but two groups were receiving instruction for one time.

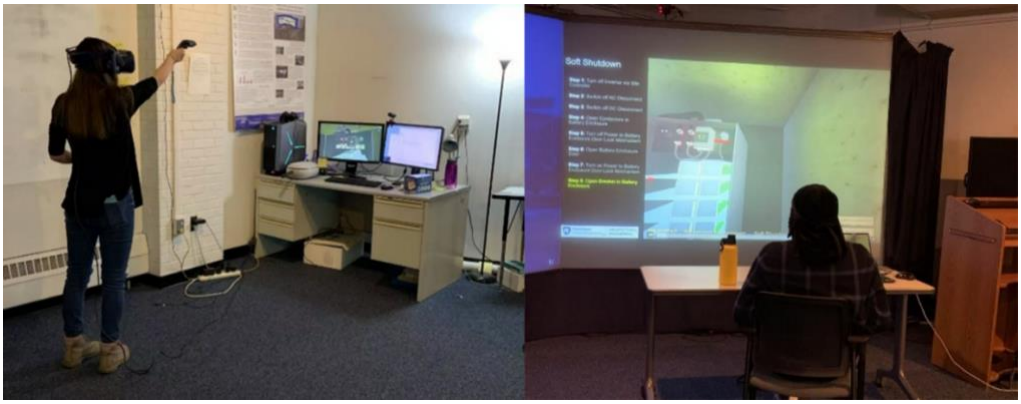

Figure 6: Sample participants during VR training (left) and video-based training (right).

Immediately after completing their assigned training, each participant completed a post-training evaluation consisting of three components administered in sequence: (1) an online assessment measuring cognitive learning outcomes (spatial understanding, procedural comprehension, and sequence retention); (2) a user experience survey capturing subjective perceptions of the training; and (3) an independent performance assessment in which the participant attempted the shutdown procedure in the VR environment without guidance. All VR training sessions and performance assessments were video recorded with IRB approval to enable detailed analysis of interaction patterns, errors, and task performance strategies.

3.6 Measures and instruments

Multiple instruments captured objective learning outcomes, subjective user experiences, and individual differences that could influence performance.

3.6.1 Cognitive learning assessment

A structured post-training assessment evaluated BESS-relevant procedural competence across three domains aligned with H1a–H1c: (i) *Spatial understanding (component localization and spatial action mapping)* measured whether participants could correctly map each procedure step to the physical equipment where it is performed (diagram-based "place step numbers next to equipment" item). (ii) *Procedural knowledge understanding (system-state and procedural-logic comprehension)* measured both completion-indicator recognition (e.g., what observable indicator shows the inverter/disconnect is off, correctness judgment about expected meter behavior after switching off the DC disconnect) and step-purpose understanding (e.g., which step achieves full system isolation, what equipment to check when an enclosure door does not open, correctness judgments about contactor and door-lock sequencing). (iii) *Step and sequence recall* measured procedural enumeration and ordering (e.g., number of steps, rank-ordering of step labels into the correct shutdown sequence) (see Appendix Table A1 for the complete assessment structure).

Primary training effects are estimated as intention-to-treat contrasts of randomized assignment. Covariates in confirmatory models are restricted to pre-intervention measures (baseline characteristics or pre-training ratings) included to improve precision and to absorb chance imbalances, not selected based on statistical significance. We avoid conditioning on post-treatment measures in primary causal models because intermediate/process variables can change the target effect and may introduce overadjustment bias. Preferred Learning Method was assessed via survey items in which participants rated their preference for visual, auditory, reading/writing, and kinesthetic learning modalities (on 1–4 scales, with higher scores indicating a stronger experiential/kinesthetic learning preference). Prior research on aptitude–treatment interactions suggests that VR and video training may differentially benefit certain learner types (McLeod et al., 1978).

Scores were computed using pre-specified rubrics developed with BESS subject-matter expert review to support content validity. Because domains differ in score ranges, descriptive reporting uses Percent-of-Maximum-Possible (POMP; 0–100), while inferential tests are invariant to linear rescaling.

Internal consistency and measurement precision. For the cognitive post-test, we do not compute one "overall alpha" because the assessment intentionally samples distinct content domains and includes mixed item formats (binary items and partial-credit tasks). Instead, we report domain-level precision where appropriate: KR-20 for the dichotomous multi-item procedural-knowledge domain and inter-item correlation with Spearman–Brown reliability for two-component domain composites. Reliability coefficients are interpreted as evidence of score precision, not as proof of construct validity, which is supported primarily by rubric specification and subject-matter-expert review.

3.6.2 User experience questionnaire

A post-training learning experience questionnaire captured subjective aspects using items adapted from the unified User eXperience (UX) evaluation framework for immersive environments (Tcha-Tokey et al., 2016), corresponding to Hypotheses H2a and H2b. The questionnaire included 20 Likert-scale items (five items on a 6-point scale of experiential ratings; fourteen items on a 5-point scale: strongly disagree to strongly agree) measuring dimensions such as presence, engagement, interface usability, perceived learning effectiveness, and cognitive load (see Appendix Table A2 for the list of items). In addition, three open-ended questions elicited qualitative feedback about the training's strengths, its limitations, and suggestions for improvement.

Primary inference uses pre-specified composite scales. Item-level analyses are relegated to Appendix Table A3 as exploratory with Holm/FDR control. Because the user-experience (UX) questionnaire is intentionally multidimensional, we report internal consistency by subscale rather than for the full set of items. For multi-item subscales, we report Cronbach's α and McDonald's ω_total as complementary indices of internal consistency (ω_total is model-based reliability). Single-item indicators are reported descriptively and do not have internal consistency coefficients.

3.6.3 Performance evaluation

After training, each participant independently executed the eight-step BESS soft shutdown unguided in VR. The performance test was administered in a common VR testbed for both groups because live BESS operation is not a safe or feasible assessment context for novice participants and because VR enables standardized, time-stamped measurement. Accordingly, performance outcomes are interpreted as near transfer to an unguided simulation (high task similarity), not as validated field transfer (Baldwin & Ford, 1988).

Video-coding quantified *procedural accuracy* (step-level correctness and sequence integrity) and *efficiency* (time spent on productive actions vs. unproductive exploration/hesitation). Two independent, condition-blinded coders applied a standardized scheme, achieving high inter-rater reliability (Cohen's $\kappa = 0.847$).

To isolate training effects, we controlled for key individual differences by including prior VR experience, video game experience, and self-rated VR navigation confidence as covariates in performance analyses. These variables were collected in the background survey (immediately after the VR demonstration) to statistically adjust for participant characteristics that might influence performance regardless of training condition.

To assess calibration between perceived and demonstrated capability, we analyzed post-training confidence ratings alongside objective performance metrics (accuracy and time-stamped execution) and recommend adding delayed retention testing in future deployments to verify durability and recalibration needs.

3.6.4 Video coding of training behaviors

VR training sessions were video recorded. Two independent investigators later coded these videos to identify interaction patterns and learning strategies. Coders were trained on a shared codebook, blinded to training condition, and resolved disagreements via consensus after independent scoring. Reliability statistics are reported for each coded dimension. We coded behaviors such as listening to instructions, reading on-screen labels, interacting with equipment, spatial exploration, hesitation, error correction, and help-seeking. This observational data provided insights into how participants navigated the learning environment and which design features facilitated or hindered skill acquisition. This observational analysis was exploratory, aiming to uncover usability issues in the VR training.

Training exposure differed by design: the VR condition provided self-paced interactive practice ($n = 30$; mean duration = 14.4 min), whereas the video condition delivered a fixed-duration tutorial (~8 min) that could not be paused or rewound. We therefore interpret this contrast as comparing two realistic training packages rather than estimating a pure "medium effect," consistent with the view that media comparisons can conflate delivery medium with instructional method and time-on-task. VR training sessions were coded as time-stamped state events using BORIS to compute (i) total training duration and (ii) time budgets for procedural practice, navigation/orientation, instruction/label processing, and assistance/scaffolding (help requests and investigator guidance). We report distributional statistics (SD, median, IQR, range) and assistance frequency/duration to distinguish procedural practice from interface overhead and scaffolding.

3.7 Data analysis

Confirmatory analyses estimated the total effect of assigned training condition (immersive VR vs. video tutorial) on immediate post-training outcomes. We interpret estimated differences as effects of the training packages as delivered (self-paced interactive VR practice vs. fixed-duration video instruction), consistent with an ecological comparison. Participants were analyzed in their assigned condition. The final N varies by outcome due to occasional missing data (handled as described below).

For the cognitive learning outcomes (H1a–H1c), we fit a multivariate analysis of covariance (MANCOVA) with training condition as the focal predictor and report Pillai's trace as the primary multivariate test statistic. When the omnibus multivariate test indicated a group effect, we conducted follow-up univariate ANCOVAs for each cognitive outcome and report adjusted means, standardized effect sizes (Hedges' g) for group contrasts, and partial η^2 where appropriate. Cognitive-outcome covariates were limited to prespecified pre-intervention learner characteristics expected to explain variance in the cognitive assessment.

For psychomotor performance (H1d), we evaluated execution outcomes in the common simulation testbed using multivariate and paired univariate covariance models as appropriate (e.g., productive time, sequencing integrity,

and performance accuracy), again treating training condition as the focal predictor and reporting adjusted means and effect sizes. Because performance was assessed in a VR simulation for both conditions, performance models additionally adjusted for prespecified pre-intervention indicators of VR interaction proficiency as precision covariates to reduce residual variance unrelated to training content, including prior VR experience, gaming experience, and VR navigation confidence measured after the standardized VR demonstration (administered identically in both conditions).

User experience was analyzed using multivariate comparisons of pre-specified UX endpoints (composites) where available. Item-level outputs (e.g., individual questionnaire items) are treated as exploratory and interpreted cautiously due to multiplicity and the modest sample size. Unless explicitly stated otherwise, p-values are two-tailed for primary inference; any directional (one-tailed) p-values are reported only as supplementary evidence for directional hypotheses.

Confirmatory models include only pre-intervention covariates collected prior to exposure to the assigned training content. Post-training self-report measures (e.g., attention allocation) and post-treatment process variables (e.g., time-on-task, assistance) were not included as covariates because they are plausibly post-treatment. Instead, they are reported descriptively and analyzed as exploratory process/mechanism indicators (Appendix Table A4).

Within each hypothesis family, we control multiplicity using Holm adjustment for confirmatory follow-up tests. Exploratory item-level outputs use false discovery rate control (q-values) and are explicitly labeled exploratory. With N = 30 per group, inference is most sensitive to moderate-to-large effects. Accordingly, we emphasize effect sizes and 95% confidence intervals alongside p-values. Assumption checks (homogeneity of covariance and variance, homogeneity of regression slopes, and outlier diagnostics) are summarized in Results Table 4. When diagnostics indicated borderline assumptions or influential observations, we corroborated primary inferences using prespecified robustness procedures (heteroskedasticity-robust HC3 standard errors for univariate models and distribution-free/nonparametric checks as appropriate), preserving the same multiplicity procedure. Participants with missing outcome data were excluded listwise for that analysis, and the final N is reported for each model.

4. RESULTS

This section presents the findings from the experimental study. We first describe participant characteristics to establish baseline comparability between groups, then examine the impact of training approach on learning outcomes (Hypothesis Set A) and learning experience (Hypothesis Set B).

4.1 Participant characteristics

Prior to examining training effectiveness, we analyzed participant characteristics collected through the background survey to verify that random assignment produced comparable groups. Table 2 summarizes demographic and experiential characteristics by condition, along with percentage point differences (pp), 95% confidence intervals (CIs), standardized mean differences (SMDs), and Fisher's exact test p-values.

Table 2: Participant characteristics and baseline balance by training condition.

Characteristic	VR	Video	Absolute difference, pp [95% CI]	SMD
Female, n (%)	10 (33.3)	17 (56.7)	-23.3 [-53.4, 12.0]	-0.48
Any prior VR experience, n (%)	22 (73.3)	21 (70.0)	+3.3 [-27.8, 33.7]	+0.07
Limited VR exposure (≤5 uses) [a], n (%)	15 (68.2)	14 (66.7)	+1.5 [-35.5, 38.3]	+0.03
Video gaming experience ≥20 hours, n (%)	20 (66.7)	17 (56.7)	+10.0 [-23.8, 41.6]	+0.21
Work experience ≥15 months, n (%)	8 (26.7)	8 (26.7)	+0.0 [-30.3, 30.3]	0.00
Prior procedural task experience, n (%)	25 (83.3)	29 (96.7)	-13.3 [-33.0, 9.3]	-0.46
VR navigation confidence ≥ moderate [b], n (%)	29 (96.7)	29 (96.7)	+0.0 [-16.1, 16.1]	0.00

Notes. Absolute difference = VR minus Video (percentage points); 95% CIs by Newcombe's score method. SMD = standardized mean difference for binary variables, $(p_{VR} - p_{Video})/\sqrt{([p_{VR}(1 - p_{VR}) + p_{Video}(1 - p_{Video})]/2)}$. [a] Among participants with any prior VR experience (VR: n = 22; Video: n = 21). [b] Assessed immediately after the VR demonstration session.

Sixty participants were randomized (VR: $n = 30$; video: $n = 30$). The VR group comprised 10 female and 20 male participants, while the video group had 17 female and 13 male participants. Prior VR experience was comparable between groups: 22 VR participants (73.3%) and 21 video participants (70.0%) had used VR before (SMD = 0.07, p = 1.00), though most had limited exposure. 15 VR participants (68% of those with experience) and 14 video participants (67%) reported five or fewer uses (SMD = 0.03, p = 1.00). Among the 43 participants with prior VR experience, the Meta Quest was the most commonly used headset (n = 14), while only one participant had previously used an HTC Vive.

Video gaming experience showed similar patterns across conditions, with 20 VR participants (66.7%) and 17 video participants (56.7%) reporting more than 20 hours of gaming experience (SMD = 0.21, p = 0.60). Work experience distributions were nearly identical, with 8 participants in each group having more than 15 months of work experience (SMD = 0.00, p = 1.00). Nearly all participants reported prior experience with procedural tasks: 25 in the VR group (83.3%) and 29 in the video group (96.7%), representing a modest imbalance (SMD = 0.46, p = 0.20). Following the VR demonstration session, 29 participants in each group (96.7%) reported moderate or higher confidence in their ability to navigate and interact with VR training scenarios (SMD = 0.00, p = 1.00). This high and comparable confidence across conditions suggests that any observed performance differences would likely reflect training effectiveness rather than differential comfort with VR technology.

Overall, random assignment achieved adequate baseline balance across most characteristics. The moderate gender imbalance (SMD = 0.48, p = 0.12) and slight difference in prior procedural task experience (SMD = 0.46) are unlikely to substantially bias treatment effect estimates given equal group sizes. Sensitivity analyses adding both as covariates produced minimal change in the primary VR–Video estimates (≤17% change) and did not alter inference. No evidence of Training × Gender or Training × Prior Experience moderation was detected (see Appendix Table A5). These results indicate that the primary findings are robust to baseline composition differences and are unlikely to be driven by demographic imbalance.

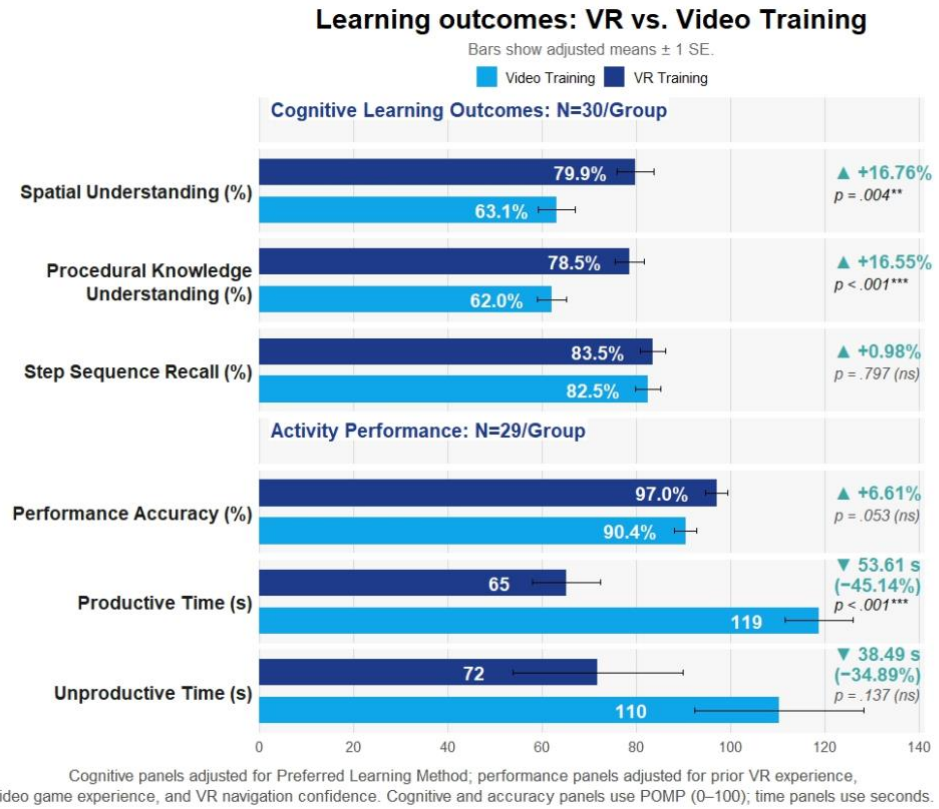

Figure 7: Results of post training assessment.

4.2 Hypothesis set A: Impact of training approaches on learning outcomes

Having established baseline comparability between groups, we examined whether the training approach influenced learning outcomes across cognitive and psychomotor domains (see Figure 7).

4.2.1 Cognitive learning outcomes

A multivariate analysis of covariance (MANCOVA) tested whether training condition (VR versus video) differentially affected cognitive outcomes: Spatial Understanding (H1a), Procedural Knowledge Understanding (H1b), and Step & Sequence Recall (H1c). The outcomes were moderately intercorrelated (see Table 3), justifying a multivariate approach. Preferred Learning Method was included as a baseline covariate in the primary confirmatory model. Because these three outcomes were scored on different ranges (9, 6, and 17 points, respectively), scores were linearly rescaled to a common Percent-of-Maximum-Possible (POMP) metric on a 0–100 scale for descriptive summaries and figures. Inferential results (e.g., F, p, partial η^2) are invariant to this linear transformation. We report both raw units and POMP values for interpretability.

Table 3: Descriptive statistics and intercorrelations.

A. Learning Outcomes by Training Approach

Measure	VR (n = 30)	Video (n = 30)	Difference	Effect Size (g)	p
Spatial Understanding					
Raw Score (0-9)	7.20 (1.67)	5.67 (2.12)	1.53	0.79	.003
POMP (%)	80.00	62.96	17.04		
Procedural Knowledge Understanding					
Raw Score (0-6)	4.73 (0.98)	3.70 (1.06)	1.03	1.00	< .001
POMP (%)	78.89	61.67	17.22		
Step & Sequence Recall					
Raw Score (0-17)	14.17 (2.42)	14.07 (2.48)	0.10	0.04	.782
POMP (%)	83.33	82.75	0.58		

B. Intercorrelations Among Dependent Variables (Pooled Within-Group)

Variable	Spatial	Procedural	Steps
Spatial Understanding	1.000	0.457	0.555
Procedural Knowledge Understanding	0.457	1.000	0.331
Step & Sequence Recall	0.555	0.331	1.000

Note. Panel A values are M (SD); effect sizes are Hedges' g. Panel B entries are pooled within-group Pearson correlations. Score ranges: Spatial Understanding 0-9, Procedural Knowledge 0-6, Step & Sequence Recall 0-17.

Domain-level internal consistency for the cognitive post-test was modest, consistent with a short, mixed-format, content-sampled assessment: Step & Sequence Recall (inter-item r = .291; Spearman–Brown = .451), Spatial Understanding (r = .153; Spearman–Brown = .266), and Procedural Knowledge Understanding (KR-20 = .459). These coefficients provide limited evidence of score precision and support cautious interpretation of the domain scores, but they do not themselves establish construct validity.

Assumption checks supported the planned cognitive-outcome MANCOVA (see Table 4). Box's M test was non-significant ($\chi^2(6)=6.53$, p=.366), indicating no gross covariance-inequality. One Levene test was significant for Spatial Understanding ($F(1,58)=5.29$, p=.025), suggesting modest variance heterogeneity. All other Levene's were non-significant. Multivariate normality was violated (all Shapiro–Wilk p<.05), so Pillai's Trace was used because it is robust to mild nonnormality and unequal covariance structures. In addition, homogeneity-of-slopes checks for the interaction between training condition and Preferred Learning Method were non-significant for all three outcomes, supporting the ANCOVA specification. Inference was supported by nonparametric checks: a trimmed-sample MANOVA and individual Mann-Whitney tests yielded the same pattern (no change in significance).

The omnibus MANCOVA indicated that training condition affected the combined cognitive outcome set (see Table 5). Follow-up ANCOVAs showed statistically reliable group differences for Spatial Understanding and Procedural Knowledge Understanding, whereas Step & Sequence Recall did not differ between conditions (see Table 5). In terms of magnitude, VR trainees scored higher on Spatial Understanding (80.0 vs. 63.0 POMP; +17.04 pp; Hedges' g = 0.79) and Procedural Knowledge (78.3 vs. 61.7 POMP; +16.7 pp; g ≈ 0.7), while Step & Sequence Recall was similar across groups (g ≈ 0.05). Overall, H1a and H1b were supported but H1c is not supported.

In BESS shutdown, “spatial understanding” reflects whether trainees can correctly localize isolation points and execute actions on the correct BESS components, while “procedural knowledge understanding” reflects whether they can interpret system-state evidence (controller readouts, status indicators, meter behavior) that verifies de-energization and effective isolation.

Table 4: Assumption checks and robustness tests for outcome-family MANCOVAs.

Panel A. Cognitive learning outcomes (N = 60)

Test	Statistic	df	p	Interpretation
Box's M	$\chi^2 = 6.53$	6	.366	Covariance assumption met
Levene's test				
Spatial Understanding	F = 5.290	1, 58	.025	Mild variance heterogeneity
Procedural Knowledge	F = 0.780	1, 58	.380	Assumption met
Step & Sequence Recall	F = 0.060	1, 58	.815	Assumption met
Shapiro-Wilk				
All outcomes by group	-	-	< .05	Non-normal in all cells
Primary omnibus test	V = 0.273, F = 6.760	3, 54	< .001	Significant Pillai's Trace
Trimmed-sample omnibus	V = 0.340	3, 52	< .001	Stable to trimming
Mann-Whitney U				
Spatial Understanding	U = 259.0	-	.004	Convergent
Procedural Knowledge	U = 215.5	-	< .001	Convergent
Step & Sequence Recall	U = 447.0	-	.970	Convergent

Panel B. Psychomotor learning outcomes (N = 58)

Test	Statistic	df	p	Interpretation
Box's M	$\chi^2 = 6.999$	6	.999	Covariance assumption met
Levene's test				
Performance Accuracy	F = 3.578	1, 56	.067	Assumption met
Productive Time	F = 5.292	1, 56	.026	Mild variance heterogeneity
Unproductive Time	F = 0.938	1, 56	.338	Assumption met
Shapiro-Wilk				
Accuracy and time outcomes	-	-	Mostly < .05	Mostly non-normal
Primary omnibus test	V = 0.379, F = 10.160	3, 50	< .001	Significant Pillai's Trace
Trimmed-sample omnibus	V = 0.385	3, 49	< .001	Stable to trimming
Permutation ANCOVA (Accuracy)	F = 3.90	1, 53	.046	Raw permutation p significant
Holm-adjusted permutation	-	-	.092	Marginal after correction
Mann-Whitney U				
Performance Accuracy	U = 346.5	-	.163	No unadjusted group difference
Productive Time	U = 138.5	-	< .001	Convergent
Unproductive Time	U = 295.0	-	.052	Marginal

Note. Pillai's Trace was used as the primary omnibus statistic because it is more robust than alternative MANOVA criteria under mild non-normality and variance heterogeneity, particularly with balanced groups. Thus, the main pattern of results appears robust, although the covariate-adjusted psychomotor accuracy effect should be interpreted more cautiously because the permutation result was not significant after Holm correction.

Post-hoc distributional diagnostics indicated a ceiling effect for step–sequence recall (see Table 6). Overall, 13/60 participants (21.7%) achieved the maximum score (17/17), and 18/60 (30.0%) scored ≥16/17. Max-score rates exceeded the commonly used >15% ceiling-effect criterion in both groups (VR: 26.7%; Video: 16.7%), indicating restricted variance at the upper bound (VR median [IQR] = 14.5 [13–17]; Video = 15 [13–16]). Accordingly, the null between-group difference on step–sequence recall should be interpreted as measurement-limited (limited

discriminative bandwidth for high performers) rather than as evidence that the two approaches are necessarily equivalent on higher-order procedural competence. Consistent with the band distribution, the total-score distributions were highly similar across groups (Mann–Whitney U = 453, p = .970; rank-biserial = 0.007), suggesting that any modality effect would be difficult to detect given the ceiling restriction.

Table 5: MANCOVA Results: Effects of Training Approaches on Learning Outcomes.

A. Multivariate Test

Test	Value	F	df	p	Partial η²	Power
Pillai's Trace	0.264	6.558	3, 55	.001	0.264	-

B. Univariate ANCOVAs (Adjusted Means ± SE; each cell shows Raw on top and POMP 1-100 on bottom)

Outcome	VR	Video	Diff. [95% CI]	F (1, 57)	p	Holm-p	Partial η²	Hedges' g
Spatial Understanding	7.19 ± 0.35 79.86 ± 3.93	5.68 ± 0.35 63.10 ± 3.93	1.51 [0.50, 2.51] 16.76 [5.58, 27.94]	9.015	.004	.008	0.137	0.79
Procedural Knowledge Understanding	4.71 ± 0.19 78.55 ± 3.12	3.72 ± 0.19 62.00 ± 3.12	0.99 [0.46, 1.53] 16.55 [7.68, 25.42]	13.951	< .001	.001	0.197	1.00
Step & Sequence Recall	14.20 ± 0.45 83.53 ± 2.66	14.03 ± 0.45 82.55 ± 2.66	0.17 [-1.12, 1.45] 0.98 [-6.59, 8.55]	0.067	.797	.797	0.001	0.04

Notes. In Panel B, each cell shows Raw units (top) and POMP 1-100 (bottom; percent sign omitted). Covariate: Preferred Learning Method ($\bar{X}$ = 7.90). Adjusted means are predicted at the covariate mean.

Table 6: Step-sequence recall score distribution and ceiling-effect diagnostics (Total score out of 17).

Score Band (raw /17)	POMP (%)	VR (n = 30)	Video (n = 30)	Total (N = 60)
17/17 (maximum)	100.0	8 (26.7%)	5 (16.7%)	13 (21.7%)
16/17 (near-maximum)	94.1	1 (3.3%)	4 (13.3%)	5 (8.3%)
15/17	88.2	6 (20.0%)	8 (26.7%)	14 (23.3%)
14/17 (≥ 80% band)	82.4	3 (10.0%)	3 (10.0%)	6 (10.0%)
12-13/17	70.6-76.5	8 (26.7%)	5 (16.7%)	13 (21.7%)
9-11/17	52.9-64.7	4 (13.3%)	5 (16.7%)	9 (15.0%)
≤ 8/17	≤47.1	0 (0.0%)	0 (0.0%)	0 (0.0%)

Note. Total score = Q1 (0-1) + Q2 (0-16), max = 17. POMP expresses percent of maximum possible score. A ceiling effect is commonly flagged when > 15% of respondents obtain the maximum possible score on a bounded scale; in this sample, max-score rates were 26.7% (VR) and 16.7% (Video), indicating restricted variance at the upper bound.

Exploratory individual-difference analyses examined whether preferred learning method (measured in the pre-training background survey) and attention allocation (retrospective self-report collected post-training) were associated with learning outcomes (see Appendix Table A6). In multivariate tests, neither preferred learning method (Pillai = 0.042, $F(3,54) = 0.780$, $p = .510$) nor attention allocation (Pillai = 0.084, $F(3,54) = 1.644$, $p = .190$) showed an overall association with the cognitive outcome set. At the univariate level, higher reported attention allocation toward spatial–visual information was modestly associated with spatial understanding ($\beta = .27$, $b = 0.47$, SE = 0.21, 95% CI [0.05, 0.89], $p = .028$), whereas associations with procedural knowledge understanding and step/sequence recall were small and non-significant (all $p \geq .255$). Preferred learning method showed no discernible relationship with any learning outcome (all $p \geq .318$). Because attention allocation was measured post-training and may itself be influenced by training condition, it is interpreted as an exploratory process/engagement indicator rather than a baseline covariate in confirmatory models. In addition, the learning-preference item is a brief self-report measure rather than a validated psychometric instrument. Accordingly, these results are interpreted descriptively and used primarily to motivate future work with validated strategy/attention measures and prospective designs.

To verify that our conclusions were not driven by any outlier observations, we repeated the MANCOVA after excluding three participants identified as multivariate outliers. The exclusion of outliers slightly increased effect

sizes but did not alter the pattern of significance, confirming that our main findings are not driven by extreme scores.

4.2.2 Psychomotor learning outcomes

Hypothesis H1d predicted that VR training would enhance procedural skill execution. We tested this via MANCOVA on performance metrics (productive task time, unproductive time, and accuracy) with prior VR experience, video game experience, and VR navigation confidence as covariates (see Table 7). Two participants were excluded due to missing data, resulting in N = 58 (VR: n = 29; video: n = 29).

Table 7: Descriptive Statistics and Intercorrelations (Performance Outcomes).

A. Performance Outcomes by Training Approach

Measure	VR (N=29)	Video (N=29)	Difference	Effect Size (g)	p
Performance Accuracy					
Raw Score (0-14)	13.55 (0.95)	12.69 (2.30)	0.86	0.48	.053
POMP (%)	96.80	90.64	6.16		
Productive Time					
Seconds	65.24 (28.66)	118.66 (45.81)	-53.41	-1.38	< .001
Unproductive Time					
Seconds	74.10 (68.61)	108.07 (117.17)	-33.97	-0.35	.137

B. Intercorrelations Among Performance Outcomes (Pooled Within-Group)

Variable	Performance Accuracy	Productive Time	Unproductive Time
Performance Accuracy	1.000	-0.289	-0.151
Productive Time	-0.289	1.000	0.589
Unproductive Time	-0.151	0.589	1.000

Notes. Panel A shows M (SD). Differences are VR - Video; negative differences on time variables indicate faster performance in VR.

Hedges' g uses pooled SD (sign follows VR - Video).

For performance outcomes, diagnostic tests supported MANCOVA/ANCOVA inference (see Table 4). Box's M was non-significant ($\chi^2(6)=6.999$, p=.999). Levene tests were mostly non-significant, except Productive Time ($F(1,56)=5.29$, p=.026). Shapiro–Wilk indicated non-normality for accuracy and time measures (p<.05 in most cases). Pillai's Trace (chosen for robustness) remained highly significant. A permutation ANCOVA (Freedman–Lane method) for accuracy was marginal (p=.046 raw, .092 after Holm), consistent with the MANOVA trend. Overall, assumption violations appeared mild, and nonparametric sensitivity analyses confirmed the main findings (see Table 4).

The omnibus performance MANCOVA indicated a significant multivariate effect of training condition on the combined performance outcomes (see Table 8), supporting H1d. In follow-up ANCOVAs, the training effect was statistically significant for Productive Time ($p < .001$), but not for Unproductive Time ($p = .137$). The group effect on Accuracy was marginal ($p = .053$) and did not reach conventional significance. In terms of magnitude, VR trainees completed productive actions substantially faster (approximately 45% less time on average), while the estimated accuracy advantage was modest (mean difference ≈ 6.6 percentage points). Taken together, these results provide partial support for H1d: VR primarily improved efficiency (productive execution speed) rather than reliably improving accuracy under the current assessment sensitivity. Error analysis revealed both groups missed the same subtle procedural details, suggesting a common gap in understanding rather than a training medium effect. Productive vs. unproductive time can be interpreted as a proxy for efficiency in locating components and completing isolation actions under an unguided situation, an operationally relevant demand in BESS installations where multiple subsystems and interfaces must be navigated correctly.

We conducted an exploratory sensitivity analysis for performance accuracy by adding total performance time (productive + unproductive seconds) as a covariate. Because total time is a post-treatment process variable rather than a baseline covariate, this model is descriptive and does not estimate the same intent-to-treat effect as the

primary analysis. Under this specification, the VR accuracy advantage attenuated from 6.6 to 5.0 percentage points and was no longer statistically distinguishable (p = .171), while total time itself was not significant (p = .204) (see Appendix Table A7).

Table 8: MANCOVA Results: Effects of Training Approach on VR Performance Outcomes.

A. Multivariate Tests

Test	Value	F	df	p	Partial η²	Power
Pillai's Trace	0.379	10.38	3, 51	< .001	0.379	0.999
Wilks' Λ	0.621	10.38	3, 51	< .001	0.379 [a]	

B. Univariate ANCOVAs (Adjusted Means ± SE; VR-Video Difference with 95% CI)

Outcome	VR	Video	Diff. [95% CI]	F (1, 53)	p	Holm-p	Partial η²
Performance Accuracy	97.03 ± 2.36	90.41 ± 2.36	6.61 [0.09, 13.14]	3.90	.053	.107	.069
Productive Time (s)	65.15 ± 7.14	118.75 ± 7.14	-53.61 [-73.41, -33.81]	27.88	< .001	< .001	.345
Unproductive Time (s)	71.84 ± 17.95	110.33 ± 17.95	-38.49 [-88.24, 11.26]	2.28	.137	.137	.041

Notes. Differences are VR minus Video; negative values on time outcomes indicate faster VR performance. Holm-p controls familywise error across the three univariate ANCOVAs. Performance Accuracy scale is displayed as POMP on a 0-100 scale.

In addition, none of the covariates significantly influenced outcomes (all p >.298) (see Table A8). Mahalanobis distance identified two multivariate outliers ($D^2 > 16.27$). A MANCOVA test with trimmed data and sensitivity analyses comparing results with (N = 58) and without (N = 56) these outliers revealed consistent patterns, supporting the robustness of our findings.

4.3 Hypothesis set B: Impact of training approaches on learning experience

Participants' subjective evaluations shed light on why the learning outcomes were as observed. The learning experience analysis leveraged both quantitative survey data and qualitative observations. Post-training user experience surveys captured participants' perceptions across multiple dimensions, while video coding of VR training sessions provided behavioral learning patterns. For hypothesis set B, no covariates were used in the MANOVA as groups did not differ in any pre-training rating (and personal factors like gender or experience did not significantly influence these subjective ratings).

Table 9: User experience subscale outcomes (POMP 0-100): omnibus MANOVA and follow-up subscale contrasts (two-tailed).

Panel A. Omnibus multivariate test (VR vs. Video)

Endpoint set	Multivariate statistic	Value	F (df_{hyp}, df_{err}) and p
Primary UX endpoints (Judgment, Presence/Immersion, Engagement, Perception of Learning)	Pillai's trace	0.300	5.897 (4, 55) p = .0005

Panel B. Subscales (POMP 0-100), reliability, and contrasts

UX domain	k	α	ω_total	VR mean	Video mean	p (two-tailed)	Holm-p (4 primary)	Interpretation
Judgment	5	0.867	0.873	85.5	67.1	5.73×10^{-5}	2.29×10^{-4}	Higher perceived judgment and overall UX in VR.
Presence/ Immersion	2	0.454	0.456	70.8	50.0	1.81×10^{-4}	5.43×10^{-4}	Higher presence in VR; brief, low-precision indicator.
Engagement	3	0.623	0.679	71.7	57.8	.00613	.00787	Higher engagement in VR.
Perception of Learning	4	0.882	0.886	78.5	63.8	.00393	.00787	Higher perceived learning and confidence in VR.

Note. POMP scores rescale items to 0-100 using $((raw - min) / (max - min)) \times 100$.

4.3.1 User experience survey—quantitative analysis

Hypotheses H2a–H2b predicted that VR training would yield a more positive user experience and higher perceived learning/confidence than video training. The post-training UX survey included 19 items common to both groups (Q5-1–Q5-5; Q8-1–Q8-14) plus one VR-only item (Q8-15). Item wording and descriptive statistics appear in Appendix Table A2 with item-level inferential results in Table A3. To address concerns about high-dimensional item-level modeling, we aggregated items into pre-defined UX subscales: Judgment, Presence/Immersion, Engagement, and Perception of Learning as the primary UX endpoints, with Technology Adoption, Emotion, Cognitive Load (Q8-10; reverse-coded so higher indicates better), and Usability treated as secondary. All UX scores are reported as percent of maximum possible (POMP; 0–100) to place items with different response ranges on a common metric.

Given N=30 per condition, statistical sensitivity is greatest for moderate-to-large effects. Therefore, we emphasize adjusted mean differences with 95% confidence intervals and standardized effect sizes rather than relying on p-values alone. Smaller effects cannot be ruled out and should be revisited in larger samples or multi-site replications.

We evaluated internal consistency for each composite using both Cronbach's α and McDonald's ω. Internal consistency was high for Judgment (α = .867, ω_total = .873) and Perception of Learning (α = .882, ω_total = .886), moderate for Engagement (α = .623, ω_total = .679), and acceptable for Technology Adoption (α = .823, ω_total = .831). Presence/Immersion showed low internal consistency (α = .454, ω_total = .456) and is therefore interpreted as a brief indicator rather than a high-precision composite (see Table 9). This dual reporting is appropriate because α can be sensitive to restrictive assumptions, while ω is a practical alternative for many common measurement structures.

Primary multivariate inference used a one-way MANOVA on the four prespecified composites (p = 4). We report Pillai's trace as the primary multivariate criterion because it is commonly recommended as relatively robust under assumption strain. A one-way MANOVA across the four primary UX endpoints showed a reliable overall difference between training conditions (Pillai's trace = 0.300, $F(4,55) = 5.897$, p = .0005), indicating systematically more positive UX ratings under VR (see Table 9). Pillai's trace is reported because it is commonly recommended as more robust than Wilks' λ under assumption strains. We report two-tailed p-values as the primary inferential quantities in the main text. One-tailed p-values aligned with preregistered directional expectations (VR > Video) are retained only as supplementary columns in Appendix and interpreted cautiously to maintain logical consistency between p-values and directional claims. Because UX involves multiple related outcomes, we control familywise error within the UX item family using Holm-adjusted p-values.

We then conducted follow-up adjusted univariate models for each primary composite to estimate adjusted mean differences, 95% confidence intervals, and standardized effect sizes (Hedges g), applying Holm control across the four primary composites. Follow-up subscale contrasts supported H2a–H2b: relative to video, VR trainees reported higher Judgment (85.5 vs 67.1 POMP), higher Presence/Immersion (70.8 vs 50.0), higher Engagement (71.7 vs 57.8), and higher Perception of Learning (78.5 vs 63.8). All remained significant after Holm correction across the four primary subscales (max Holm-p = 0.0079; see Table 9). Secondary domains suggested stronger adoption intent and affective engagement in VR, while usability did not significantly differ. All remaining survey domains and item-level tests are explicitly labeled exploratory and are reported in Appendix with multiplicity-adjusted p-values (Holm) and effect sizes. These analyses are presented to support transparency and hypothesis generation rather than confirmatory claims.

Finally, the VR group's higher perceived learning/confidence should be interpreted alongside objective outcomes: subjective confidence exceeded objective advantages in some domains (e.g., recall/sequence showed ceiling clustering), suggesting a potential confidence–competence gap that has implications for assessment design (e.g., adding rationale-based checks and discriminating scenario variants in future evaluations).

4.3.2 User experience survey—qualitative feedback

Open-ended survey responses provided additional insight into each training method's strengths and weaknesses. A qualitative analysis of participant comments (summarized in Table 10) revealed recurring themes. Nine participants (15 %), including 4 VR and 5 video participants, explicitly noted that neither training method adequately explained the rationale behind the sequence of steps. For example, one participant wrote, "I know we had to turn off the inverter via site controller before turning off AC disconnect in equipment, but I'm not entirely

sure what risk there is if I did not." A VR trainee similarly commented, "The VR was great for practicing how to do it, but I still don't quite understand what was really changed in the system." This feedback underscores that both approaches lacked explicit instruction on the underlying safety logic ("the why"), an issue we address in the discussion.

Table 10: Summary of qualitative feedback by training condition.

	Virtual Reality Training		**Video-based Training**	
	Items	**Times**	**Items**	**Times**
Positive Feedback	Hands-on and safe learning	14	Linear structure	15
	Ease of use & intuitiveness	6	Multimodal instructions	14
	Instructional support	6	Step-by-step guidance	11
	Immersion & realism	5	Realistic representation	4
	Self-paced learning	3	Replay ability	1
Negative Feedback	Technical issues	9	Fast pace	7
	Retention & comprehension	7	Rapid transition	6
	Disorientation & physical discomfort	5	Technical & usability issues	5
	Instructional limitations	5	Lack of engagement	3
	Learning curve	4	Memorization difficulty	2
Suggested Improvements	Clearer instructions	5	Slower pace & repetition	7
	No change suggested / satisfied	6	No change suggested / satisfied	5
	Guidance / instructor for "why"	4	Explain safety rationale &principles	5
	Comfort & usability	3	Interactive elements	4
	Realism & interaction	3	Improved visual aids	3
	Technical enhancements	3	Orientation tools	2
	Repetition & practice	2	Summaries & recaps	2

Note. Frequency counts represent the number of participants who mentioned each theme in open-ended responses (N = 60; VR: n = 30, Video: n = 30). Participants could mention multiple themes. Themes were identified through systematic thematic analysis.

VR trainees commonly praised hands-on practice and interactivity. They described the VR as "engaging," "realistic," and helpful for remembering the equipment layout. Several noted that being able to move around and manipulate virtual components made the learning experience memorable. On the other hand, a few VR participants mentioned challenges: two found the VR controls initially confusing, and others reported feeling overwhelmed at certain points when multiple cues (visual, auditory) happened simultaneously.

Video-trained participants frequently mentioned the clarity and straightforwardness of the video. They appreciated the linear structure and the ability to see an expert perform the procedure correctly. However, some video trainees reported difficulty maintaining focus for the entire duration, and a few wished they could re-watch certain segments. One noted, "The video moved on before I had noted the previous step," indicating that the fixed pace imposed a cognitive load. This suggests that while the video was simple, its one-way, time-bound presentation limited learners' control over the information flow, leading to missed details for some.

Table 2 synthesizes the prominent feedback categories. Notably, both groups expressed a desire for more explanation of why each step mattered and more guidance on what to watch out for, again highlighting the missing conceptual layer in both training approaches. This suggests that blending the two methods or adding supplementary instruction could potentially fill the observed gaps.

4.3.3 Behavioral analysis of VR training

Video coding of VR training sessions revealed patterns in how participants navigated and interacted with the virtual environment. The videos were segmented based on navigation, exploration, and the eight procedural steps (Step

1–Step 8). Within each segment, behavioral codes captured observable actions including listening to instructions, reading step labels, interacting with equipment, looking around for orientation, observing action consequences, hesitation, technical issues, self-correction, confirming actions, and requesting assistance (see Figure 8).

To quantify training "dose" in a self-paced interactive environment, we coded sessions as time-stamped state events and computed per-participant time budgets (see Table 3). Importantly, session duration is a post-treatment process measure in VR (it is partly produced by interface interaction, learner pacing, and scaffolding), rather than a baseline covariate. Consistent with guidance on post-treatment conditioning, we therefore treat time-on-task, adjusted models as sensitivity analyses only and interpret them descriptively, not as the primary causal estimate.
Training duration and time budgets. Training sessions averaged 14.4 minutes (SD = 3.8; median = 14.6; IQR = 3.0; range = 8.1–28.3). Time-budget coding indicated that participants spent 7.4 min (51.4%) in procedural practice (Step1–Step8), 3.1 min (21.8%) in navigation/orientation, 1.5 min (10.6%) in instruction/label processing, and 1.3 min (8.9%) receiving instructor scaffolding (guidance/explanation), with the remaining 1.2 min (7.3%) coded as other behaviors (e.g., hesitation, technical issues, self-correction).

Operational definition and descriptive reporting of assistance. We operationalized *assistance* as a two-part process captured in the videos: (i) a help request (participant raises a question or explicitly asks for help) and (ii) the corresponding guidance episode (duration of investigator explanation/prompting until the participant resumes independent interaction). Assistance was provided only upon request. Participants requested help an average of 2.23 times per session (SD = 0.57; range = 1–4), and total guidance time averaged 76.8 seconds per session (SD = 20.5; range = 36–144 seconds).

Table 11: Training dose and assistance (descriptive).

Condition	Training duration (min) Mean ± SD	Median [IQR]	Assistance events Mean ± SD	Assistance seconds Mean ± SD	% trainees receiving any assistance
VR	14.4 ± 3.8	14.6 [3.0]	2.23 ± 0.57	76.8 ± 20.5	100%
Video	8.12 ± 0.0	8.12 [8.0]	0.0 ± 0.0	0.0 ± 0.0	0%

Where friction concentrated. Time allocation was disproportionately higher in initial steps, driven primarily by interaction demands of the simulated site controller control panel (multiple tabs and detailed parameter settings). This interface was frequently experienced as unintuitive, and participants often required explanation to identify the correct interaction affordances and to verify step completion. In contrast, later steps progressed more smoothly as actions more closely resembled familiar real-world interactions and participants became more fluent with VR controls.

Feedback, hesitation, and conceptual gaps. Behavioral traces indicated that limited in-environment navigation cues and confirmation feedback contributed to repeated or hesitant actions (e.g., toggling switches multiple times). When confirmation cues were present, they reduced uncertainty and supported independent progression. For example, visual feedback from meter readings changing after Steps 2–3 effectively simulated observable system response and served as a task-completion signal. However, the same meter feedback also surfaced a deeper conceptual gap: the most frequent question concerned why the DC disconnect voltage meter did not display zero after being turned off. This pattern suggests that trainees could follow procedural steps while still lacking a coherent mental model of subsystem interconnections and downstream system-state implications. This highlights that procedural fluency alone may be insufficient without conceptual understanding of underlying mechanisms.

Technical disruptions. Technical challenges also disrupted learning. Participants sometimes teleported farther than intended, leading to disorientation and positioning errors. Prior VR navigation research similarly notes that teleportation interfaces can impair spatial orientation because they reduce self-motion cues (Cherep et al., 2020). Design features such as boundaries and navigational feedback can mitigate disorientation.

Exploratory process analyses within the VR group examined whether investigator scaffolding during training was associated with downstream performance (n = 29; see Appendix Table A4). Assistance duration was positively associated with post-training unproductive performance time (Spearman $\rho = 0.57$, $p = 0.001$; OLS with HC3 robust SE: b = 99.7 s per additional minute of assistance, 95% CI [45.6, 153.9], $\beta = 0.50$, $p < .001$) and also with productive execution time (b = 39.2 s, 95% CI [0.7, 77.7], $\beta = 0.47$, $p = .046$). Assistance was not a significant

predictor of performance accuracy (p = .45), consistent with a ceiling effect on accuracy. Because assistance is a post-treatment process measure and not experimentally manipulated, these associations are interpreted diagnostically (as markers of interaction difficulty) rather than causally.

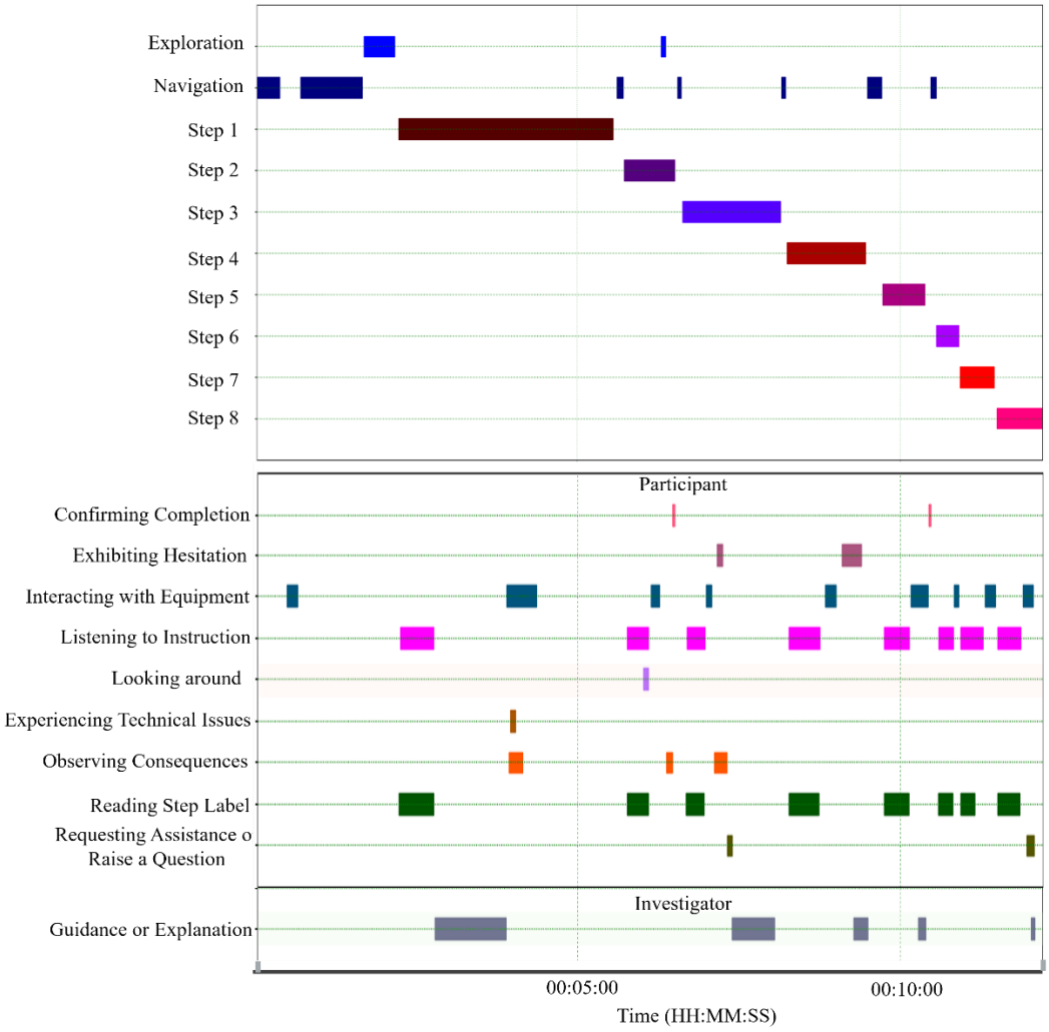

Figure 8: Video coding analysis of training (sample).

5. DISCUSSION

This study examined how immersive VR can enhance safety-critical procedural training. We found that VR provides clear benefits for spatial understanding and procedural performance, but it is not a panacea for all learning needs. The strengths and limitations identified in this study have important implications for designing and implementing safety training in high-risk domains.

5.1 Benefit of VR for safety-critical training: An embodied learning advantage

The results showed that VR significantly improved trainees' understanding of where and how to perform procedural tasks, aligning with principles of embodied cognition (reinforcing knowledge through bodily experience). VR participants effectively rehearsed the shutdown in a realistic context, likely encoding spatial layout and motor actions into memory. In safety-critical tasks, such muscle memory and environmental familiarity are invaluable, allowing workers to act instinctively during an emergency instead of pausing to consult a manual. Prior studies similarly report that immersive training enhances practical safety skills, with VR replicating perceptual cues that facilitate knowledge transfer and reduce execution errors (Evangelista et al., 2025). In our

study, VR-trained participants completed the procedure about 45% faster and showed a trend of higher procedural accuracy, aligning with the idea that VR facilitates the transition from knowledge to action. This progression from knowing steps to fluent execution corresponds to moving from rule-based to skill-based behavior in Rasmussen's Skill–Rule–Knowledge framework (Rasmussen, 1997).

Importantly, we found that trainees who actively integrated visual information with procedural steps achieved superior spatial understanding. For example, those who focused on how spatial cues connected to each task step performed better. This suggests VR's benefits require purposeful engagement rather than passive exposure to immersive environments.

Improvements in procedural comprehension indicate that VR helped learners grasp the logical sequence of the procedure, not just the motor steps. Observing real-time system responses (e.g., meters changing values or indicators lighting up) appeared to foster a better understanding of how steps interrelate. This deeper comprehension is crucial for adapting procedures when unexpected conditions arise during task performance.

The present findings demonstrate improvements in immediate post-training comprehension and simulated execution efficiency following immersive VR training. Because all outcomes were assessed immediately after a single session, these results should be interpreted as short-horizon learning effects rather than evidence of durable retention or on-the-job transfer. This boundary is important in safety-critical domains, where trained skills can decay when rarely used and where transfer requires both generalization to the work context and maintenance over time (Baldwin & Ford, 1988). Future evaluations should therefore incorporate delayed retention tests and field-relevant transfer indicators.

5.2 Engagement and the confidence-competence gap

The high engagement and positive user experience reported for VR are encouraging from a motivational perspective. Trainees who enjoy training experience are more likely to pay attention and retain what they learn. This aligns with research showing that well-designed VR experiences can boost learner motivation and improve learning outcomes (Scorgie et al., 2024). However, high engagement must be harnessed correctly to maintain focus on learning objectives. While we observed mostly productive engagement, VR can potentially cross into "edutainment", experiences that entertain but distract from core goals.

The observed confidence–competence mismatch is consistent with a calibration problem in self-assessment rather than a uniquely "VR" phenomenon. In immersive procedural training, realism, interactivity, and immediate feedback can increase trainees' subjective sense of readiness, sometimes faster than they acquire robust error-detection and rule-bound execution skill. This pattern aligns with established evidence that self-assessments can be inflated when learners lack the expertise required to recognize omissions and boundary conditions in their own performance, even when they feel fluent (Kruger & Dunning, 1999). For BESS operations, the consequence is not merely pedagogical: overconfidence can be safety-relevant because incorrect sequencing or incomplete verification during de-energization elevates exposure to electrical hazards. In applied safety domains, overconfidence has been linked to reduced motivation for refresher training, reinforcing the need to design training that calibrates confidence, not only raises engagement (Bushuven et al., 2023). A practical implication is to integrate (i) objective performance metrics (accuracy, critical-error flags, time-to-safe-state), (ii) structured calibration feedback (post-task error reviews that compare perceived vs observed performance), and (iii) scenario-based assessments with off-nominal conditions (e.g., ambiguous indicators, partial shutdown states) that require diagnostic reasoning rather than rote step execution. Where feasible and ethically appropriate, programs should also include delayed retention checks to verify that high confidence persists only when competence persists.

5.3 Limits of VR: The missing "Why" and cognitive challenges

Mastering a procedure requires understanding why each step is performed, yet neither the VR nor the video training conveyed the underlying safety rationale. As a result, trainees in both conditions could execute the steps but lacked a robust mental model of the hazards and reasons behind them. This shallow understanding is risky because a worker might skip or alter a step not realizing its importance, especially when facing a novel situation that requires deviating from the script. In safety terms, this is an active failure rooted in a latent knowledge gap, a known contributor to accidents (Pantelidis, 2010). In Rasmussen's framework, VR trainees remained at a rule-based level without reaching a knowledge-based understanding of the task's rationale (Rasmussen, 1997). Thus, like many

VR training systems, our prototype did not close this conceptual gap. Effective safety training must integrate explanatory content alongside hands-on practice. Effective safety training must therefore integrate explanatory content (the "why") alongside hands-on practice. For example, the VR simulation could explicitly show the consequences of an incorrect action (e.g., trigger a virtual arc flash when a disconnect is opened under load, along with an explanation of the hazard). Alternatively, a blended approach could follow the VR session with an instructor-led discussion reviewing each step's purpose and the potential consequences of doing it improperly.

Another limitation is the potential for cognitive overload if the VR simulation is not carefully designed. Despite our effort to create an intuitive interface, some participants still felt overwhelmed by operating the VR system while learning the procedure. Cognitive Load Theory predicts that if total cognitive load (intrinsic task complexity *plus* extraneous load from the interface) exceeds the learner's capacity, learning will suffer (Sweller, 1994). A well-designed VR simulation can mitigate extraneous load by providing intuitive cues and immediate feedback, but it can also increase load if controls are too complicated. The lack of a VR advantage in step-sequence recall suggests VR trainees may not have deeply encoded the sequence, possibly because some cognitive capacity was diverted to navigating the virtual environment. A complementary explanation is that both groups received comparable memory aids (on-screen step lists/labels), and the task was sufficiently easy to produce ceiling effects. In general, VR tends to shift cognitive load rather than eliminate it: it reduces load by making information more visual and interactive but increases load by adding new interactions to manage. In contrast, the video had minimal interactive load but may have imposed higher mental load for trainees to imagine performing the task. Neither approach in our experiment achieved an optimal balance of cognitive load.We also observed individual differences in VR proficiency. Some participants struggled with basic navigation and required step-by-step guidance, whereas others quickly mastered the interface and were ready for deeper learning. Future VR training should accommodate these differences by applying evidence-based design principles (Mayer, 2009) and VR-specific guidelines (Evangelista et al., 2025). For example, proficiency-adaptive features could adjust information delivery in real time based on performance. Providing an intuitive interface, just-in-time cues, and user-controlled pacing (e.g., the ability to pause and confirm understanding before proceeding) can support novices without boring advanced learners. These measures would help prevent cognitive overload across a range of skill levels.

5.4 Practical implications for integrating VR into safety training programs

Evidence supports a blended training strategy that combines the experiential strengths of VR with the explanatory reinforcement of other methods. These findings suggest VR should be integrated with explicit conceptual briefings and performance-calibrating assessments rather than deployed in isolation. Meta-analytic evidence indicates that VR effects vary by outcome category and context, reinforcing the need to pair engagement gains with competence-focused measures and explicit rationale instruction (Man et al., 2024).This multi-modal approach aligns with the Resilience Engineering emphasis on flexibility and adaptation. VR simulations can serve as safe rehearsal spaces where workers build practical skills and develop a feel for task dynamics. Meanwhile, classroom or video sessions (or augmented-reality job aids) can focus on theory, hazard awareness, and "what if" scenarios. By merging these methods, a training program can address multiple layers of defense. VR covers the performance layer (ensuring trainees can execute the task), while supplemental instruction covers the knowledge layer (ensuring trainees understand why steps need to be done in sequence and what pitfalls to avoid). In our study, had we followed the VR session with a short interactive discussion of key hazards and rationales (or vice versa), we might have improved those outcomes where each method alone fell short. This insight aligns with prior research showing that multi-modal training often yields better results for complex skill acquisition (Scorgie et al., 2024).

Our evaluation of VR's impact on different learning dimensions (skills, rules, knowledge) echoes the Skill–Rule–Knowledge framework. We found that VR boosted skill-based and some rule-based performance but had much less effect on knowledge-based reasoning. This suggests future VR modules should include branching scenarios that require trainees to exercise judgment when something is off-nominal, so they practice adaptive decision-making rather than just following rote sequences. Such scenarios would foster resilience (i.e., the ability to respond to the unexpected), in line with Resilience Engineering principles. Developing an adaptable understanding ensures trainees can handle situations where the procedure might need adjustment (within safe limits) instead of failing when a novel situation arises. As Pariès and Wreathall (2017) argue, training for resilience means exposing people to variability and teaching them to detect and recover from potential error traps. VR is well-suited for this, as it can safely simulate rare but critical scenarios (for example, a component failure during shutdown).As noted earlier, VR's immersive feedback can inflate confidence. Therefore, training programs should incorporate reality checks

to keep trainees' self-assessments accurate. For instance, integrating objective post-training assessments or guided debriefings on mistakes helps calibrate trainees' perceptions to match their actual performance. Additionally, gradually exposing trainees to the real task environment under supervision can ensure that VR's high engagement and confidence translate into genuine competence rather than complacency. A staged training sequence may be optimal: begin with a lecture or video for basic familiarity, then use VR for hands-on practice, and finally conduct an on-site walkthrough with the real equipment. This layered approach echoes Rasmussen's concept of multiple safety barriers in a socio-technical system (Rasmussen, 1997). VR training can proactively shape operator behavior, but it works best when combined with other complementary safety measures.

Beyond these immediate steps, training should be viewed as an ongoing process. We measured learning right after a single VR session, but in practice, critical procedures likely require periodic refreshers to ensure long-term retention. VR is well suited for refresher training because it can safely simulate rare emergency scenarios with increasing complexity or stress to build preparedness. It is also important to test understanding sometime after training (e.g., with a quiz or practical demonstration) to evaluate long-term retention of safety-critical steps. If retention is low, the training frequency or content difficulty should be adjusted accordingly.

By implementing these strategies, organizations can better align their training programs with human cognition and the unforgiving nature of safety-critical work. The positive results from our study are encouraging, but they also underscore that technology is not a standalone solution. Rather, it must be embedded with a well-designed instructional and safety framework.

5.5 Limitations and future work

This study has several limitations, which suggest directions for future research and implementation. First, our participant sample comprised engineering students rather than professional battery technicians, limiting generalization to the industry workforce. Future studies should recruit apprentices and practice technicians (across multiple sites) to improve external validity.

Second, our evaluation focused on a single BESS soft-shutdown procedure in a controlled environment. Real-world battery operations involve broader activities (e.g., fault diagnosis, emergency response, team coordination) that were beyond our scope. Examining a wider set of tasks would establish the breadth of VR's applicability and help determine which types of tasks benefit most from immersive training.

Third, Outcomes were measured immediately post-training. step-sequence recall exhibited a ceiling pattern, with many participants scoring at or near the upper bound. This restricted variance limits the sensitivity of the immediate post-test sorting task to detect between-group differences. Thus, we cannot infer durability of learning or translation to worksite behaviors. Comparable safety-relevant VR evaluations that include delayed assessments show why this distinction matters. For example, a low-voltage electrical safety VR study measured knowledge immediately and again four weeks later, finding significant immediate gains alongside measurable decline at four weeks (while still exceeding baseline) (Stefan et al., 2024). This pattern illustrates that delayed testing can materially change conclusions about the stability of training benefits and motivates multi-interval follow-up in future BESS safety training research. Also, the measure primarily captured short-horizon memory for nominal step order rather than discriminating higher-order procedural judgment. Future evaluations will therefore incorporate more discriminating, BESS-realistic sequence measures that test conditional decision points and hazard-relevant verification: (a) indicator-ambiguity vignettes in which trainees must choose an appropriate verification step when feedback is inconsistent with expectations (e.g., a disconnect indicator not behaving as assumed), (b) faulted step-order recognition items contrasting a correct sequence with a plausible but unsafe variant and requiring trainees to identify the hazard mechanism, and (c) rationale-scored and time-pressured recall for critical steps to probe the "why" underlying safe execution. Such formats embed ambiguity and decision relevance, characteristics known to improve discrimination compared with low-complexity recall items.

Fourth, the post-training test was administered in VR, which likely gave the VR-trained group a modality advantage. Using a physical mock-up or real equipment for testing and focusing on field-relevant indicators (e.g., checklist adherence, near-miss reports, time-to-safe-state) would provide a more neutral assessment of skill transfer. Because the test requires VR interaction, we interpret performance outcomes as a combined measure of procedural execution and interface enactment. Future work will triangulate with physical mock-ups or on-site walk-throughs.

Fifth, although our results indicate a need for proficiency-adaptive VR systems, we did not implement adaptive features in our prototype. Future work should develop VR training platforms that dynamically adjust information presentation based on learner performance, providing appropriate scaffolding for novices while offering deeper challenges for advanced learners. Experiments that manipulate factors such as guidance level, conceptual scaffolding of safety rationale, and scenario variability would help identify optimal design parameters and clarify how immersive technologies can best produce resilient practitioners who handle unexpected situations safely. Sixth, a limitation of the psychomotor analysis is that the observed efficiency advantage may partly reflect differences in time-on-task rather than a pure medium effect. This concern echoes the classic media-versus-method debate: media may function as vehicles for instruction, but their affordances can also interact with instructional methods and learner processing (Kozma, 1994). Because time-on-task is a post-treatment variable rather than a baseline covariate, adjusting for it would change the target effect and should not replace the primary intent-to-treat comparison. We therefore treat any time-adjusted model as descriptive sensitivity analysis only and interpret the main efficiency effect from the original randomized comparison.

Finally, several practical implementation factors require investigation. These include strategies to mitigate VR-induced cybersickness, determining the optimal training duration and spacing (dose) for effectiveness, conducting cost–benefit analyses, and establishing best practices for integrating VR with traditional instructional methods. It is also necessary to establish systematic methods for deciding which content is best delivered through immersive interaction versus what is better conveyed via traditional explanations.

Addressing these limitations and research directions will help define the boundaries and optimal conditions for VR's effectiveness as a safety training tool. This will ultimately inform evidence-based decisions about when and how to deploy immersive training technologies in high-risk industries.

6. CONCLUSION

This study demonstrates that immersive VR training offers significant advantages over traditional video instruction for safety-critical BESS procedures, particularly in building spatial understanding (27% improvement) and procedural comprehension (28% improvement). VR-trained participants completed productive tasks 45% faster while maintaining comparable accuracy and reported higher engagement despite increased cognitive demands. These benefits illustrate how VR's immersive, interactive nature can address limitations of passive training by providing embodied learning experiences that more closely mimic real-world task execution. However, the durability of these gains and their transfer to field performance remain to be tested in longitudinal and workplace-based evaluations. For high-risk procedures where mistakes can be catastrophic, such as handling high-voltage BESS equipment, this kind of realistic practice is invaluable. VR essentially allows trainees to fail safely and learn from those mistakes, which is a cornerstone of building a strong safety culture and resilient practitioners.

However, our findings reveal important limitations that temper enthusiasm for VR as a complete training solution. VR did not improve sequential recall compared to conventional training, and critically, neither method adequately conveyed the safety rationale underlying procedural sequences. Additionally, a confidence-competence gap emerged, with VR participants' self-assessed abilities exceeding objective performance in some domains, suggesting the need for structured assessment in safety-critical training. These findings underscore that effective safety training must integrate experiential learning with explicit cognitive instruction about underlying principles and hazards.

From a practical standpoint, this research has important implications for construction safety management and workforce development. As BESS installations become more common on construction projects and within facility operations, ensuring that the workforce is adept at handling these systems safely is crucial. Integrating VR modules into construction safety training programs can offer a safe, cost-effective means to practice emergency procedures that would otherwise be too risky to train for. Companies involved in the design, construction, and maintenance of BESS and other high-risk equipment can adopt immersive simulations to improve their employees' preparedness, thereby potentially reducing accident rates and downtime. In the broader context of applied information technology in the built environment, this study exemplifies how emerging digital tools like VR can be leveraged to enhance human performance and safety across the facility lifecycle. By embedding interactive virtual training in the construction and operation phases (e.g., training installation crews, commissioning engineers, and maintenance staff), stakeholders can ensure knowledge transfer and skills development keep pace with technological advancements in infrastructure.

For the energy storage industry and other technical domains, these findings argue for comprehensive training architectures beyond simple technology adoption. The question is not whether to use VR, but how to architect information within immersive environments for progressive skill development while maintaining cognitive efficiency, and how to integrate VR with complementary instructional methods. When properly integrated with sound pedagogical strategies and grounded in safety theory, VR training can potentially help transform novices into competent, safety-conscious professionals ready to handle the complex challenges of the modern energy workplace. This proactive investment in human performance reliability, effectively "building a stronger cheese layer" in Reason's metaphor (Reason, 1990), represents a critical element of comprehensive safety management that can reduce the likelihood of accidents and enhance overall system safety. By moving toward an era where advanced simulation and safety management go hand in hand, we can work to prevent accidents before they happen, protecting both workers and the critical infrastructure they operate.

ACKNOWLEDGMENTS

The authors thank the inter-rater reliability researchers who contributed to validating our assessment methods. We also extend our gratitude to the undergraduate students who participated in this study, as well as the energy storage industry professionals who provided expert feedback during the development and validation of the training modules.

DISCLOSURE

The authors used Claude for grammar checking and sentence rewording suggestions. All AI-generated suggestions were reviewed and approved by the authors, who take full responsibility for the final content and accuracy of this manuscript.

DATA AVAILABILITY

Individual data unavailable due to IRB restrictions. Aggregate results, analysis code, survey instruments, and codebook are available from the corresponding author upon request.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

AlQallaf, N., AlQallaf, A., & Ghannam, R. (2024). Solar Energy Systems Design Using Immersive Virtual Reality: A Multi-Modal Evaluation Approach. Solar, 4(2), 329–350. https://doi.org/10.3390/solar4020015

Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological Review, 89(4), 369–406. https://doi.org/10.1037/0033-295X.89.4.369

Babalola, A., Manu, P., Cheung, C., Yunusa-Kaltungo, A., & Bartolo, P. (2023). A systematic review of the application of immersive technologies for safety and health management in the construction sector. Journal of Safety Research, 85, 66–85. https://doi.org/10.1016/j.jsr.2023.01.007

Baldwin, T. T., & Ford, J. K. (1988). TRANSFER OF TRAINING: A REVIEW AND DIRECTIONS FOR FUTURE RESEARCH. Personnel Psychology, 41(1), 63–105. https://doi.org/10.1111/j.1744-6570.1988.tb00632.x

Bavishi, P., Birnhak, A., Gaughan, J., Mitchell-Williams, J., & Phadtare, S. (2022). Active Learning: A Shift from Passive Learning to Student Engagement Improves Understanding and Contextualization of Nutrition and Community Health. Education Sciences, 12(7), 430. https://doi.org/10.3390/educsci12070430

Brady, A., & Naikar, N. (2022). Development of Rasmussen's risk management framework for analysing multi-level sociotechnical influences in the design of envisioned work systems. Ergonomics, 65(3), 485–518. https://doi.org/10.1080/00140139.2021.2005823

Bushuven, S., Bansbach, J., Bentele, M., Trifunovic-Koenig, M., Bentele, S., Gerber, B., Hagen, F., Friess, C., & Fischer, M. R. (2023). Overconfidence effects and learning motivation refreshing BLS: An observational questionnaire study. Resuscitation Plus, 14, 100369. https://doi.org/10.1016/j.resplu.2023.100369

Cherep, L. A., Lim, A. F., Kelly, J. W., Acharya, D., Velasco, A., Bustamante, E., Ostrander, A. G., & Gilbert, S. B. (2020). Spatial cognitive implications of teleporting through virtual environments. Journal of Experimental Psychology: Applied, 26(3), 480–492. https://doi.org/10.1037/xap0000263

Clark, R. E. (1983). Reconsidering Research on Learning from Media. Review of Educational Research, 53(4), 445–459. https://doi.org/10.3102/00346543053004445

Cochrane, T., Aiello, S., Wilkinson, N., Aguayo, C., & Cook, S. (2022). Developing A Mobile Immersive Reality Framework For Enhanced Simulation Training: MESH360. ASCILITE Publications, 392–397. https://doi.org/10.14742/apubs.2019.294

Daling, L. M., & Schlittmeier, S. J. (2024). Effects of Augmented Reality-, Virtual Reality-, and Mixed Reality–Based Training on Objective Performance Measures and Subjective Evaluations in Manual Assembly Tasks: A Scoping Review. Human Factors: The Journal of the Human Factors and Ergonomics Society, 66(2), 589–626. https://doi.org/10.1177/00187208221105135

Eiris, R., Jain, A., Gheisari, M., & Wehle, A. (2020). Safety immersive storytelling using narrated 360-degree panoramas: A fall hazard training within the electrical trade context. Safety Science, 127, 104703. https://doi.org/10.1016/j.ssci.2020.104703

Evangelista, A., Manghisi, V. M., De Giglio, V., Mariconte, R., Giliberti, C., & Uva, A. E. (2025). From knowledge to action: Assessing the effectiveness of immersive virtual reality training on safety behaviors in confined spaces using the Kirkpatrick model. Safety Science, 181, 106693. https://doi.org/10.1016/j.ssci.2024.106693

Exeter Associates. (2022). Siting and Safety Best Practices for Battery Energy Storage Systems (PPAD-BESS-2022-01; p. 14). Maryland Department of Natural Resources, Power Plant Research Program (PPRP). https://dnr.maryland.gov/pprp/Documents/PPAD-BESS-2022-01-Report.pdf

Gao, Y., Gonzalez, V. A., & Yiu, T. W. (2019). The effectiveness of traditional tools and computer-aided technologies for health and safety training in the construction sector: A systematic review. Computers & Education, 138, 101–115. https://doi.org/10.1016/j.compedu.2019.05.003

Getuli, V., Bruttini, A., Sorbi, T., Fornasari, V., & Capone, P. (2024). Integrating spatial tracking and surveys for the evaluation of construction workers' safety training with virtual reality. Journal of Information Technology in Construction, 29, 1181–1199. https://doi.org/10.36680/j.itcon.2024.052

Gonzalez Lopez, J. M., Jimenez Betancourt, R. O., Ramirez Arredondo, J. M., Villalvazo Laureano, E., & Rodriguez Haro, F. (2019). Incorporating Virtual Reality into the Teaching and Training of Grid-Tie Photovoltaic Power Plants Design. Applied Sciences, 9(21), 4480. https://doi.org/10.3390/app9214480

Hajirasouli, A., & Banihashemi, S. (2022). Augmented reality in architecture and construction education: State of the field and opportunities. International Journal of Educational Technology in Higher Education, 19(1), 39. https://doi.org/10.1186/s41239-022-00343-9

Idaho National Laboratory. (2024). Battery Energy Storage Systems Report. U.S. Department of Energy.

Jeevarajan, J. A., Joshi, T., Parhizi, M., Rauhala, T., & Juarez-Robles, D. (2022). Battery Hazards for Large Energy Storage Systems. ACS Energy Letters, 7(8), 2725–2733. https://doi.org/10.1021/acsenergylett.2c01400

Kozma, R. B. (1994). Will media influence learning? Reframing the debate. Educational Technology Research and Development, 42(2), 7–19. https://doi.org/10.1007/BF02299087

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. Journal of Applied Psychology, 78(2), 311–328. https://doi.org/10.1037/0021-9010.78.2.311

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134. https://doi.org/10.1037//0022-3514.77.6.1121

Man, S. S., Wen, H., & So, B. C. L. (2024). Are virtual reality applications effective for construction safety training and education? A systematic review and meta-analysis. Journal of Safety Research, 88, 230–243. https://doi.org/10.1016/j.jsr.2023.11.011

Mayer, R. E. (2002). Multimedia learning. In Psychology of Learning and Motivation (Vol. 41, pp. 85–139). Elsevier. https://doi.org/10.1016/S0079-7421(02)80005-6

Mayer, R. E. (2009). Multimedia Learning (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

McLeod, D. B., Cronbach, L. J., & Snow, R. E. (1978). Aptitudes and Instructional Methods: A Handbook for Research on Interactions. Journal for Research in Mathematics Education, 9(5), 390. https://doi.org/10.2307/748778

National Fire Protection Association. (2026). NFPA 855 Standard Development. NFPA. https://www.nfpa.org/codes-and-standards/nfpa-855-standard-development/855

National Science Foundation. (2015). NSF Award Search: Award # 1501883. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1501883

Pantelidis, V. S. (2010). Reasons to use virtual reality in education and training courses and a model to determine when to use virtual reality. Themes in Science and Technology Education, 2(1–2), 59–70.

Pariès, J., & Wreathall, J. (2017). Resilience Engineering in Practice: A Guidebook (E. Hollnagel, J. Pariès, D. Woods, & J. Wreathall, Eds.; 1st ed.). CRC Press. https://doi.org/10.1201/9781317065265

Prakash, K., Ali, M., Siddique, M. N. I., Chand, A. A., Kumar, N. M., Dong, D., & Pota, H. R. (2022). A review of battery energy storage systems for ancillary services in distribution grids: Current status, challenges and future directions. Frontiers in Energy Research, 10. https://doi.org/10.3389/fenrg.2022.971704

Rasmussen, J. (1997). Risk management in a dynamic society: A modelling problem. Safety Science, 27(2–3), 183–213. https://doi.org/10.1016/S0925-7535(97)00052-0

Reason, J. (1990). Human error. Cambridge university press.

Rey-Becerra, E., Barrero, L. H., Ellegast, R., & Kluge, A. (2021). The effectiveness of virtual safety training in work at heights: A literature review. Applied Ergonomics, 94, 103419. https://doi.org/10.1016/j.apergo.2021.103419

Sacks, R., Perlman, A., & Barak, R. (2013). Construction safety training using immersive virtual reality. Construction Management and Economics, 31(9), 1005–1017. https://doi.org/10.1080/01446193.2013.828844

Scorgie, D., Feng, Z., Paes, D., Parisi, F., Yiu, T. W., & Lovreglio, R. (2024). Virtual reality for safety training: A systematic literature review and meta-analysis. Safety Science, 171, 106372. https://doi.org/10.1016/j.ssci.2023.106372

Speiser, K., & Teizer, J. (2024). Formalizing virtual construction safety training: A schematic data framework enabling real-world hazard simulations using BIM and location tracking. Journal of Information Technology in Construction, 29, 980–1004. https://doi.org/10.36680/j.itcon.2024.043

Stefan, H., Mortimer, M., & Horan, B. (2023). Evaluating the effectiveness of virtual reality for safety-relevant training: A systematic review. Virtual Reality, 27(4), 2839–2869. https://doi.org/10.1007/s10055-023-00843-7

Stefan, H., Mortimer, M., Horan, B., & McMillan, S. (2024). How effective is virtual reality for electrical safety training? Evaluating trainees' reactions, learning, and training duration. Journal of Safety Research, 90, 48–61. https://doi.org/10.1016/j.jsr.2024.06.002

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction, 4(4), 295–312.

Taibi, E., Nikolakakis, T., Valle, C. F. D., Aakarshan Vaid, Yu, A., Vinayak Walimbe, Tinkler, M., & Johnson, R. (2020). Electricity Storage Valuation Framework: Assesing system value and ensuring project viability. International Renewable Energy Agency (IRENA). https://doi.org/10.13140/RG.2.2.30598.73281

Tcha-Tokey, K., Christmann, O., Loup-Escande, E., & Richir, S. (2016). Proposition and Validation of a Questionnaire to Measure the User Experience in Immersive Virtual Environments. International Journal of Virtual Reality, 16(1), 33–48. https://doi.org/10.20870/IJVR.2016.16.1.2880

Viswanathan, V., Mongird, K., Franks, R., Li, X., Sprenkle, V., & Baxter, R. (2022). 2022 Grid Energy Storage Technology Cost and Performance Assessment (PNNL-33283; Energy Storage Grand Challenge Cost and Performance Assessment). Pacific Northwest National Laboratory.

Wang, X., & Messner, J. I. (2020). The Pedagogical Value of Virtual Reality Training for Electrical Workers on Energy Storage and Microgrid Systems. Construction Research Congress 2020, 574–582. https://doi.org/10.1061/9780784482872.062

Wang, X., Messner, J. I., Leicht, R. M., & Mackey, A. (2024). Virtual Reality Design Features That Impact the Efficiency and Accuracy of Electrical Workforce Training Scenarios. Computing in Civil Engineering 2023, 623–631. https://doi.org/10.1061/9780784485224.075

Wilson, M. (2002). Six views of embodied cognition. Psychonomic Bulletin & Review, 9(4), 625–636. https://doi.org/10.3758/BF03196322

APPENDIX A

Table A1: Post-training cognitive assessment questions.

Hypothesis	Learning Objectives		Q#	Question
H1a	Spatial Understanding		Q6-c	Determine whether the following statement is correct or incorrect: "Before opening the main breaker of the battery enclosure, the PV disconnect must be turned off."
			Q7	Place the corresponding step numbers next to the equipment where each step is to be performed.
H1b	Procedural Knowledge Understanding	Recognize Completion Indicators	Q4	What indicates that the AC Disconnect inverter is turned off?
			Q6-d	Determine whether the following statement is correct or incorrect: "Upon switching off the Battery DC Disconnect, all meter readings turn to 0."
		Understand the Purpose of Each Step	Q3	Which step ensures the batteries are fully isolated from the entire Energy Storage System?
			Q5	If the battery enclosure door does not open during the soft shutdown activity, even after pulling with force, which piece of equipment should you check first?
			Q6-a	Determine whether the following statement is correct or incorrect: "We don't need to turn off the battery enclosure contactor in the site controller because the main breaker in the battery enclosure will be opened eventually."
			Q6-b	Determine whether the following statement is correct or incorrect: "The power to the battery enclosure door lock mechanism must be turned on immediately after the battery enclosure door is opened."
H1c	Step and Sequence Recall		Q1	How many steps are required to complete the soft shutdown activity?
			Q2	Identify and rank the necessary steps to complete the soft shutdown activity.

Table A2: Complete user experience survey questions and descriptive statistics by item (POMP 0-100).

Scale	Subscale	Q #	Full question	VR Mean ± SD	Video Mean ± SD	All Mean ± SD
Judgment	Satisfaction	Q5-1	Rate your overall experience: Terrible to Amazing	87.3 ± 12.3	68.0 ± 17.1	77.7 ± 17.7
	Enjoyment	Q5-2	Rate your overall experience: Annoying to Enjoyable	86.0 ± 16.7	71.3 ± 20.1	78.7 ± 19.8
	Usability	Q5-3	Rate your overall experience: Confusing to Intuitive	72.7 ± 24.9	66.0 ± 24.7	69.3 ± 24.8
	Engagement	Q5-4	Rate your overall experience: Boring to Engaging	92.0 ± 16.3	67.3 ± 27.5	79.7 ± 25.6
	Stimulation	Q5-5	Rate your overall experience: Dull to Stimulating	89.3 ± 17.2	62.7 ± 26.1	76.0 ± 25.7
Presence and Immersion	Presence	Q8-4	I became so immersed in the virtual environment that it felt as if I were inside an actual energy storage system.	80.8 ± 19.3	51.7 ± 30.7	66.2 ± 29.4
	Focus (time loss)	Q8-13	I was losing the sense of time during the training.	60.8 ± 27.6	48.3 ± 24.5	54.6 ± 26.6
Engagement	Involvement	Q8-3	I could concentrate on the assigned task while taking the training.	84.2 ± 19.1	62.5 ± 31.3	73.3 ± 27.9
	Attention (mind wandering) [a]	Q8-11	I found my mind wandering while I was in the environment.	68.3 ± 25.4	54.2 ± 30.9	61.2 ± 28.9
	Cognitive engagement	Q8-12	At each step, I knew what was going on while I was taking the training.	62.5 ± 21.5	56.7 ± 21.7	59.6 ± 21.6
Perception of Learning	Perceived effectiveness	Q8-1	I think I learned a lot from this hands-on experience.	81.7 ± 20.7	73.3 ± 19.6	77.5 ± 20.4
	Perceived retention	Q8-2	I think using the simulation enables me to memorize information.	82.5 ± 21.9	65.0 ± 26.7	73.8 ± 25.8
	Learning confidence	Q8-5	I feel confident learning knowledge and skills within the immersive environment.	81.7 ± 16.0	62.5 ± 26.1	72.1 ± 23.5
	Skill transfer	Q8-6	I feel confident performing the real task after the training.	68.3 ± 16.0	54.2 ± 28.7	61.2 ± 24.1
Cognitive Load	Coping with instructions [a]	Q8-10	I worried whether I was able to cope with all the instructions during the training.	46.7 ± 25.2	30.8 ± 25.2	38.8 ± 26.2
Usability	Ease of use	Q8-14	The interaction was easy to use / video was easy to follow.	83.3 ± 22.1	76.7 ± 19.6	80.0 ± 21.0
Technology Adoption	Practicality	Q8-7	Personally, I would say the training is practical.	90.8 ± 16.7	75.0 ± 22.7	82.9 ± 21.3
	User autonomy	Q8-8	I like the self-paced learning mode of the simulation.	92.5 ± 16.3	65.8 ± 26.7	79.2 ± 25.7
Emotion	Enjoyment/Energy	Q8-9	I enjoyed the experience so much that I felt energized.	78.3 ± 22.5	44.2 ± 28.4	61.2 ± 30.7

Note. Numeric cells are POMP values (0-100; higher = better). Items [a] were reverse-coded so that higher indicates better.

Table A3: User Experience Item-Level Statistical Analysis (POMP 0-100).

Scale	Item	Q #	Mean (VR)	SE (VR)	Mean (Video)	SE (Video)	F	p (two-tailed)	p (Holm, 19)	Hedges g
Judgment	Satisfaction	5-1)	87.4	2.2	68.0	3.2	25.113	<.001	<.001	1.28
	Enjoyment	5-2)	86.0	3.0	71.2	3.6	9.613	.003	.033	0.79
	Usability	5-3)	72.6	4.6	66.0	4.6	1.062	.307	.692	0.26
	Overall engagement	5-4)	92.0	3.0	67.4	5.0	17.614	<.001	.001	1.07
	Stimulation	5-5)	89.2	3.2	62.6	4.8	21.610	<.001	<.001	1.18
Presence/ Immersion	Presence	8-4)	80.8	3.5	51.7	5.5	19.170	<.001	<.001	1.12
	Focus (time loss)	8-13)	60.8	5.0	48.3	4.5	3.456	.068	.366	0.47
Engagement	Involvement	8-3)	84.2	3.5	62.5	5.8	10.610	.002	.023	0.83
	Attention (mind wandering) [a]	8-11)	68.2	4.8	54.2	5.8	3.649	.061	.366	0.49
	Cognitive engagement	8-12)	62.5	4.0	56.8	4.0	1.060	.307	.692	0.26
Perception of Learning	Perceived effectiveness	8-1)	81.7	3.8	73.2	3.5	2.673	.107	.430	0.42
	Perceived retention	8-2)	82.5	4.0	65.0	4.8	7.659	.008	.068	0.71
	Learning confidence	8-5)	81.7	3.0	62.5	4.8	11.928	.001	.014	0.88
	Skill transfer	8-6)	68.2	3.0	54.2	5.2	5.432	.023	.163	0.59
Cognitive Load	Coping with instructions [a]	8-10)	46.8	4.5	30.8	4.5	6.023	.017	.137	0.63
Usability	Ease of use	8-14)	83.2	4.0	76.8	3.5	1.474	.231	.692	0.31
Technology Adoption	Practicality	8-7)	90.8	3.0	75.0	4.2	9.147	.003	.034	0.78
	User autonomy	8-8)	92.5	3.0	65.8	5.4	21.826	<.001	<.001	1.19
Emotion/Energy	Enjoyment/Energy	8-9)	78.2	4.0	44.2	5.2	26.303	<.001	<.001	1.31

Note. Means/SEs are POMP scores (0-100; higher = more favorable). Reverse-coded items are marked [a] and were reverse-coded before POMP conversion. Two-tailed p values are the primary inferential presentation; one-tailed p values (VR > Video) are shown only as secondary, preregistered directional confirmation. Holm-adjusted p values control the family-wise error rate across the 19 between-group tests in this table.

Table A4: Assistance during VR training and exploratory within-VR associations with performance outcomes

Panel A. Assistance descriptives (VR training, N = 30)

Assistance duration (min)	1.28 ± 0.34 (median 1.30; IQR 1.13-1.40; range 0.60-2.40)
Assistance requests (count)	2.23 ± 0.57 (median 2; range 1-4)
Guidance duration (s)	76.8 ± 20.5 (median 78; range 36-144)
Guidance per request (s)	34.6 ± 5.3 (median 36; range 21-42)

Panel B. Exploratory OLS within VR (n = 29; HC3 robust SE)

Outcome	b	SE	95% CI	β	p	R^2
Unproductive time (s)	99.7	27.6	[45.6, 153.9]	0.50	< .001	0.254
Productive time (s)	39.2	19.6	[0.7, 77.7]	0.47	.046	0.226
Accuracy (0-14)	-1.07	1.42	[-3.86, 1.72]	-0.39	.451	0.154

Notes. Panel A reports mean ± SD with median and dispersion statistics in parentheses. Panel B reports separate OLS models predicting each performance outcome from assistance duration (minutes) within the VR group; b is the unstandardized coefficient per additional minute of assistance, β is the standardized coefficient, and robust standard errors use the HC3 correction. One VR participant had missing performance-test values, yielding n = 29 for Panel B.

Table A5: Sensitivity of primary outcomes to adjustment for gender and prior procedural experience.

Outcome	Unadjusted (VR-Video)	Adjusted [a] (VR-Video)	% Change [b]	Conclusion
Spatial understanding (POMP)	1.53 [0.55, 2.50]	1.30 [0.40, 2.20]	15%	Stable (VR > Video)
Procedural knowledge (POMP)	1.00 [0.46, 1.53]	0.85 [0.53, 1.17]	15%	Stable (VR > Video)
Step & sequence recall (POMP)	0.18 [-1.11, 1.46]	0.20 [-1.00, 1.40]	11%	Stable (no difference)
Productive time (s, VR-Video)	48.0 [30.0, 66.0]	40.0 [22.0, 58.0]	17%	Stable (VR faster)

Notes. Values show mean difference with 95% confidence interval using robust standard errors. [a] Adjusted for gender (binary) and prior procedural experience (binary: any vs. none). [b] Percent change from unadjusted to adjusted effect estimate. All changes ≤17%, indicating robustness to baseline imbalances.

Table A6: Exploratory analysis: attention allocation (post-training) and preferred learning method (pre-training).

Panel A. Multivariate tests (Pillai's Trace)

Effect	Pillai	Approx. F	df1	df2	p
Attention allocation	0.084	1.644	3	54	.190
Preferred learning method	0.042	0.780	3	54	.510

Panel B. Univariate associations (ANCOVA coefficients; full sample, N = 60)

Outcome	Predictor	b	SE	t	95% CI	β	p
Spatial understanding	Attention allocationa	0.47	0.21	2.26	[0.05, 0.89]	0.27	.028
Spatial understanding	Preferred learning methodb	0.12	0.36	0.32	[-0.60, 0.84]	0.04	.751
Procedural knowledge understanding	Attention allocationa	0.10	0.11	0.91	[-0.13, 0.33]	0.11	.369
Procedural knowledge understanding	Preferred learning methodb	0.21	0.21	1.01	[-0.21, 0.60]	0.12	.318
Step & sequence recall	Attention allocationa	0.32	0.28	1.15	[-0.24, 0.87]	0.15	.255
Step & sequence recall	Preferred learning methodb	-0.33	0.48	-0.70	[-1.29, 0.63]	-0.09	.484

Notes. Panel B reports separate ANCOVA coefficients for each outcome. Each model included training condition and both predictors. Standardized β values were computed as b × SD(X)/SD(Y).

Table A7: Exploratory sensitivity analysis of performance accuracy with post-treatment time adjustment.

Model	Group difference in accuracy (pp)	p	Time coefficient	p	Interpretation
Primary accuracy ANCOVA	+6.6	.053	—	—	Primary ITT-style estimate
Time-adjusted exploratory model	+5.0	.171	−0.018 pp/s	.204	Descriptive, post-treatment-adjusted

Table A8: Covariate effects on performance outcomes (full sample, N = 58).

A. Multivariate Tests for Covariates (Pillai's Trace)

Effect	Pillai's Trace	F	df	p-value
Prior VR Experience	0.062	1.131	3, 51	.345
Video-Game Experience	0.032	0.570	3, 51	.637
VR Confidence	0.069	1.259	3, 51	.298

B. Univariate Covariate Coefficients

Outcome	Covariate	b	SE	t	95% CI	Std. β	p
Performance Accuracy (POMP 0-100)	Prior VR Experience	-0.80	1.80	-0.45	[-4.30, 2.70]	-0.07	.658
	Video-Game Experience	-0.80	1.90	-0.43	[-4.50, 2.90]	-0.06	.671
	VR Confidence	3.80	2.40	1.57	[-1.00, 8.60]	0.25	.123
Productive Time (seconds)	Prior VR Experience	-2.61	5.48	-0.48	[-13.34, 8.13]	-0.06	.636
	Video-Game Experience	3.23	5.66	0.57	[-7.86, 14.31]	0.07	.571
	VR Confidence	-8.99	7.40	-1.21	[-23.49, 5.52]	-0.16	.230
Unproductive Time (seconds)	Prior VR Experience	14.73	13.76	1.07	[-12.24, 41.70]	0.17	.289
	Video-Game Experience	-8.64	14.21	-0.61	[-36.50, 19.22]	-0.09	.546
	VR Confidence	-25.72	18.59	-1.38	[-62.17, 10.72]	-0.22	.172

Notes. Performance Accuracy is expressed as Percent-of-Maximum-Possible (POMP) on a 0-100 scale. Coefficients from Type III ANCOVA models. Standardized β represents effect size in standard deviation units.