FAIR LABS
FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

RESEARCH REPORT ◆ AI & HEALTH ◆ 2025

# Mixed Reality in Medical Training: Evidence Review

*Reviewing the evidence on volumetric and mixed-reality communication for remote medical procedure training—what it can show, what it has shown, and what remains unproven.*

PROGRAM
AI & Health

DOCUMENT TYPE
Research Report

AUDIENCE
Medical Educators & Health Systems

PUBLISHED
2025

FAIR Labs · A 501(c)(3) nonprofit research institute · Washington, DC

CONTENTS

# Table of Contents

EXECUTIVE SUMMARY

# What the evidence supports, and what it does not

IN ONE PARAGRAPH

Immersive simulation is a real and useful addition to the medical educator's toolkit, best understood as a maturing complement to—not a replacement for—cadaveric, bench, and supervised clinical training. The evidence most strongly supports its value for early procedural rehearsal, anatomical understanding, and repeatable deliberate practice, particularly where a learner can fail safely and often. The evidence is thinner, and should be read cautiously, for durable skill retention, transfer to live patients, and the newest frontier—volumetric telepresence for remote procedural mentoring—where enthusiasm currently outruns controlled data. This report maps that uneven landscape and proposes how to strengthen it.

**4**
levels of the Kirkpatrick hierarchy we use to grade what an outcome actually measures

**6**
technology families reviewed, from immersive VR to volumetric telepresence

**5**
recurring threats to validity that recur across the immersive-training literature

## What we conclude

1. **Grade every claim by what it measured.**
A result about learner satisfaction or a simulator score is not a result about patient outcomes. Most positive findings in this field sit at the lower levels of the training-evaluation hierarchy; read them as encouraging signals about learning, not proof of clinical benefit.

2. **Match the technology to the competency.**
Immersive media are not uniformly useful. They excel where spatial understanding, repetition, and safe failure matter most, and add least where the binding constraint is haptic fidelity or genuine tissue variability that current systems cannot reproduce.

**3** **Treat telepresence mentoring as promising but unproven.**
Volumetric and mixed-reality remote guidance addresses a real problem—expertise is unevenly distributed—but the controlled evidence that it improves procedural learning at a distance is early. Fund it as research, deploy it as a pilot, and measure it honestly.

**4** **Build the evaluation in from the start.**
The field's central weakness is methodological, not technological. Institutions adopting these tools should commit, in advance, to blinded assessment, adequate follow-up, and outcomes that reach beyond the simulator—because that evidence will not appear on its own.

The sections that follow develop each conclusion, describe the technologies and their evidence in detail, and set out a research agenda for moving the field from plausible to proven.

SECTION I

# Background: XR and Mixed Reality in Medical Education

Medicine has always taught procedural skill through a chain of imperfect substitutes for the patient. The cadaver stood in for the living body; the animal model stood in for human tissue; the plastic task-trainer stood in for the cadaver; the standardized patient stood in for the clinical encounter. Each substitute traded some fidelity for some combination of safety, availability, repeatability, and cost. Extended reality is the newest link in that chain, and it is best understood not as a break with this tradition but as its continuation by computational means. What is genuinely new is the axis along which the trade is made: immersive media can offer repeatability and safe failure at a scale no physical model matches, while still struggling to reproduce the one thing the cadaver never lacked—the feel of real tissue.

The pedagogical case for simulation of any kind rests on well-established foundations. Deliberate practice—focused, repetitive rehearsal of a skill with immediate feedback and progressive difficulty—is among the most robust findings in the science of expertise, and procedural medicine is a domain where it applies cleanly. Mastery learning, in which a trainee advances only upon demonstrating a defined standard rather than after a fixed period of time, has a credible track record in health-professions education. Simulation exists to make deliberate practice and mastery learning feasible for skills that are too dangerous, too rare, or too resource-intensive to rehearse on patients. Immersive technologies extend that logic: they promise the repetition of a task-trainer with richer context, the spatial realism of a cadaver without its scarcity, and the possibility of rehearsing rare emergencies on demand.

## A short history of the immersive turn

Virtual reality entered surgical education in earnest through minimally invasive surgery, where the operative field is already mediated by a screen and the instruments already move through a restricted set of degrees of freedom. Laparoscopy and endoscopy were, in a sense, waiting for the simulator: because the surgeon already operates by watching a monitor and manipulating long instruments, a virtual rendering of that same experience loses less in translation than an open procedure would. Early virtual laparoscopic trainers demonstrated that time on a simulator could improve measured performance, and this became one of the field's foundational results. Augmented and mixed reality arrived later, riding improvements in head-mounted displays, spatial tracking, and depth sensing, and opened a different possibility: not a fully synthetic world, but digital information—anatomy, guidance, a mentor's hands—registered onto the real one.

**A WORKING VOCABULARY**

Throughout this report, **extended reality (XR)** is the umbrella term for virtual, augmented, and mixed reality together. **Virtual reality (VR)** replaces the user's visual world entirely; **augmented reality (AR)** overlays digital content on the real world without necessarily anchoring it; and **mixed reality (MR)** anchors digital content to the real environment so that it behaves as if physically present. We reserve **volumetric** for capture and display of a person or scene as three-dimensional data that can be viewed from multiple angles, rather than as flat video.

## Why the enthusiasm, and why the caution

The enthusiasm is easy to understand and partly justified. Immersive simulation can compress the logistics of training—no operating room to book, no cadaver to source, no patient to consent—into a headset and a controller. It can present the same scenario identically to a hundred learners, which is a precondition for fair assessment. It can let a novice fail catastrophically and learn from it, which the clinical environment rightly forbids. And it can, in principle, capture rich data about how a learner moves, hesitates, and errs, opening the door to feedback that a human observer could not provide. These are real advantages, and they explain why medical schools, surgical colleges, and health systems have invested in the technology.

The caution is equally warranted and is the throughline of this report. Enthusiasm for a new educational technology tends to arrive before the evidence that it works, and to be reinforced by a novelty effect—learners engage more with something new, which inflates early measures of learning that fade as the novelty wears off. The history of educational technology is littered with tools that promised transformation and delivered a modest, context-dependent improvement, if that. Immersive media may well be different, but the burden of demonstrating so falls on evidence, not on the vividness of a demonstration. A striking simulation is a claim, not a proof.

> *The question is never whether a simulation is impressive. It is whether the learner who trained on it is better with a patient—and how we would know.*
>
> — The evaluative stance of this review

## What this report covers, and what it does not

Our scope is procedural and psychomotor training: the acquisition of skills a clinician performs with the hands, under time pressure and often under stress, where spatial understanding and fine-motor control are decisive. We touch on anatomical education because it underpins procedural competence, and on team and communication training because remote mentoring is inseparable from it. We do not attempt a comprehensive review of XR for patient-facing therapeutics—pain distraction, exposure therapy, rehabilitation—which is a distinct literature with its own evidence base.

And we are explicit that our subject is the *evidence*: not what the technology could someday do, but what studies to date actually establish, and with what confidence.

SECTION II

# The Technologies: VR, AR, MR, Volumetric Capture, Haptics, Telepresence

The word "immersive" hides a family of technologies with very different capabilities, costs, and evidentiary maturity. Grouping them together is convenient and misleading; the evidence for one says little about another.

Before weighing what immersive training can do, it is worth separating the technologies that the term conflates. Each occupies a distinct point in a trade-off space defined by how much of the real world it preserves, how faithfully it renders touch, how much it costs to deploy, and—crucially for this report—how much controlled evidence exists for its use in procedural training. A single number cannot capture that space, but a structured comparison can orient the reader before we turn to the evidence itself.

### THE IMMERSIVE TECHNOLOGY FAMILIES, COMPARED

| Technology | What it does | Best-fit training use | Chief limitation | Evidence maturity |
|---|---|---|---|---|
| **Immersive VR** | Replaces the visual world with a synthetic one | Rehearsing scenarios, spatial navigation, rare-event drills | Weak or absent haptics; simulator sickness | Moderate |
| **Augmented reality** | Overlays digital content on the real view | Guidance cues, anatomy overlays, checklists in situ | Registration accuracy; visual clutter | Emerging |
| **Mixed reality** | Anchors digital content spatially to the real world | Anatomy on a mannequin, remote mentor's hands in place | Field of view; tracking drift; cost | Early |
| **Volumetric capture** | Records a person or scene as 3-D data | Viewing an expert's motion from any angle | Bandwidth; capture rig cost; fidelity | Early / research |
| **Haptic devices** | Renders force and touch feedback | Needle insertion, drilling, tissue resistance | Fidelity ceiling; per-modality tuning | Moderate (task-specific) |
| **Telepresence** | Places a remote person in a shared space | Remote mentoring, tele-proctoring of procedures | Latency; presence; regulatory status | Early / research |

## Virtual reality and the problem of touch

Fully immersive VR is the most mature of the immersive modalities and, for many procedural tasks, the most limited in a specific and important way. VR excels at reproducing what the eyes see and the body navigates: a rendered operative field, an anatomical structure that can be walked around, an emergency that unfolds in a rehearsable sequence. Where it struggles is touch. Procedural medicine is saturated with haptic information—the give of a vessel wall, the pop of a needle through fascia, the changing resistance as a drill approaches a nerve—and rendering that information convincingly is extraordinarily difficult. A VR trainer with no haptics teaches the visual and cognitive components of a procedure while leaving its most distinctive channel of information unaddressed. This is not a fatal flaw; much of procedural skill is cognitive and visual. But it bounds what a headset alone can claim to teach.

## Augmented and mixed reality: keeping the real world in view

AR and MR take the opposite approach: rather than substituting for reality, they add to it. For procedural training this is often exactly what is wanted, because the learner still handles real instruments and real or physical-model tissue while receiving digital guidance—an anatomical overlay registered to a mannequin, a target trajectory for a needle, a checklist that floats in the corner of vision. Mixed reality's defining feature, spatial anchoring, is what makes it interesting for remote mentoring: a guidance cue or a mentor's rendered hand can be pinned to a location on the patient and stay there as the learner moves. The chief technical burdens are registration accuracy—digital and physical must align precisely, or the guidance misleads—and field of view, which in current head-mounted displays is narrower than natural vision and can crop away context.

## Volumetric capture and telepresence

The frontier this report is most concerned with combines two capabilities. **Volumetric capture** records a person—an expert's hands, a whole operator—as three-dimensional data that a remote viewer can inspect from any angle, rather than from the single fixed viewpoint of a camera. **Telepresence** then places that captured person into the learner's space as if present. Together they promise something video calling cannot: a remote mentor who can point *in three dimensions*, demonstrate a motion that the learner sees from the correct perspective, and share a spatial frame of reference with the trainee. The appeal for procedural mentoring is obvious, because so much of what an expert conveys at the bedside is spatial and gestural—"come in at this angle," "your hand should be here"—and is lost or degraded over a flat video link.

**THE LATENCY CONSTRAINT**

Remote procedural guidance lives or dies by latency. A mentor's correction that arrives a fraction of a second late is not merely annoying; for a moving needle or a live vessel it can be worse than no guidance at all. Any honest appraisal of telepresence mentoring must treat network latency, jitter, and reliability not as engineering footnotes but as first-order determinants of whether the approach is safe to use on a patient at all.

## Haptics as the crosscutting variable

Haptic feedback deserves separate mention because it cuts across the other categories and is often the true determinant of whether an immersive trainer teaches a manual skill or merely its visual shell. Haptic realism is highly task-specific: a device tuned to render the resistance of a lumbar puncture is not the same device that renders bone drilling, and building convincing force feedback for each procedure is a substantial engineering effort. Where haptics are good, the evidence for skill transfer strengthens; where they are absent or crude, the trainer's claims should shrink accordingly. A reader evaluating any immersive procedural trainer should ask, first, what the system does about touch.

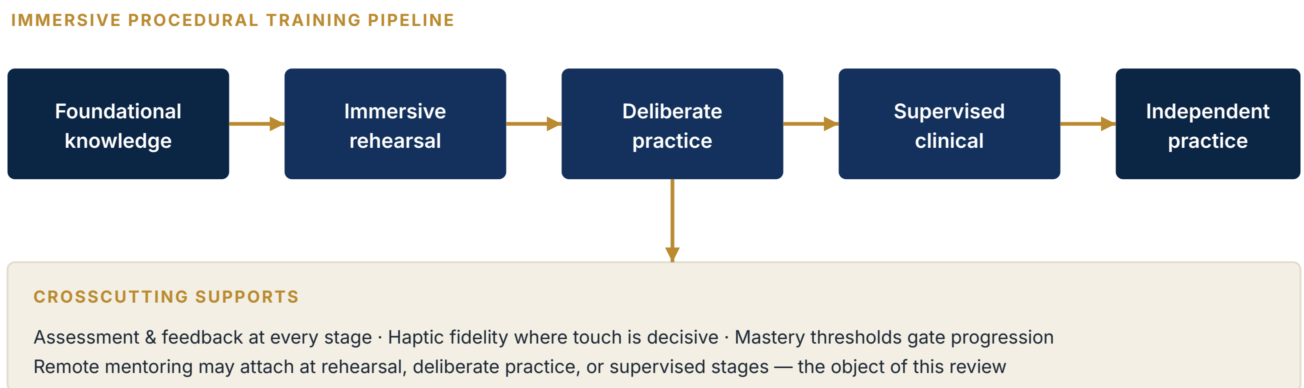


**Figure 1.** Immersive media occupy the middle of the procedural-training pipeline, between foundational knowledge and supervised clinical work. They augment deliberate practice; they do not remove the supervised clinical stage that precedes independent practice.

SECTION III

# Evidence on Learning Outcomes

To evaluate the evidence coherently, we need a way to grade what any given study actually measured. The most durable framework for this is the Kirkpatrick hierarchy of training evaluation, which distinguishes four levels of outcome: reaction, learning, behavior, and results. A learner who enjoyed a VR module has demonstrated a reaction. One who scored better on a knowledge test or a simulator task has demonstrated learning. One who performs better in the actual clinical workplace has demonstrated a change in behavior. And one whose patients fare better has demonstrated results. These levels are not interchangeable, and the single most common error in reading this literature is to treat a lower-level finding as if it settled a higher-level question.

THE KIRKPATRICK LEVELS APPLIED TO IMMERSIVE MEDICAL TRAINING

| Level | What it measures | Typical immersive-training evidence | Strength to date |
|---|---|---|---|
| **1 · Reaction** | Learner engagement and satisfaction | Surveys of enjoyment, perceived usefulness, presence | Consistently positive — but weakest evidence of benefit |
| **2 · Learning** | Knowledge and skill gained | Knowledge tests; simulator metrics; skills-lab checklists | Substantial and generally favorable |
| **3 · Behavior** | Performance in the clinical workplace | Rated performance on real procedures; time to competence | Limited; fewer, smaller studies |
| **4 · Results** | Patient and system outcomes | Complication rates; error rates; efficiency | Sparse; rarely powered to detect it |

## Skill acquisition: the strongest ground

The evidence is at its most persuasive at the level of skill acquisition. Across a range of procedural domains, learners who train on immersive simulators tend to improve on the metrics those simulators capture, and often outperform learners who received no simulation or only conventional instruction on comparable skills-lab assessments. This is a real and useful finding: it means that immersive practice can move a novice up the early part of a learning curve efficiently, safely, and repeatably, which is precisely the region where practicing on patients is least appropriate. The laparoscopic literature, the oldest and deepest, is the clearest example, but similar patterns appear for endoscopy, certain orthopedic and neurosurgical tasks, and anatomical understanding.

Two qualifications keep this finding in proportion. First, the comparison group matters enormously. Immersive training compared against *no* training will almost always look good; compared against an equal dose of conventional simulation or bench practice, the advantage often shrinks or disappears.

The interesting question is rarely "is it better than nothing" but "is it better than the next-best alternative for the same cost." Second, the outcome measure matters. When the outcome is a metric internal to the simulator, a positive result partly reflects learners becoming better at the simulator, which is not the same as becoming better clinicians.

THE TRANSFER QUESTION

The pivotal concept in this literature is **transfer**: whether skill built in the simulator carries over to performance on a real task. A trainer that improves simulator scores but not real performance has taught the simulator, not the procedure. Transfer studies—especially transfer to live clinical performance—are the evidence that matters most and, unsurprisingly, the evidence that is hardest and most expensive to produce.

## Retention: better than assumed, but under-measured

Skill, once acquired, decays. How fast immersive-trained skill decays, and whether it decays more or less than skill acquired by other means, is a question the literature answers only partially. Many studies assess learners immediately after training, when performance is at its peak and novelty is freshest, and never return. The subset that includes delayed reassessment generally finds that simulator-acquired skill persists at least in part over weeks to months, and that refresher practice restores it efficiently—a genuinely encouraging pattern, since one of simulation's promises is low-cost maintenance of perishable skills. But retention findings should be read with the caveat that the studies powered to detect meaningful decay, over clinically relevant intervals, remain relatively few.

## Transfer to clinical performance: the decisive gap

The evidence thins markedly as we move from the simulator to the patient. A modest number of well-designed studies have shown that trainees who first learned on a simulator subsequently perform better on early real procedures than peers who did not—the transfer result that gives the whole enterprise its point. These studies are the field's most valuable evidence precisely because they are the hardest to run: they require access to real clinical performance, blinded expert rating, and enough participants to detect a difference that time and supervision will eventually erase for everyone. Their scarcity is not evidence that transfer does not occur; it is evidence that the field has not yet measured it at the scale the claims require.

For augmented, mixed-reality, and volumetric modalities specifically, the transfer evidence is younger and sparser still. Much of it consists of feasibility studies, single-arm demonstrations, and small comparative trials with the methodological features—short follow-up, unblinded assessment, enthusiastic developers as evaluators—that warrant caution. This does not mean these modalities do not work. It means that, at present, the honest summary is "promising and under-tested," and that anyone claiming more is reaching beyond the data.

> *A simulator score is a hypothesis about a clinician. Transfer studies are how the hypothesis is tested; without them, the score is just a score.*
>
> — On reading procedural simulation research

SECTION IV

# Remote Training & Telementoring

The strongest argument for immersive remote mentoring is not technological novelty. It is that procedural expertise is scarce and unevenly distributed, and the people who most need guidance are often the farthest from it.

Consider the underlying problem the technology is meant to solve. A clinician in a rural hospital, a low-resource setting, a military forward position, or simply a busy department without a subspecialist on hand may face a procedure they perform rarely and would perform better with an expert beside them. Telementoring—an expert guiding a procedure remotely—predates immersive media and has an established, if modest, evidence base using ordinary video and audio. The immersive proposition is that volumetric and mixed-reality communication can restore the spatial richness that flat video strips away, closing the gap between "an expert on a screen" and "an expert in the room." This is the specific frontier connected to the NSF-funded research on volumetric and mixed-reality communication for remote medical procedure training that informs this report.

## What volumetric communication adds, in principle

The theoretical advantages are concrete and worth stating precisely, because they are what any evaluation must eventually test. First, **shared spatial reference**: a mentor can indicate a location or a trajectory in three dimensions, anchored to the patient, rather than gesturing at a two-dimensional image and hoping the trainee maps it correctly. Second, **perspective independence**: because the scene is captured volumetrically, the mentor can view the field from the angle they need rather than from wherever a camera happens to sit. Third, **gestural demonstration**: an expert's hand motion, the most natural vocabulary of procedural teaching, can be shown and seen from the learner's own viewpoint. Each of these addresses a documented failure mode of conventional telementoring, in which spatial instructions are misunderstood and the mentor cannot see what they need to see.

MODES OF REMOTE PROCEDURAL GUIDANCE, FROM ESTABLISHED TO EXPERIMENTAL

| Mode | Spatial fidelity | Evidence base | Principal risk |
|---|---|---|---|
| **Audio-only guidance** | None | Long-standing, limited | Ambiguity; no shared view |
| **Two-way video** | Low (2-D, fixed view) | Established, moderate | Perspective mismatch |
| **AR annotation on video** | Moderate (marks on a flat field) | Emerging | Registration; latency |
| **Mixed-reality telementoring** | High (anchored 3-D cues) | Early / research | Latency; hardware; presence |
| **Volumetric telepresence** | Highest (captured 3-D expert) | Research / feasibility | Bandwidth; latency; unproven benefit |

## What the evidence currently shows

The candid summary is that the evidence for immersive remote mentoring is at an early, largely feasibility stage. Studies to date more often establish that a system *can* be built and used than that it *improves learning or outcomes* relative to a fair comparator. Feasibility and usability findings are genuinely useful—they establish that the approach is not merely a rendering, and they surface the practical obstacles of latency, hardware ergonomics, and workflow integration. But feasibility is a low bar for a claim about training effectiveness, and the field should resist the temptation to present a successful demonstration as a demonstrated benefit. Where comparative studies exist, they are typically small, short, and conducted by the developers of the systems they evaluate, a configuration that invites optimism.

A CAUTION SPECIFIC TO REMOTE MENTORING

Remote procedural guidance is not only an educational intervention; when it guides a live procedure, it is a clinical one. The evidentiary bar therefore rises: it is not enough to show that learners like it or perform better in a lab. A system used to mentor a real procedure at a distance must be evaluated for patient safety, for its behavior when the network degrades, and for who is accountable when guidance and outcome diverge. Educational promise does not confer clinical license.

## The human factors that will decide it

Whether immersive telementoring succeeds will likely turn less on rendering fidelity than on human factors that are easy to underestimate. Trust between mentor and trainee, built over a link rather than in person, shapes whether guidance is followed appropriately—neither ignored nor deferred to blindly. Communication protocols matter: who speaks, who acts, and how a disagreement is resolved in real time. Cognitive load matters: an immersive overlay that crowds the learner's field of view or demands attention at the wrong moment can degrade performance even as it adds information. These are the variables that distinguish a system that helps from one that merely impresses, and they are under-studied relative to the hardware.

VOLUMETRIC TELEMENTORING — SCHEMATIC

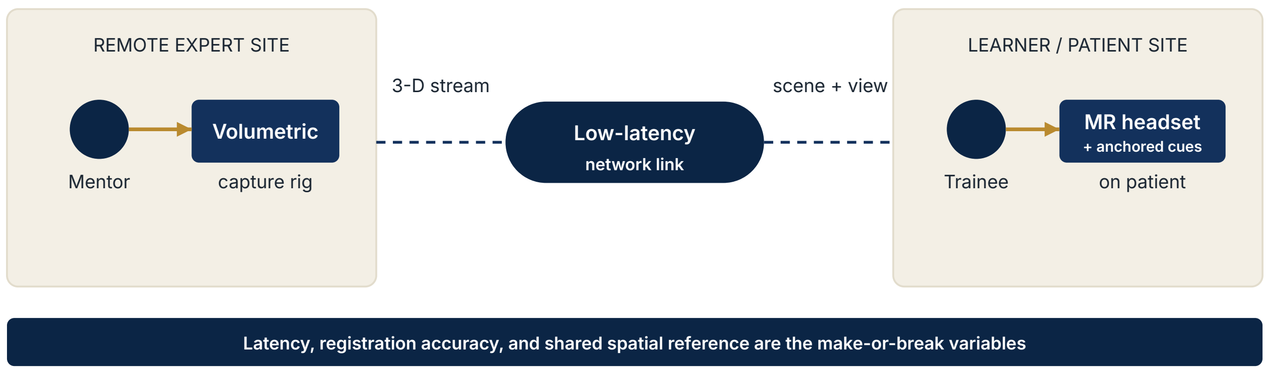


**Figure 2.** A schematic volumetric telementoring setup. The expert is captured in three dimensions and rendered into the learner's mixed-reality view, anchored to the patient. The dashed path—the network link—is where most of the practical and evidentiary risk concentrates.

SECTION V

# Strength & Quality of the Evidence

A review that only tallied positive findings would mislead. The more important question is how much confidence each finding can bear, and the answer depends on how the studies were done.

Medical education research is difficult to conduct well, and immersive-training research inherits all of its usual challenges while adding a few of its own. The interventions are hard to blind—a learner knows whether they are wearing a headset—the outcomes are hard to standardize, the samples are typically small cohorts of trainees at a single institution, and the technology changes faster than the studies evaluating it can be completed. None of this makes the evidence worthless. It makes the evidence something that must be read with an eye to specific, recurring threats to validity rather than taken at face value.

## Five recurring threats to validity

**A** **Small, single-site samples.**
Many studies enroll a few dozen trainees at one program. Such studies are underpowered to detect anything but large effects, vulnerable to the idiosyncrasies of one setting, and prone to false negatives as readily as false positives.

**B** **Weak or absent blinding.**
Participants usually cannot be blinded to the intervention, but assessors often can be—and frequently are not. Unblinded rating of a subjective performance measure is an open door for expectation to shape the result.

**C** **Surrogate outcomes.**
Simulator-internal metrics stand in for clinical performance because they are cheap to collect. They are legitimate as intermediate measures but should never be read as the clinical outcome they approximate.

D **Short follow-up and novelty.**
Assessment immediately after training captures peak performance inflated by engagement with a novel tool. Without delayed reassessment, retention and the fading of novelty are simply not measured.

E **Conflicted evaluation and publication bias.**
Systems are often evaluated by the teams that built them, and negative or null results are less likely to be published. Both forces push the visible literature toward optimism.

## What good evidence looks like

Against these threats, the features of a trustworthy study are easy to name and hard to deliver: a randomized design where feasible; an active comparator receiving an equal dose of the next-best training, not a do-nothing control; blinded assessment by raters with no stake in the result; validated outcome measures that reach as high up the Kirkpatrick hierarchy as the question requires; adequate statistical power; delayed reassessment to capture retention; and, ideally, evaluation by a party independent of the developers. Studies meeting most of these criteria exist, and they are the backbone of whatever confident claims the field can make. They are simply outnumbered by studies that do not.

**A STUDY-QUALITY APPRAISAL GRID FOR IMMERSIVE-TRAINING RESEARCH**

| Appraisal criterion | Stronger study | Weaker study |
|---|---|---|
| **Design** | Randomized or well-matched comparative | Single-arm or convenience comparison |
| **Comparator** | Equal-dose active alternative | No training / unmatched dose |
| **Assessment** | Blinded, validated instrument | Unblinded, ad-hoc rating |
| **Outcome level** | Behavior or results (Kirkpatrick 3–4) | Reaction or simulator score only (1–2) |
| **Follow-up** | Delayed reassessment for retention | Immediate post-test only |
| **Independence** | Evaluators separate from developers | Developer-run evaluation |
| **Reporting** | Pre-registered; nulls reported | Selective, post-hoc reporting |

**ON SYSTEMATIC REVIEWS AND META-ANALYSIS**

Systematic reviews of immersive medical training exist and are valuable, but they inherit the heterogeneity of the studies they pool: different technologies, procedures, outcomes, and comparators are aggregated under one banner. A pooled effect across such diversity can obscure as much as it reveals. The reader should weight a well-conducted review's *description of the evidence's limitations* at least as heavily as its headline effect.

## Reproducibility and the moving target

A structural difficulty deserves emphasis: the technology evolves faster than rigorous evaluation can keep pace. By the time a well-designed trial of a particular headset or system reports, the system has often been superseded, and the finding applies to hardware no one uses. This does not excuse weak studies, but it does explain a chronic mismatch between the pace of development and the pace of proof, and it argues for evaluating *capabilities and mechanisms*—does anchored 3-D guidance improve spatial understanding?—that outlast any particular device, rather than only specific products.

SECTION VI

# Safety, Ethics & Patient Considerations

A training technology is not ethically neutral simply because it operates on simulations rather than patients. Immersive medical training raises questions of learner safety, patient safety when the training touches live care, data governance, equity of access, and the honest representation of what a trainee has actually learned. These are not peripheral concerns to be handled after the technology proves effective; they shape whether it should be deployed at all, and how.

## Learner safety and simulator sickness

The most immediate safety consideration is the learner's own well-being. Immersive VR can induce simulator sickness—nausea, disorientation, eye strain—in a meaningful fraction of users, with effects that can persist after the headset comes off. For a training modality, this matters in two ways: it can degrade the learning experience and exclude susceptible learners, and it raises a practical safety question if a trainee moves to a real task while still affected. Prolonged headset use also raises ergonomic and visual-fatigue concerns that are still being characterized. These effects are usually manageable, but they are real, unevenly distributed across individuals, and a legitimate reason to design training with breaks, monitoring, and alternatives for those who cannot tolerate immersion.

## Patient safety when training meets live care

The ethical stakes rise sharply when immersive technology crosses from pure simulation into live care, as remote telementoring does by definition. Guiding a real procedure at a distance introduces failure modes that a skills lab does not: a network that degrades at the critical moment, guidance that is misregistered or misunderstood, an overlay that distracts at the wrong instant, and a diffusion of responsibility between the hands on the patient and the expert on the link. Any deployment that guides live procedures must treat these as clinical risks with named mitigations, must define what happens when the technology fails mid-procedure, and must be clear about who holds ultimate responsibility for the patient—which, in almost every defensible arrangement, is the clinician physically present.

**THE ACCOUNTABILITY QUESTION**

When a remote mentor guides a procedure and the outcome is poor, who is answerable—the hands on the patient, the voice on the link, or the institution that connected them? This is not a technicality to be resolved after adoption. It is a precondition for responsible use, and it must be settled before the first patient, not litigated after an adverse event.

## Data, privacy, and the captured procedure

Immersive and volumetric systems are, among other things, sophisticated capture devices. They can record a learner's every motion, a patient's anatomy in three dimensions, and the full spatial context of a procedure. This generates rich data with obvious educational value and equally obvious governance obligations. Trainee performance data can become a surveillance record and a basis for consequential judgments; patient-derived volumetric data is health information with all the sensitivity that entails. Who owns these recordings, how long they are retained, whether they can be used to train algorithms, and how patient consent is obtained for capture during care are questions that must be answered before capture begins, not after.

## Equity and honest representation

Two further considerations round out the ethical picture. The first is equity: immersive systems are expensive, and a technology promoted as a way to bring expertise to under-resourced settings could, if access tracks wealth, instead widen the gap between well-funded and struggling institutions. Whether telementoring narrows or widens disparities depends on how it is deployed and funded, not on the technology alone. The second is honesty about competence. A trainee who has mastered a simulator has demonstrated something valuable but bounded; presenting simulator proficiency as clinical readiness would be a disservice to the trainee and a hazard to patients. Immersive training earns its place as a step in a longer path to competence, and the ethical use of it depends on not overstating what that step establishes.

> *The right question is never "is the simulation realistic?" but "does the trainee, and everyone relying on them, correctly understand what the simulation has and has not taught?"*
>
> — On the ethics of representing competence

SECTION VII

# Implementation in Institutions

Between a promising study and a working program lies a set of decidedly unglamorous questions—cost, integration, faculty time, and equity—that determine whether a technology delivers value or gathers dust in a storage room.

Institutions considering immersive training tend to focus on the headset and underestimate everything around it. The recurring lesson from health-professions education technology is that the device is the cheap part. The durable costs and the determinants of success lie in curriculum integration, faculty development, maintenance, and the honest matching of a tool to a genuine educational need. A technology adopted because it is impressive, rather than because it solves a defined problem better than the alternative, is the most expensive kind of failure.

## The true cost, and the total cost of ownership

The purchase price of hardware is a fraction of the lifetime cost. Content must be acquired or built, and high-quality procedural content is expensive to produce and to keep current. Devices must be maintained, cleaned to clinical standards where they touch the training environment, updated, and eventually replaced on a cycle shorter than most institutional planning assumes. Technical support must exist, or the equipment fails in ways faculty cannot fix and abandon. And the largest hidden cost is faculty time—to learn the system, to integrate it into a curriculum, to supervise its use, and to assess the learning it produces. An honest business case counts all of this, not the sticker price.

COST AND INTEGRATION FACTORS BEYOND THE HARDWARE

| Factor | Why it is underestimated | What it demands |
|---|---|---|
| **Content** | Assumed to come with the device | Procedure-specific development and updates |
| **Faculty development** | Treated as a one-time briefing | Ongoing training and curricular integration |
| **Maintenance & support** | Invisible until it fails | Staff, spares, update and hygiene cycles |
| **Curriculum integration** | Confused with mere availability | Explicit placement in objectives and assessment |
| **Assessment linkage** | Deferred indefinitely | Defined competencies and mastery thresholds |
| **Refresh cycle** | Planned like durable equipment | Budgeting for short hardware lifespans |

## Integration is the hard part

A simulator that sits outside the curriculum is a novelty; one woven into it is a tool. Integration means deciding where in a program the technology belongs, what objective it serves, how a learner's progress on it counts, and what mastery standard gates advancement. It means faculty who can teach with it, not merely operate it, and assessment that connects simulator performance to the competencies the program actually certifies. This is curricular and cultural work, and it is where most implementations succeed or fail—long after the procurement decision that gets all the attention.

**A BLUNT IMPLEMENTATION HEURISTIC**

Before acquiring an immersive system, an institution should be able to state the specific competency it will improve, the alternative it is chosen over, how learning will be measured, and who will maintain and teach with it. An adoption that cannot answer these four questions is buying a demonstration, not a capability.

## Faculty development and the risk of deskilling

Immersive tools change the educator's role as much as the learner's. Faculty need time and support to become fluent, and programs need to guard against a subtler risk: that reliance on simulation erodes the traditional apprenticeship channels through which tacit skill has always passed. The goal is augmentation—freeing scarce expert time for the parts of teaching that only a human can do—not the quiet substitution of a headset for the mentorship that builds judgment. Whether a program achieves augmentation or substitution is a choice, and it should be made deliberately.

## Equity of access within and between institutions

Finally, implementation has a distributive dimension. Within an institution, immersive training should not become a privilege of the programs that can fund it while others make do; across institutions, the promise of democratizing expertise is real but conditional. Shared regional facilities, pooled content, and deliberate subsidy can spread access; unmanaged market adoption will tend to concentrate it. An institution that cares about the equity argument for this technology must build equity into how it deploys it, because the technology will not deliver that outcome on its own.

SECTION VIII

# A Research Agenda & Recommendations

The field's central weakness is not its technology but its evidence. Closing that gap is a tractable, if unglamorous, program of work—and it is the work most worth funding.

Our review returns, at every turn, to the same diagnosis: capability has outrun proof. The technologies are real and improving; the demonstrations are compelling; the enthusiasm is understandable. What is missing is the body of rigorous, higher-level, independently conducted evidence that would let institutions and educators know not merely that immersive training *can* work but *when*, *for whom*, and *compared to what*. The recommendations below are aimed at producing that evidence and at deploying the technology responsibly in the meantime.

## Priorities for research

1. **Climb the Kirkpatrick hierarchy.**
   Fund and design studies that measure behavior in the clinical workplace and, where possible, patient results—not only reactions and simulator scores. The higher-level evidence is scarce precisely because it is hard and expensive, which is exactly why it needs deliberate support.

2. **Prioritize transfer and retention.**
   Make transfer to real performance, and retention over clinically meaningful intervals, the default outcomes rather than afterthoughts. A study that stops at the immediate post-test has left the most important questions unanswered.

3. **Test mechanisms, not just products.**
   Because devices are superseded faster than they can be evaluated, study the durable mechanisms—anchored 3-D guidance, shared spatial reference, haptic fidelity—that outlast any single headset and generalize across the technology's future forms.

4. **Build independence and pre-registration into the norm.**
   Evaluations conducted by parties without a stake in the outcome, pre-registered and reporting null results, are the antidote to the optimism that developer-run studies invite. The field's credibility depends on it.

## Priorities for the telepresence frontier specifically

**A** **Characterize the latency envelope.**
Establish, empirically, the network conditions under which volumetric telementoring is safe and useful, and the conditions under which it is not. This is foundational and currently unsettled.

**B** **Study the human factors.**
Investigate trust, communication protocols, cognitive load, and role clarity between remote mentor and local trainee. These may matter more than rendering fidelity and are badly under-studied.

**C** **Deploy as monitored pilots, not products.**
Where immersive remote mentoring touches live care, deploy it as a carefully evaluated pilot with patient-safety oversight and honest measurement, not as a finished clinical service.

**THE BOTTOM LINE**

Mixed reality earns a real, if bounded, place in medical training today: strongest for early procedural rehearsal and spatial understanding, weaker for durable transfer, and earliest of all for volumetric remote mentoring. The responsible posture is neither dismissal nor hype. It is to adopt what the evidence supports, to pilot what it does not yet support under honest measurement, and to fund the studies that will tell us which is which.

## What good looks like

An institution engaging this technology well is not one that has bought the most headsets. It is one that can say, for any immersive tool it uses, what competency it improves, what evidence supports that claim and at what Kirkpatrick level, how it will know whether the learning transfers, and how it will measure the tool honestly over time. It treats impressive demonstrations as hypotheses, holds telepresence mentoring to a clinical bar when it touches patients, and contributes to—rather than only consuming—the evidence base. That posture will do more for patients than any particular device, because it is the posture that turns a promising technology into a proven one.

The promise of volumetric and mixed-reality communication for remote medical training is genuine, and the problem it addresses—the uneven distribution of procedural expertise—is one worth solving. This report's argument is not that the promise is empty, but that it is unredeemed, and that redeeming it is a matter of evidence and discipline as much as engineering. The technology has earned the right to be taken seriously. It has not yet earned the right to be taken on faith.

APPENDIX

# Notes & Further Reading

This review synthesizes established work in medical education, simulation science, and extended-reality research. The sources below are entry points into the literature that informs it; they are indicative rather than exhaustive, and are cited for their conceptual role in the argument rather than as the source of any specific numerical claim.

1. Kirkpatrick, D. L., & Kirkpatrick, J. D. *Evaluating Training Programs: The Four Levels.* The foundational hierarchy—reaction, learning, behavior, results—used throughout this report to grade what an outcome measures.
2. Ericsson, K. A. *Deliberate Practice and Acquisition of Expert Performance.* On focused, feedback-rich repetition as the engine of expertise, including in medicine.
3. McGaghie, W. C., et al. *A Critical Review of Simulation-Based Medical Education Research.* On effectiveness, mastery learning, and the methodological standards this literature should meet.
4. Cook, D. A., et al. *Technology-Enhanced Simulation for Health Professions Education: A Systematic Review and Meta-Analysis.* A landmark synthesis of simulation's effects and its heterogeneity.
5. Issenberg, S. B., et al. *Features and Uses of High-Fidelity Medical Simulations that Lead to Effective Learning.* On the design features—feedback, repetition, integration—that make simulation work.
6. Seymour, N. E., et al. *Virtual Reality Training Improves Operating Room Performance.* An early and influential demonstration of transfer from a virtual laparoscopic trainer to real performance.
7. Gallagher, A. G., et al. *Virtual Reality Simulation for the Operating Room: Proficiency-Based Training.* On proficiency-based progression and the science of surgical simulation.
8. Milgram, P., & Kishino, F. *A Taxonomy of Mixed Reality Visual Displays.* The reality–virtuality continuum that underlies the VR/AR/MR vocabulary used here.
9. Barsom, E. Z., et al. *Systematic Review on the Effectiveness of Augmented Reality Applications in Medical Training.* On the emerging and uneven evidence for AR in education.
10. Slater, M., & Sanchez-Vives, M. V. *Enhancing Our Lives with Immersive Virtual Reality.* On presence, immersion, and the human experience of virtual environments.
11. Ericsson, K. A. *Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine.* On skill maintenance, decay, and refresher practice relevant to retention.
12. Augestad, K. M., & Lindsetmo, R. O. *Overcoming Distance: Video-Assisted Surgical Telementoring.* On the established, pre-immersive evidence base and challenges of remote procedural mentoring.

 It draws on the authors' NSF-funded research on volumetric and mixed-reality communication for remote medical procedure training.