

FIELDWORK NO. 044 • REHEARSAL AND TRANSFER

When Does AI Practice Actually Make Someone Better?

Practice is about to get very cheap. We looked at what happened historically when people gained the ability to rehearse something consequential before doing it for real, to work out which kinds of practice change the outcome and which kinds only change how ready someone feels.

13

CASES EXAMINED

1934—
2026

PERIOD COVERED

12

AI TRIALS REVIEWED

In the late 1970s, researchers ran one of the largest education experiments ever attempted in the United States. About 18,000 high school students in DeKalb County, Georgia were randomly assigned to different driver education programs.

One group got a seventy-hour course. Classroom work, simulation, closed-range practice, real driving. It was designed off a detailed analysis of what drivers actually have to do. It was the best program anyone knew how to build.

Another group got a minimal course. A third got nothing.

The students in the intensive program were more likely to get licensed. They were also more likely to crash and more likely to be cited than the students who got no training at all.

The explanation is not that the training made them worse drivers. It got them onto the road sooner. More practice moved a real behavior. It moved the outcome in the wrong direction.

18,000

students randomly assigned

70 hrs

of instruction in the intensive program

More

crashes and violations than the untrained group

That result is worth holding onto right now, because the cost of practice is collapsing.

AI can generate an interview, a sales call, a presentation, a negotiation, a difficult conversation with an employee, a clinical scenario, at close to zero marginal cost. Every product team in reach of this is being asked to add an AI coach.

The assumption underneath all of it is that more realistic practice makes people better. History says that assumption is too simple.

THE PATTERN

The most realistic simulation had the weakest evidence. The less realistic one transferred.

Early laparoscopic surgeons had a specific problem. They had to move instruments inside a patient while watching their own hands indirectly, on a flat screen, with the movement reversed. Everything about the coordination is wrong until you have done it enough times.

Someone built a virtual reality trainer for exactly that. It reproduced almost nothing about an operating room. No patient. No bleeding. No team. No consequence if it went badly. Just the hand movement, over and over, with feedback.

In a **randomized, blinded study**, residents who trained on it were compared with residents who did not, on real gallbladder surgery, scored by reviewers who did not know who had trained. The untrained group made roughly three times as many errors and took about 58 percent longer.

Now take the opposite case.

A paper trading account reproduces almost everything. Real prices. Real securities. Real order types. Real market movement in real time. On the screen it is close to indistinguishable from the real thing. The one thing missing is that the money is not yours.

We looked for independent controlled evidence that paper trading performance carries into live trading performance, across academic databases, brokerage disclosures and trade press. We did not find any. That is not the same as saying paper trading has been proven useless. It means the evidence people assume exists does not appear to.

REPRODUCED VERY LITTLE	REPRODUCED ALMOST EVERYTHING
Surgical VR trainer	Paper trading account
No patient	Real prices
No team	Real securities
No consequence	Real order types

Reproduced the hand-eye coordination	Removed the money at risk
Transfer measured under randomization.	No independent transfer evidence found.

What matters is not how much the practice looks like the real thing. What matters is whether it contains the thing that makes the real thing hard.

THE RULE, AND A CORRECTION TO IT

Practice works when it contains the reason you would otherwise fail

The useful version of this is short. Every consequential moment has one thing that is actually causing failure. For a junior surgeon it is hand control. For a retail trader it is what happens to your judgment when the money is real. Call that the binding constraint.

Practice is more likely to transfer when it contains the thing actually limiting that person. Realism everywhere else is a poor substitute.

That was where this research landed first. Then a set of studies turned up that made the rule better.

The interview case

Virtual Reality Job Interview Training has been through five randomized controlled trials. It is not a confidence product. The trials measured whether people got jobs.

The participants were people with severe mental illness, autism, PTSD, substance use histories, and people leaving prison. In one follow-up, about 51 percent of trainees received a job offer against 25 percent of the comparison group. A later analysis found that the improvement in interview skill carried the effect.

This should have been hard. A job interview is a consequential human interaction. The simulation removes the actual employer and the actual judgment. The prediction would have been weak transfer.

Transfer happened anyway. The likely reason is that for these participants, the thing causing failure was often mechanical rather than emotional. How to structure an answer. How to talk about a two-year gap or a conviction. How to hold the ordinary rhythm of a professional conversation. Those can be practiced, and the simulation contains them.

That is a real correction and it changes the rule:

Practice works when it contains the binding constraint for the person doing it. The same activity can be hard for different people for completely different reasons.

What this does not establish. None of these trials studied a general job seeker who already interviews competently. For that person the limiting factor may well be the thing the simulation removes. The finding is real and its population is specific.

THE MEASUREMENT TRAP

Confidence is not performance, and behavior is not performance either

There are four separate things a practice product can move, and moving one does not move the next.

WHAT PRACTICE CAN CHANGE

01	Confidence	How ready the person feels. Moves almost always.
02	Behavior	What they actually do differently in the real situation.
03	Performance	How well they do it, judged by someone other than the product.
04	Outcome	Whether the thing you cared about happened. The offer, the sale, the recovery.

Many products measure the first level while the buyer assumes they are buying the fourth.

Public speaking gives the cleanest illustration. A randomized trial of a virtual reality exposure program found large reductions in anxiety and in gaze avoidance during later real speeches, and no improvement in speech quality as rated by an outside committee.

Real behavior changed. It was measurable, in a real speech, by independent raters. Performance did not move.

CPR training runs the same split faster. In one study of high school students, the share who said they felt confident performing CPR went from about 27 percent before training to about 99 percent after. Measured skills decay within about three months, and knowledge was not associated with skill level at all.

None of this means rehearsal is useless. It means a product can produce excellent internal numbers while the thing you were trying to change sits exactly where it was.

The opportunity is real and it is uneven

Practice used to require instructors, facilities, actors, or expensive purpose-built software. AI removes most of that cost. What it does not do is change which kinds of practice work.

STRONGEST EVIDENCE

Procedural and technical skills

Clinical procedures · surgical technique · anything with an observable correct action

A [meta-analysis of twelve randomized trials](#), searched through April 2026, compared AI-assisted clinical simulation against traditional instruction. AI significantly improved surgical and invasive procedural skills. It showed no advantage for clinical assessment and diagnostic skills. Vision-based systems beat traditional training. Large language models did not.

In plain terms: AI worked when it could see the action and correct it. It was much less impressive when the task needed judgment under uncertainty.

Worth knowing before anyone quotes this: every one of those trials measured training performance. None of them measured patient outcomes.

PROMISING, UNPROVEN

Sales, customer service, language, general interview prep

AI role-play platforms · conversation practice · call rehearsal

There is real commercial activity here. Companies are paying somewhere in the range of sixty to a hundred and twenty dollars per seat per month for AI sales role-play, and the products are credible.

What is missing is independent evidence. We searched for studies showing that people assigned to AI sales rehearsal go on to produce higher win rates or more revenue than comparable people who were not, and did not find them. Vendor evidence pages in this category tend to cite research on human role-play, or figures about learners feeling more confident, neither of which is evidence about this.

This is unproven, which is a different thing from failed. The data to settle it already sits inside these companies' customers.

HIGHER TRANSFER RISK

Consequential human interactions where the product removes the consequence

Negotiation · difficult workplace conversations · confidence-led speaking products

For some people, the difficult part of these moments is that the other person's judgment persists after the conversation ends. That is exactly the part the simulation removes.

The risk is a user who becomes comfortable inside the practice while the hard part of the real event stays outside it. The public speaking result is the

warning shot: a well-built intervention moved real behavior and left rated performance flat.

Higher risk is not the same as cannot work. The interview trials are proof that some human interactions do transfer, when the product happens to contain the thing that is actually stopping that particular person.

THE BUSINESS QUESTION

Whoever runs the evaluation usually decides whether there is a market

Separate from whether practice works is whether anyone can build a durable business selling it. Across the historical cases, the outcome was set less by the quality of the practice than by what the evaluating institution did.

01 **The evaluator gives practice away**

The independent market gets much harder. The College Board, which owns the SAT, [put free official practice online with Khan Academy in 2015](#). A current directional example: Google retired its standalone Interview Warmup tool in 2026 and now points job seekers toward its broader assistant. The sourcing on that is competitor reporting rather than a Google statement, so treat it as a direction rather than a proven cause.

02 **The evaluator changes the test**

The market usually moves rather than dies. When [USMLE Step 1 went pass/fail in 2022](#), preparation shifted to Step 2. Technical interview prep has adapted repeatedly as employers changed formats, and it is still growing.

03 **The evaluator requires practice**

This is the one that builds something large. Simulator hours are written into pilot certification, which turned aviation rehearsal from an optional product into required infrastructure and a [multibillion dollar market](#). Nobody in the historical record built a big rehearsal business by outrunning the evaluator. They built it by being required.

04 **The evaluator cannot credibly coach the people it judges**

That leaves an independent market standing. Bar examiners cannot become every candidate’s personal tutor without a conflict, and roughly sixty years later [one provider still holds around 60 percent of bar preparation](#). This is a more defensible position than a role-play tool anyone can rebuild.

GENERAL ASSISTANTS

The simulation is going to be cheap. The system around it is where value sits.

If the core of a product is “talk to an AI person and practice a conversation,” assume a general assistant will reproduce most of that experience, and reasonably soon.

That does not mean every standalone company disappears. It does mean the conversation is the wrong thing to own. Once the conversation becomes cheap, value has to move somewhere the assistant cannot easily reproduce: the record of real performance, an institutional requirement, or the workflow around the practice.

The AI sales platforms are the visible case. They do not really sell the role-play. They sell a scorecard applied to real recorded calls and a certification gate a revenue leader can defend in a meeting. A general assistant can produce an objection-handling role-play this afternoon. It cannot produce the record.

The historical version of the same thing is aviation. Simulator hours are required, so the practice is not competing for attention against a free alternative. It is competing to be the thing that satisfies the requirement.

THE TEST

Three questions before building or funding any of this

BEFORE YOU BUILD THE REHEARSAL

Question one

What actually causes this person to fail when the moment counts?

Not what makes the task complicated. What makes them fall over. If you cannot answer it specifically, for a specific kind of user, you are not ready to build the practice yet.

Question two

Does the practice contain that thing?

If it does not, adding realism everywhere else may not help. The surgical trainer got this right by reproducing one thing and ignoring the rest. Paper trading gets it wrong by reproducing everything except the part that matters.

Question three

What changed outside the product?

Usage, completion rates, confidence scores and improvement inside the simulation are all measures of the rehearsal. None of them is a measure of the thing you were trying to improve. Ask for one number from outside:

- Sales: win rate or booked revenue, by cohort
- Hiring and interview products: offers received
- Clinical training: eventually a real clinical outcome

- Customer service: escalation or resolution, not satisfaction with the training

WHAT COMES NEXT

The history points to a new possibility

Every established rehearsal market in this research involved something people did often enough, or were required to practice often enough, to justify the cost. Pilots fly. Surgeons operate. Students take the test.

The moments that were never practiced are the rare ones. Your first time firing someone. A negotiation that happens once in a career. A procedure a rural physician performs twice a year. An operational failure a team has read about and never lived through.

Those were not skipped because they did not matter. They were skipped because a simulation for something that happens once is impossible to justify. That arithmetic is what AI actually changes.

Rarity by itself is not a business, and this is a prediction rather than a finding. Someone still has to work out what makes that particular moment hard for that particular person, and put that specific thing inside the practice. Most of what is being built right now skips that step and reproduces the visible shape of the event instead.

People are about to be able to practice far more moments before they count. The winners will be the ones that know which part of the moment needs practicing.

AI can make almost anything rehearsable. Most of what gets built will make people feel more prepared, and it will have the engagement numbers to prove it.

Feeling prepared is easy to produce and easy to measure. Changing what happens when the moment counts is harder. That is the outcome executives should demand evidence for.

Sources

- [Lund, Williams and Zador, further evaluation of the DeKalb County driver education study](#), Accident Analysis and Prevention, 1986.
- [Seymour and colleagues, virtual reality training and operating room performance](#), Annals of Surgery, 2002.

- [Smith and colleagues, virtual reality job interview training and job offers at six months](#). Note that the University of Michigan receives royalties on the tool evaluated in related trials.
- [Smith and colleagues, mechanism of action for obtaining job offers with virtual interview training](#), Psychiatric Services, 2017.
- [Randomized trial of virtual reality gaze exposure for public speaking anxiety](#), 2023.
- [Systematic review of CPR knowledge, self-efficacy and psychomotor skills](#), covering 34 articles and more than 35,000 participants.
- [Longitudinal study of basic life support skills in high school students](#), including the confidence figures.
- [Meta-analysis of AI-assisted versus traditional clinical simulation training](#), Frontiers in Medicine, 2026.
- [College Board announcement of free official SAT practice with Khan Academy](#), 2015, and the [College Board summary of the coaching research](#).
- [Review of residency selection after the USMLE Step 1 pass/fail change](#), JMIR Medical Education, 2023.
- [Civil aviation flight training and simulation market analysis](#), including regulatory simulator-hour requirements.
- [National Jurist guide to bar preparation providers and market share](#), 2026.
- [Comparison of app-based and classroom language instruction](#), Studies in Second Language Acquisition, summarizing the independent oral proficiency findings.
- [Reporting on the retirement of Google Interview Warmup](#). Competitor sourcing, no statement from Google beyond a brief note about retiring the tool.