

A escada da evidência: o que conta como prova de que se aprendeu

Uma afirmação pública sobre aprendizagem vale o degrau de evidência que a sustenta, e nada mais. Cinco degraus, do racional à eficácia independente: sobe-se um degrau por prova produzida, nunca antes. Vale para fornecedores, e vale para o **IA para escolas**.

1

Racional

A solução cita as ciências da aprendizagem e um mecanismo plausível. É o nível "demonstra um racional" da lei americana ESSA (Every Student Succeeds Act).

2

Validação técnica

A leitura por critérios da IA (sem nota) concorda com a de corretores humanos, competência a competência (kappa), com os erros publicados.

3

Ciclo documentado

Iterações registradas: o que mudou, quando, por quê e em que versão. Melhoria contínua com data.

4

Ganho com comparação

Pré e pós com grupo de comparação, protocolo registrado antes da coleta e coautoria acadêmica independente.

5

Eficácia independente

Quase-experimento ou ensaio randomizado conduzido por terceiros, com pré-registro e dados abertos.

Caso de fornecedor não é evidência

A escada é uma regra editorial: nenhuma frase acima do seu degrau.

OS NÍVEIS QUE EXISTEM

A lei americana (ESSA) distingue evidência forte, moderada, promissora e "demonstra um racional", com o What Works Clearinghouse como árbitro; o AI Act europeu põe a maioria das aplicações educacionais no **alto risco**; a UNESCO pede validar a utilidade pedagógica antes da adoção (*Validação de ferramentas*).

O QUE A MAIORIA TEM

Menos de 2% das intervenções avaliadas pelo What Works Clearinghouse (WWC) tinham evidência forte ou moderada (UNESCO, 2023); a EEF (Education Endowment Foundation, 2019) conclui que o resultado depende do contexto de implementação. O *Caderno 4* já pergunta: "**há estudo independente do resultado?**".

QUEM LIDERA

Carnegie Learning e ASSISTments (ESSA Tier 1), Eedi (ensaio com 3.000 alunos em 20 escolas), Khan Academy (efeito 0,36 entre quem atingiu a dose; nenhuma evidência do Khanmigo até 2024). Ninguém diz "somos engenharia da aprendizagem"; todos dizem **o degrau em que estão**.

No IA para escolas. Hoje o sistema está no degrau 1: fundamentos publicados, mecanismo plausível, tempo devolvido medido. A escada diz o que pode ser dito agora e o que só depois de cada prova; e a frase "comprovado" fica proibida até o degrau 4.

1 Racional

Projetado com base na ciência: verdadeiro, e só isso.

O QUE AUTORIZA DIZER

"Projetado com base nas ciências da aprendizagem"; "processo de engenharia da aprendizagem em cinco fases"; "nenhuma nota é lançada sem validação humana, e a nota de toda produção autoral é do professor". Descrições de **desenho**, não de resultado.

O QUE EXIGE

Os fundamentos publicados e uma página que explique o ciclo e como o sistema o implementa. É o nível "demonstra um racional" da ESSA: **mecanismo plausível**, sem medida ainda.

NA FRICÇÃO ÓTIMA

O racional é forte: Bjork, Sweller, Kapur, Bastani, Kosmyna (*Fricção cognitiva ótima*, Seções dois a seis). Mas racional não é resultado: "**falar bem não é saber**" (hábito 5) vale também para quem vende.

No IA para escolas. É o degrau atual, e já é mais do que a maioria dos fornecedores tem. O risco é a palavra "comprovado" aparecer antes da hora; a escada é o freio.

2 Validação técnica

A IA concorda com o professor? Em qual competência?

O QUE AUTORIZA DIZER

"Leitura por critérios validada contra corretores humanos: concordância de X por competência; Y% das contestações confirmam a nota do professor". Números por competência, **inclusive os ruins**.

O QUE EXIGE

Amostra de ao menos 300 redações; dois corretores humanos cegos, com a rubrica do ENEM (Exame Nacional do Ensino Médio) ou a da escola; a evidência apontada pela leitura da IA, sem nota, é comparada com a dos corretores; kappa ponderado por competência, com meta de **0,70**, o padrão da literatura; publicar onde a IA erra mais.

NA FRICÇÃO ÓTIMA

Viés humano e viés algorítmico se calibram um contra o outro, sob auditoria: a validação é essa auditoria com número. A explicabilidade (critério aplicado, não nota opaca) é condição (Princípios éticos, princípio 6).

No IA para escolas. É a prova mais barata e mais vendável: seis a oito semanas, o custo dos corretores. A contestação já registrada dá a segunda métrica (material 3).

3 Ciclo documentado

Melhoria contínua com data e motivo.

O QUE AUTORIZA DIZER

"Melhoria contínua: N iterações no último semestre, com o que mudou e por quê". A partir daqui, e só a partir daqui, "engenharia da aprendizagem" deixa de ser slogan: **ciclo publicado** é o que o campo reconhece.

O QUE EXIGE

Registro de decisões (LEED, Learning Engineering Evidence and Decision tracker) resumido publicamente; changelog pedagógico por semestre; versões da ontologia numeradas. Totino e Kessler (2025) mostram que até os engenheiros da aprendizagem precisam de disciplina para isso.

NA FRICÇÃO ÓTIMA

"Uma ontologia que se versiona" (hábito 6) só é verificável se as versões e os motivos forem públicos. Governança viva, não documento na gaveta (*Caderno 3*).

No IA para escolas. A reunião semanal já produz iterações; o que falta é a ata com evidência (material 3) e um resumo semestral por escola. É o degrau mais próximo, e o que muda o discurso comercial.

4 Ganho com comparação

Melhoraram quanto, em relação a quem?

O QUE AUTORIZA DIZER

"Alunos que passaram por três ciclos de devolutiva melhoraram Z pontos na competência W, em relação a turmas comparáveis", com as ressalvas de amostra e contexto.

O QUE EXIGE

Estudo pré e pós com grupo de comparação (correção cruzada ou gêmea), **protocolo registrado** antes da coleta, coautoria acadêmica independente, efeito com intervalo de confiança, não só média.

NA FRICÇÃO ÓTIMA

A fase sem IA da correção gêmea é o desfecho: o que o aluno produz sozinho depois dos ciclos. É a única medida que responde a "**desempenho não é aprendizagem**".

No IA para escolas. Um semestre de coleta com um parceiro acadêmico; o material 5 mostra o desenho. Publicar também o resultado nulo: o campo premia isso, e diferencia de quem só publica sucesso.

5 Eficácia independente

Quando outro mede, o número muda de valor.

O QUE AUTORIZA DIZER

"Efeito de $d = \dots$ em estudo independente", e a entrada em listas de ferramentas validadas, como a UNESCO recomenda antes de responsabilizar o professor.

O QUE EXIGE

Quase-experimento ou ensaio randomizado conduzido por terceiros (universidade, hub de avaliação); pré-registro; dados abertos anonimizados; **resultados publicados** sejam quais forem.

NA FRICÇÃO ÓTIMA

Mesmo aqui o efeito depende da implementação: o Khanmigo em 18 escolas por dois anos rendeu 0,06 a 0,08 desvio-padrão ao ano, com 1,7 sessão semanal; os efeitos grandes (Harvard, 0,73 a 1,3) vieram de tutores com salvaguardas em **contexto curto**.

No IA para escolas. Horizonte de dois anos, com edital ou parceria. Chegar ao degrau 5 sem passar pelos anteriores é o erro clássico da edtech: a escada existe para não pular.

Seis perguntas sobre a prova

- ? Em que degrau está a afirmação: racional, validação, ciclo, comparação ou estudo independente?
- ? Quem mediu: o próprio fornecedor, um parceiro acadêmico ou um terceiro independente?
- ? A concordância entre IA e corretores humanos é conhecida por competência, com o número?
- ? A medida é de aprendizagem com a IA retirada, ou de uso, satisfação e tempo economizado?
- ? O protocolo foi registrado antes da coleta, e os resultados nulos também foram publicados?
- ? O que mudou no produto por causa dos resultados, e onde isso está escrito?

Créditos e referências

Quarto material da série *Engenharia da aprendizagem*. Responde à quarta pergunta do *Caderno 4* ("há estudo independente do resultado?") e ao critério de evidência pedagógica da *Validação de ferramentas* (UNESCO), e usa a correção gêmea e a cruzada (*Tipos de correção com IA*) como desenhos de medida. Os níveis da ESSA e do What Works Clearinghouse são a referência internacional; a escada de cinco degraus e as frases autorizadas por degrau são do **IA para escolas**.

UNITED STATES. DEPARTMENT OF EDUCATION. Non-regulatory guidance: using evidence to strengthen education investments. Washington, DC, 2016 (rev. 2023).

UNESCO. Global education monitoring report 2023: technology in education, a tool on whose terms? Paris: UNESCO, 2023.

EDUCATION ENDOWMENT FOUNDATION. Using digital technology to improve learning: evidence review. London: EEF, 2019.

KHAN ACADEMY. Khan Academy efficacy results. Mountain View: Khan Academy, 2024.

OREOPOULOS, P.; LOW, N. One click away: AI tutoring with Khanmigo in a two-year school experiment. EdWorkingPaper AI26-1551, 2026.

KESTIN, G. et al. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, v. 15, 17458, 2025.

MIAO, F.; CUKUROVA, M. AI competency framework for teachers. Paris: UNESCO, 2024.

Oldimar Cardoso · iaparaescolas.com.br · 2026