

Experimentar sem virar laboratório: como comparar devolutivas na escola

Um experimento na escola não precisa de laboratório: precisa de uma pergunta, duas versões, uma alocação honesta, uma medida antes e depois, e transparência com quem participa. Cinco decisões de desenho, com a correção gêmea e a cruzada como ponto de partida.

1

A pergunta

Uma hipótese que um dado pode responder: "a devolutiva por critérios melhora a competência 3 mais do que a devolutiva com trecho reescrito?".

2

As variantes

Duas ou três versões da devolutiva, iguais em tudo o mais; a célula da Escala de Avaliação fixa o resto.

3

A alocação

Por turma ou por aluno, sorteada e registrada; nunca escolhida a dedo. É o que separa comparação de anedota.

4

A medida

Antes e depois, com a IA retirada: a fase sem IA da correção gêmea ou o segundo instrumento da cruzada.

5

A transparência

Famílias, alunos e professores sabem que há comparação, podem sair, e recebem o resultado. O comitê de IA aprova.

Toda escola já experimenta; poucas registram

Mudar a rubrica e ver o que acontece é um experimento sem controle.

MAIS DE 205 TRILHÕES

Koedinger, Booth e Klahr (2013) estimam esse número de combinações de decisões instrucionais. A resposta do campo é embutir experimentos onde o ensino acontece: **E-TRIALS** (ASSISTments), **UpGrade** (Carnegie Learning), **Terracotta** (Indiana University), com consentimento automatizado.

PEQUENO E HONESTO

Um quase-experimento por turma, com pré e pós, vale mais do que um depoimento. A matriz do LEVI (Learning Engineering Virtual Institute; Whitmer e Andres, 2024) aceita métodos de cinco níveis, desde que o desfecho seja declarado e o método seja **adequado ao estágio** do produto.

O QUE JÁ EXISTE

A correção gêmea (fase sem IA, fase com IA) e a correção cruzada (dois instrumentos para a mesma aprendizagem) são desenhos **quase-experimentais** à espera de instrumentação (*Tipos de correção com IA*, tipos 5 e 6).

No IA para escolas. O sistema já permite variar o tipo de devolutiva por atividade e registrar a refeição. Faltam a alocação sorteada e a medida sem IA. Com isso, a correção cruzada vira experimento de verdade, e o degrau 4 da evidência (material 4) fica ao alcance de um semestre.

1 A pergunta

Uma pergunta que um dado pode responder.

O QUE É

Hipótese com variável (tipo de devolutiva), desfecho (nota na competência, na fase sem IA), população (9º ano, redação dissertativa) e prazo (seis semanas, três ciclos). É o **desenho centrado em evidências** (evidence-centered design, Mislevy): partir da evidência que se quer produzir.

COMO SE FAZ

Escrever a pergunta em uma frase e o resultado que a **refutaria**. Registrar o protocolo antes de coletar, com data, no changelog ou em repositório aberto (OSF, Open Science Framework).

NA FRICÇÃO ÓTIMA

A pergunta certa mede o produto cognitivo: "com a IA retirada, o aluno faz sozinho o que antes não fazia?" (*Fricção cognitiva ótima na prática*, tela 23).

No IA para escolas. Primeira pergunta candidata: "a devolutiva por critérios, sem trecho reescrito, produz refações melhores do que a devolutiva com exemplo reescrito?". Ela testa a diretriz 4 (estrutura, não solução) com dado.

2 As variantes

Iguais em tudo, exceto no que se testa.

O QUE É

Duas ou três versões da mesma devolutiva: detalhada por critério; sintética; socrática (só perguntas). Tudo o mais igual: rubrica, prazo, tipo de correção, **célula da Escala**.

COMO SE FAZ

Um marcador por turma ou por aluno que escolhe a variante; o professor revisa todas do mesmo modo; as variantes vivem na ontologia como **regras alternativas**, com versão.

NA FRICÇÃO ÓTIMA

As variantes são doses de fricção: perguntas (mais), critérios (média), exemplo reescrito (menos). O experimento mede qual dose serve a qual aluno: **calibragem com dado** (tela 14).

No IA para escolas. A metacorreção já usa uma amostra corrigida pelo professor como exemplo de referência (sem treinar modelos com dados dos alunos) e prepara, nesse padrão, as demais leituras: é uma variante natural (leitura no padrão do professor contra rubrica padrão) que ninguém precisa inventar; em ambas, a nota é do professor.

3 A alocação

Sortear é o que separa comparação de anedota.

O QUE É

Alocação aleatória por turma (mais simples, evita contaminação) ou por aluno (mais poder estatístico); registro de quem recebeu o quê; **equilíbrio** verificado (nota inicial, série). Sem sorteio, a diferença pode ser da turma, não da devolutiva.

EXEMPLO

O UpGrade randomiza por grupo dentro do produto; o E-TRIALS randomiza dentro do mesmo conjunto de exercícios; a Khan Academy só adota uma mudança com **regra de decisão de 95%**.

NA FRICÇÃO ÓTIMA

Equidade: nenhuma variante pode ser pior do que o que a escola já faz; se uma se mostrar melhor, **todos a recebem** no ciclo seguinte. É a iteração (material 2).

No IA para escolas. O sorteio pode ser feito na criação da correção (um modelo institucional com duas variantes) e registrado como evento; o professor não precisa saber a variante ao revisar.

4 A medida

Antes, depois, e sem a IA.

O QUE É

Pré: a primeira redação do ciclo.
Pós: a redação final, na fase sem IA (gêmea) ou no segundo instrumento (cruzada). Nota por competência por dois corretores cegos, com a leitura por critérios da IA (validada no degrau 2) como apoio.
Efeito: **diferença das diferenças** entre variantes.

COMO SE FAZ

Tamanho realista: quatro turmas por variante permitem ver efeitos médios; menos que isso é **piloto**, e deve ser chamado assim. Publicar o intervalo, não só a média.

NA FRICÇÃO ÓTIMA

Medir também a refeção concluída e a tentativa registrada antes da ajuda: o desfecho da fricção é o **produto do aluno**, não a fluência da devolutiva.

No IA para escolas. O histórico de refeições e a nota por competência já existem; a fase sem IA é a única mudança de desenho: uma redação em sala, à mão, fecha o ciclo e vira a medida.

5 A transparência

Quem participa sabe que participa.

O QUE É

Famílias, alunos e professores informados de que há comparação entre devolutivas, com direito de não participar; dados anonimizados; resultado comunicado ao final. LGPD (Lei Geral de Proteção de Dados) e o dever de **transparência** (*Princípios éticos*, princípio 6).

COMO SE FAZ

Um parágrafo na carta às famílias (*Caderno 3*, tela 7): o que se compara, por que, o que muda para o aluno, como sair. O **comitê de IA** da escola (*Guia*, Parte II) aprova o protocolo.

NA FRICÇÃO ÓTIMA

Letramento em IA (Caderno 7): o aluno que sabe que está sendo comparado aprende a perguntar "o que a IA mudou no meu texto?". A transparência é **conteúdo**, não só obrigação.

No IA para escolas. Os aceites legais visíveis e o encarregado de dados (DPO, Data Protection Officer) já existem; o protocolo do experimento entra como anexo do termo com a escola, e o resultado, no Relatório de Evidências (material 7).

Seis condições para um experimento honesto

- ✓ A pergunta cabe em uma frase e diz que resultado a refutaria; o protocolo tem data anterior à coleta.
- ✓ A alocação foi sorteada e registrada; o equilíbrio inicial entre os grupos foi verificado.
- ✓ Famílias, alunos e professores foram informados e podem sair; o comitê de IA aprovou.
- ✓ As variantes diferem em uma coisa só, e nenhuma é pior do que a prática atual da escola.
- ✓ O desfecho é medido antes e depois, com a IA retirada, por corretores cegos, com apoio da leitura por critérios validada.
- ✓ O resultado, inclusive nulo, será publicado e virará a versão seguinte da ontologia.

Créditos e referências

Quinto material da série *Engenharia da aprendizagem*. Transforma dois *Tipos de correção com IA* (gêmea e cruzada) em desenhos de estudo, usa a *Escala de Avaliação com IA* como variável de controle, a carta às famílias do *Caderno 3* e o comitê de IA do *Guia* como governança do experimento, e o *Caderno 7* como lado do aluno. Os exemplos de plataformas com experimentos embutidos (E-TRIALS, UpGrade, Terracotta) são a referência do campo; o desenho para redação é do **IA para escolas**.

KOEDINGER, K. R.; BOOTH, J. L.; KLAHR, D. Instructional complexity and the science to constrain it. *Science*, v. 342, n. 6161, p. 935–937, 2013.

MISLEVY, R. J.; STEINBERG, L. S.; ALMOND, R. G. On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, v. 1, n. 1, p. 3–62, 2003.

MOTZ, B. A. et al. Terracotta: a tool for conducting experimental research on student learning. *Behavior Research Methods*, v. 56, n. 3, p. 2519–2536, 2024.

ROE, J.; PERKINS, M.; GIRAY, L. Assessment twins: strengthening assessment validity in the age of generative AI. *Journal of Applied Learning & Teaching*, v. 9, n. 2, 2026.

WHITMER, J.; ANDRES, J. Creating a responsive, systematic approach to early edtech evidence: the LEVI Trialing Hubs matrix. Learning Engineering Virtual Institute, 2024.

ASSISTMENTS. E-TRIALS: educational technology research infrastructure to advance learning science. Worcester: Worcester Polytechnic Institute, 2026. Disponível em: assistments.org/e-trials.

CARNEGIE LEARNING. UpGrade: an open source platform for A/B testing in education. Pittsburgh, 2026. Disponível em: upgradeplatform.org.

Oldimar Cardoso • iaparaescolas.com.br • 2026