

O que a prova sabe e a nota esconde

Toda prova corrigida guarda, item por item, um diagnóstico da turma: que habilidade falhou, que erro se repetiu, que questão não funcionou. A nota soma tudo num número e perde o resto. Cinco leituras da analytics da avaliação, e o lugar do **IA para escolas** em cada uma.

1

Assessment analytics

A avaliação é o dado mais abundante da escola, e o mais ligado ao que se quis ensinar.

2

Análise de itens

Dificuldade, discriminação e distratores: três números por questão que a Teoria Clássica dos Testes já oferece.

3

Resposta ao item

O modelo das avaliações nacionais põe provas diferentes numa mesma escala, mas pede amostras grandes.

4

A matriz Q

Cada item ligado às habilidades que exige: no Brasil, ao descritor ou à habilidade da base curricular.

5

Triangulação

Dois instrumentos para a mesma aprendizagem: quando concordam, o diagnóstico se firma.

A nota é a média de muitos diagnósticos perdidos

Dois alunos com 6,0 podem ter errado provas completamente diferentes.

O PROBLEMA

A nota responde a uma pergunta administrativa (aprovado ou não?) e silencia as pedagógicas. Um 6,0 pode ser quem errou toda a geometria ou quem errou um pouco de cada coisa. Para a próxima aula, são **turmas diferentes** dentro da mesma média.

O QUE SE PERDE

Cada item tem uma história: quantos acertaram, quem acertou, que alternativa errada atraiu mais gente. Somada à nota, essa história desaparece. O material 2 desta série pediu que o dado fechasse o ciclo; aqui, o dado mais útil está **no item**, não no total.

A CORREÇÃO

Ler a prova duas vezes: por aluno, para a nota que o professor valida; e por item, para decidir o que reensinar. Bambrick-Santoyo (2010) fez dessa **análise por item**, seguida de reunião de dados e reensino, o núcleo do trabalho com avaliação na escola.

No IA para escolas. A correção assistida já produz o dado por item: o supergabarito, gerado a partir do gabarito simples, organiza a leitura das respostas, e o professor valida cada resultado, com registro, antes do lançamento. O painel da correção mostra o desempenho da turma questão a questão e por critério.

1 Assessment analytics

O dado que a escola já tem.

O QUE É

Ellis (2013) defendeu a **assessment analytics**: usar os dados da avaliação, que toda escola já produz, como fonte central da learning analytics. São dados abundantes, ligados ao currículo e mais próximos da aprendizagem do que qualquer registro de acesso.

A TRÍADE

Knight, Buckingham Shum e Littleton (2014) lembram que toda analytics carrega uma **epistemologia** (o que conta como saber), uma avaliação e uma pedagogia. Medir só o acerto final supõe que saber é acertar; medir o raciocínio supõe outra coisa.

EXEMPLO

O Gradescope (Singh e col., 2017) corrige por questão e agrupa respostas semelhantes; o dado por item vira **subproduto** da correção. Ali, uma decisão vale para o grupo inteiro; é exatamente o ponto em que o desenho precisa de cuidado.

No IA para escolas. As respostas idênticas são agrupadas para orientar a revisão, mas a validação continua resultado a resultado, nunca em bloco (*Fundamento 4*, correção validada). O agrupamento economiza leitura; não transfere a decisão.

2 Análise de itens

Três números por questão.

O QUE É

A TCT (Teoria Clássica dos Testes) oferece três leituras por item: a **difículdade** (proporção de acerto), a **discriminação** (quanto o item separa quem foi bem de quem foi mal na prova) e a análise dos distratores (quem escolheu cada alternativa errada).

COMO SE FAZ

Um item que quase todos erram pode medir algo não ensinado. Um item com discriminação perto de zero, ou negativa, em que os melhores erram mais, pode estar **mal formulado**. Às vezes é o item, e não o aluno, que está errado.

O CASO

A Eedi escreve perguntas em que cada distrator codifica uma **concepção equivocada**; diz ter mais de 8.000 catalogadas. Seus dados sustentaram o desafio do NeurIPS 2020 (Wang e col., 2020) e uma competição no Kaggle com Vanderbilt e The Learning Agency Lab.

No IA para escolas. Duas propostas de produto: anotar no supergabarito o erro que cada alternativa revela, para que o painel diga "12 alunos confundiram área com perímetro", e não só "item 7: 40%"; e sinalizar ao professor o item de discriminação baixa antes de a nota ser lançada.

3 Teoria de Resposta ao Item

O que a escala nacional permite comparar.

O QUE É

A TRI (Teoria de Resposta ao Item) modela a probabilidade de acerto a partir da proficiência do aluno e de parâmetros do item. O SAEB (Sistema de Avaliação da Educação Básica) e o ENEM (Exame Nacional do Ensino Médio), este desde 2009, usam a TRI.

O QUE PERMITE

Pôr provas diferentes numa **mesma escala**, e assim comparar edições e redes. No ENEM, o modelo de três parâmetros (discriminação, dificuldade, acerto ao acaso) faz o padrão de respostas pesar: acertar as difíceis e errar as fáceis rende menos.

O LIMITE

A TRI exige muitos respondentes e itens **calibrados** em pré-teste. Numa prova de turma, com 30 alunos e itens novos, ela não se sustenta. Na escola, a TCT é a ferramenta honesta; a TRI fica para os dados públicos.

No IA para escolas. A prova da escola é lida pela TCT, item a item; a leitura por área do ENEM, a partir dos dados públicos, serve à conversa com cada escola. Nenhuma nota simulada é calculada, e nenhuma nota futura é prevista.

4 A matriz Q

Cada item ligado ao que exige.

O QUE É

Tatsuoka (1983) formalizou a ideia: uma tabela que liga cada item às habilidades que ele exige, a **matriz Q**. Modelos diagnósticos como o DINA (Deterministic Input, Noisy "And" gate; Junker e Sijtsma, 2001) partem dela para estimar que habilidades faltam, e não só quantos itens se acertou.

NO BRASIL

A escola já tem uma matriz Q informal: o **descriptor** da matriz de referência do SAEB ou a habilidade da BNCC (Base Nacional Comum Curricular) ligada a cada questão. Basta registrá-la na hora de montar a prova.

O CUIDADO

Um item quase nunca exige uma só habilidade. Ligar cada questão a um único código simplifica a leitura e pode **esconder** a causa do erro: o aluno errou a equação ou não entendeu o enunciado?

No IA para escolas. O Monitorar lê o simulado de rede por habilidade da BNCC em déficit e entrega por turma sugestões de grupos de reforço por habilidade, revisadas e decididas pelo professor, além de orientação pedagógica e planejamento de aula.

5 Triangulação

Dois instrumentos, uma aprendizagem.

O QUE É

Nenhum instrumento sozinho mede uma aprendizagem. A triangulação cruza **dois ou mais**: a prova objetiva e a resposta dissertativa, a lição de casa e a redação, a avaliação da escola e o simulado. Quando concordam, o diagnóstico se firma; quando divergem, há algo a investigar.

EXEMPLO

A turma vai bem nos itens de interpretação da prova, mas a redação mostra argumentação frágil. Não é contradição: os itens podem medir o **reconhecimento**, e a redação, a produção. A divergência indica o que reensinar.

NA ESCADA DA EVIDÊNCIA

A triangulação também protege a própria correção: a correção cruzada (*Fundamento 4*) compara dois instrumentos para a mesma aprendizagem, e duas leituras da mesma produção por professores diferentes triangulam os leitores. É o caminho do **degrau 2** do material 4 de *Engenharia da aprendizagem*.

No IA para escolas. O chat de dados permite cruzar instrumentos por turma e habilidade ("Quais habilidades da BNCC estão mais frágeis?"), como apoio à decisão: responde com dados reais, não rotula alunos e deixa a leitura com o professor e a coordenação.

Quatro leituras por item, e duas decisões

- ✓ Dificuldade: a proporção de acerto de cada item está à vista.
- ✓ Distratores: a alternativa errada mais escolhida tem uma hipótese de erro anotada.
- ✓ Um item com problema foi revisto antes de pesar na nota de alguém.
- ✓ Discriminação: os itens em que os melhores alunos erraram mais foram revistos.
- ✓ Habilidade: cada item está ligado ao descritor ou à habilidade da BNCC.
- ✓ O reensino foi decidido e registrado, com a data da próxima verificação.

Créditos e referências

Quarto material da série *Learning analytics*. Lê a prova item a item, a partir do que a coleção já contém: a avaliação do Caderno 5, a correção validada, cruzada e gêmea (*Fundamento 4*), a escada da evidência (*Engenharia da aprendizagem*, material 4) e o ciclo do material 2 desta série. A série *Hackeamento da ensinagem* leva a análise por item ao papel escaneado no **IA para escolas**.

ELLIS, C. Broadening the scope and increasing the usefulness of learning analytics: the case for assessment analytics. *British Journal of Educational Technology*, v. 44, n. 4, p. 662–664, 2013.

KNIGHT, S.; BUCKINGHAM SHUM, S.; LITTLETON, K. Epistemology, assessment, pedagogy: where learning meets analytics in the middle space. *Journal of Learning Analytics*, v. 1, n. 2, p. 23–47, 2014.

SINGH, A. et al. Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work. In: *L@S '17*. New York: ACM, 2017. p. 81–88.

TATSUOKA, K. K. Rule space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, v. 20, n. 4, p. 345–354, 1983.

JUNKER, B. W.; SIJTSMA, K. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, v. 25, n. 3, p. 258–272, 2001.

WANG, Z. et al. Instructions and guide for diagnostic questions: the NeurIPS 2020 Education Challenge. arXiv:2007.12061, 2020.

BAMBRICK-SANTOYO, P. Driven by data: a practical guide to improve instruction. San Francisco: Jossey-Bass, 2010.