Avaliação Adaptativa com Múltiplos Estágios como Ferramenta Diagnóstica da Trajetória Escolar em Língua Portuguesa e Matemática

Mônica Cristina Weinstein, Araê Cainã, Gustavo Henrique Martins, Lucas Pereira Sperandio e Gisele Magarotto Machado

Resumo

Este estudo apresenta o desenvolvimento de uma avaliação adaptativa computadorizada com múltiplos estágios para identificar habilidades dominadas e defasagens acumuladas em Língua Portuguesa e Matemática. Foram selecionadas 132 habilidades de Língua Portuguesa e 149 de Matemática, com itens organizados por ano escolar e aplicados a cerca de 19.500 estudantes. A calibração dos itens foi realizada com o modelo 2PL da TRI e os dados analisados por meio de simulações post hoc.

Introdução

O sistema educacional brasileiro enfrenta desafios significativos na garantia da aprendizagem essencial em Língua Portuguesa e Matemática. Diversos estudos indicam que grande parte dos estudantes não adquire os conhecimentos esperados para sua idade e ano escolar. Por exemplo, dados recentes mostram que mais de 50% dos alunos do 5º ano apresentam desempenho abaixo do básico, com dificuldades em resolver operações simples de adição e leitura fluente de textos curtos (TODOS PELA EDUCAÇÃO, 2022; INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, 2023).

Tipicamente, os estudantes são avaliados com base apenas no currículo do ano em que estão matriculados, e os resultados indicam se eles dominaram ou não os conteúdos correspondentes àquela série. No entanto, esse modelo de avaliação não permite identificar em que ponto da trajetória escolar o processo de aprendizagem foi interrompido ou comprometido. Como consequência, as intervenções pedagógicas tendem a reforçar o currículo atual, sem necessariamente abordar as lacunas acumuladas de anos anteriores. Essa abordagem reduz a efetividade das ações educacionais, uma vez que as dificuldades enfrentadas pelos alunos muitas vezes têm origem em etapas anteriores de escolarização (DIAS; RAMOS, 2022).

Nesse contexto, as avaliações adaptativas computadorizadas (*Computerized Adaptive Testing* – CAT; VAN DER LINDEN; GLASS, 2010), assim como os testes em múltiplos estágios (*Multistage Testing* – MST; YAN; VON DAVIER; LEWIS, 2014), surgem como alternativas promissoras, pois ajustam a dificuldade dos itens ao nível de proficiência do

respondente, aumentando a precisão da medida e otimizando o tempo de aplicação. Além disso, permitem que o aluno seja direcionado a diferentes estágios da avaliação (no caso deste trabalho, a conteúdos de anos anteriores ou subsequentes).

Objetivo

O objetivo desse trabalho foi desenvolver uma avaliação adaptativa computadorizada com múltiplos estágios, visando avaliar o nível de proficiência dos alunos em Língua Portuguesa e Matemática ao longo da trajetória escolar, identificação de habilidades dominadas e defasagens acumuladas nos anos anteriores ao ano de matrícula do aluno.

Metodologia

Inicialmente, especialistas em Língua Portuguesa e Matemática da Parceiros da Educação, selecionaram dentre as habilidades da BNCC, habilidades consideradas como essenciais e estruturantes em cada um dos anos escolares. No total, 132 habilidades de Língua Portuguesa foram selecionas e 149 de Matemática, em média 10 habilidades por ano escolar em cada um dos componentes. Itens foram desenvolvidos para cada uma dessas habilidades.

Os itens foram inseridos na plataforma digital CATvante e organizados em 13 estágios, cada um correspondente a um ano escolar. A aplicação inicial de validação dos itens e construção do algoritmo foi estruturada de modo que o aluno iniciasse a avaliação respondendo aos itens do estágio imediatamente anterior ao seu ano de matrícula. A depender de seu desempenho (incialmente baseado em quantidade de acertos e erros), ele poderia ser direcionado a estágios anteriores (em caso de baixo desempenho) ou subsequentes (em caso de alto desempenho). Participaram deste estudo aproximadamente 1.500 alunos matriculados em cada um dos 13 anos escolares (~19.500 alunos).

Para fins de calibração dos itens e desenvolvimento do algoritmo adaptativo e de múltiplos estágios, foram analisadas apenas as respostas aos itens do ano anterior ao de matrícula. A calibração foi realizada separadamente para cada bloco por meio da Teoria de Resposta ao Item (TRI), utilizando o modelo de dois parâmetros (2PL) e a correlação ponto bisserial dos itens foi calculada. Itens com índice de discriminação inferior a 0,30 e/ou com correlação bisserial negativa foram excluídos. Após exclusões, os itens foram calibrados novamente e, por fim, realizamos uma simulação post hoc do algoritmo de CAT para cada estágio. A calibração dos itens foi realizada utilizando o pacote mirt (CHALMERS, 2012) e a simulação post hoc utilizando o pacote simCAT (JALOTO, 2025) no software R.

Para fins do desenvolvimento da parte de estágios múltiplos, os níveis de proficiência nos itens de cada ano foram categorizados em quatro níveis: muito crítico, crítico, intermediário e adequado. A partir disso, criamos um algoritmo direciona os alunos com

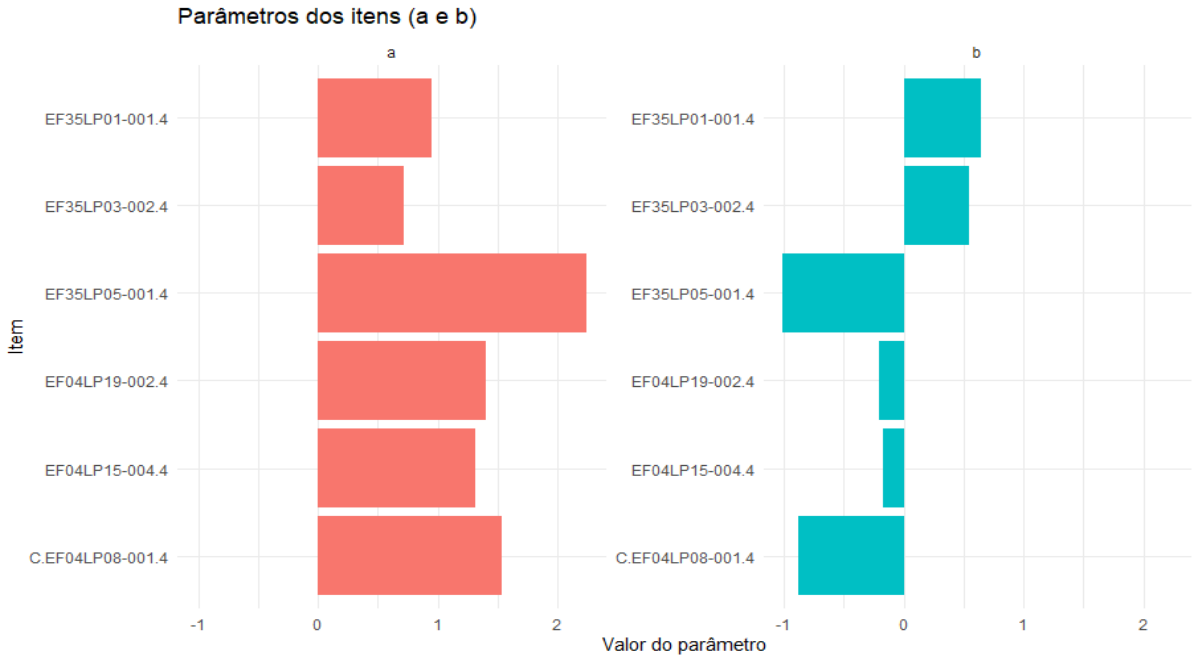
desempenho classificado como "muito crítico" a estágios compostos por conteúdos de anos anteriores; alunos com desempenho "adequado", para itens de anos seguintes; alunos com desempenho “intermediário” permanecem em itens do mesmo ano e alunos com desempenho “crítico” são direcionados para anos anteriores, porém se no ano anterior tiverem um desempenho “adequado” são redirecionados de volta para o ano no qual tiveram o desempenho “crítico”.

Resultados

A partir da primeira parametrização dos itens, 26 itens de Língua Portuguesa e 16 itens de matemática foram excluídos com base nos critérios de discriminação e correlação pontobisserial. Nós calibramos os itens de cada bloco novamente utilizando o modelo 2PL. Para fins ilustrativos, mostramos aqui apenas os resultados relativos aos itens de Língua Portuguesa relativos a habilidades do 4º do Ensino Fundamental. Os resultados dos itens dos demais anos escolares e componentes (Matemática) seguiu o mesmo padrão dos resultados aqui apresentados.

Dentre os 10 itens desenvolvidos inicialmente para 4º ano, 6 foram mantidos após as exclusões mencionadas. Após exclusões, a correlação pontobisserial dos itens variou entre 0.52 e 0.63 ($p<0.001$). Os parâmetros de discriminação (a) e dificuldade (b) da nova calibração dos itens estão apresentados na Figura 1. A média de discriminação dos itens retidos foi de 1.358 (DP = 0.527), com valores variando entre 0.71 e 2.24. A média dos parâmetros de dificuldade (b) foi de -0.18 (DP = 0.691), variando entre -1,00 e 0,646.

Figura 1.
Parâmetros a e b dos itens de LP - 4º ano.



A Tabela 1 apresenta os resultados da simulação post hoc da CAT para os itens de Língua Portuguesa do 4º ano, com diferentes critérios de parada baseados no erro padrão da estimativa (SE). Para cada valor de SE (de 0,4 a 0,9), foram realizadas 10 simulações por aluno. A habilidade final foi obtida pela média final das estimativas.

Tabela 1.
Resultados da simulação post hoc da CAT com habilidades do 4º ano de Língua Portuguesa com diferentes critérios de parada (SE)

Conteúdo	SE	SE real	RMSE	Média de Itens	DP de Itens	Min Itens	Max Itens	r
4º Ano - LP	0,4	0,609	0	6	0	6	6	1
4º Ano - LP	0,5	0,609	0	6	0	6	6	1
4º Ano - LP	0,6	0,618	0,109	5,199	0,914	4	6	0,99
4º Ano - LP	*0,7*	*0,675*	*0,28*	*3,526*	*1,19*	*2*	*6*	*0,935*
4º Ano - LP	0,8	0,745	0,429	2	0	2	2	0,84
4º Ano - LP	0,9	0,861	0,607	1	0	1	1	0,642

Nota. *SE = erro padrão alvo estipulado como critério de parada; SE real = erro padrão médio observado nas simulações; RMSE = raiz do erro quadrático médio entre a habilidade estimada e a habilidade verdadeira; Média de Itens = número médio de itens administrados; DP de Itens = desvio-padrão do número de itens administrados; Min Itens = número mínimo de itens administrados; Máx Itens = número máximo de itens administrados; r = correlação entre a habilidade estimada e a habilidade verdadeira.*

Conforme esperado, à medida que o critério de parada se torna mais exigente (isto é, à medida que o erro padrão desejado diminui), observa-se um aumento no número de itens aplicados e uma melhora na precisão da estimativa da habilidade. Quando o critério de parada foi definido como SE = 0,4 ou SE = 0,5, todos os seis itens disponíveis foram administrados a todos os estudantes, resultando em erro quadrático médio (RMSE) igual a zero e correlação perfeita entre a habilidade estimada e a verdadeira (r = 1). Nesses casos, a CAT operou de forma equivalente a um teste fixo. À medida que o critério de parada se torna mais permissivo, a média de itens utilizados diminui progressivamente, ao passo que o erro da estimativa aumenta. Com SE = 0,6, a média de itens caiu para 5,2, o RMSE foi de 0,109 e a correlação permaneceu alta (r = 0,99). Quando o critério foi SE = 0,7, o número médio de itens caiu para 3,5, com RMSE de 0,28 e correlação de 0,935. Já nos cenários mais lenientes (SE = 0,8 e SE = 0,9), o número médio de itens administrados foi de apenas 2 e 1, respectivamente, com aumento expressivo no erro (RMSE de 0,429 e 0,607) e queda nas correlações (r = 0,84 e r = 0,642). Esses resultados evidenciam o impacto direto do critério de parada na precisão e na duração da avaliação adaptativa.

Para o presente instrumento, o critério de parada foi definido a partir de um equilíbrio entre a precisão da estimativa (erro padrão), a correlação com o escore real e a quantidade média de itens administrados. Optamos por simulações que apresentassem o menor número possível de itens, desde que o erro não fosse excessivo e que não houvesse perdas significativas na correlação com a habilidade verdadeira. Essa escolha está alinhada ao propósito do modelo, que prevê que os estudantes sejam direcionados para responder a itens de anos anteriores ou posteriores, sem que o número total de itens ultrapasse o que é usualmente exigido em avaliações tradicionais (por exemplo, 14 itens para alunos do 4º ano). Com base nesses critérios, selecionamos como parâmetro para o critério de parada do estágio do 4º ano em Língua Portuguesa a simulação com SE = 0,7, que resultou em média de 3,5 itens administrados, erro padrão real de 0,675, RMSE de 0,28 e correlação de 0,935 com o escore real.

Discussão

Os resultados obtidos neste estudo demonstram o potencial das avaliações adaptativas computadorizadas com múltiplos estágios como possíveis ferramentas para a identificação de habilidades dominadas e defasagens acumuladas ao longo da trajetória escolar. O modelo desenvolvido apresentou evidências iniciais de validade. Sua vantagem é possibilitar um mapeamento mais preciso do nível de proficiência dos estudantes, considerando não apenas os

conteúdos do ano escolar em que estão matriculados, mas também os de anos anteriores e posteriores, o que representa um avanço relevante em relação às avaliações tradicionais.

Os resultados da calibração TRI e a correlação pontobiserial dos itens demonstraram que alguns itens não possuem propriedades adequadas, reforçando a necessidade de validação de avaliações no contexto educacional (o que ocorre com pouca frequência na prática das escolas no Brasil). A simulação post hoc do algoritmo em CAT mostrou que é possível reduzir substancialmente o número de itens aplicados sem comprometer significativamente a precisão da estimativa, desde que critérios de parada adequadamente calibrados sejam adotados. Essa economia na quantidade de itens referentes a um determinado ano escolar, deixa espaço para a aplicação de itens referentes a habilidades de anos escolares anteriores ou subsequentes, possibilitando avaliações que englobam habilidades da trajetória escolar, sem necessariamente aumentar o número de questões da avaliação (e a fadiga do aluno).

No entanto, algumas limitações devem ser consideradas. Primeiro, para alguns anos escolares, poucos itens foram retidos, limitando a quantidade de informação fornecida por eles, principalmente em áreas específicas de dificuldade (para a maioria dos casos, faltaram itens de média complexidade e alta complexidade). Segundo, o estudo se concentrou em propriedades psicométricas internas (parâmetros TRI, correlações pontobisseriais, simulações), mas não foram exploradas evidências de validade externa (ex: correlação com outras medidas de proficiência ou com desempenho escolar real), o que limita a robustez da inferência diagnóstica. Futuros estudos visando sanar essas limitações são necessários e estão mapeados no nosso plano de ação.

Conclusão

Este estudo apresentou o desenvolvimento e a evidências iniciais de validade de um instrumento adaptativo com múltiplos estágios para avaliar a proficiência de estudantes em Língua Portuguesa e Matemática, com base em habilidades essenciais da BNCC. Os resultados indicam que a abordagem proposta permite diagnósticos mais precisos e individualizados, com potencial para subsidiar intervenções pedagógicas mais eficazes. Embora ainda em fase inicial e com limitações, os achados indicam a viabilidade do uso de algoritmos adaptativos no contexto educacional brasileiro, apontando caminhos promissores para o aprimoramento das práticas avaliativas em larga escala.

Referências

CHALMERS, R. P. mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, [S.l.], v. 48, n. 6, p. 1–29, 2012. DOI: https://doi.org/10.18637/jss.v048.i06.

DIAS, É.; RAMOS, M. N. A educação e os impactos da Covid-19 nas aprendizagens escolares. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 30, n. 117, p. 859–870, out. 2022.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). *Relatório nacional da Avaliação da Alfabetização*. Brasília: INEP, 2023.

JALOTO, A. *simCAT: Implements Computerized Adaptive Testing Simulations*. R package. Versão 1.0.1, 2025. Disponível em: https://github.com/alexandrejaloto/simcat. Acesso em: 14 maio 2025.

TODOS PELA EDUCAÇÃO. *Anuário brasileiro da educação básica 2022*. São Paulo: Todos Pela Educação, 2022.

VAN DER LINDEN, W. J.; GLAS, C. A. W. *Elements of adaptive testing*. New York: Springer, 2010.

YAN, D.; VON DAVIER, A. A.; LEWIS, C. *Computerized multistage testing: theory and applications*. Boca Raton: CRC Press, 2014.