


Design e implementação de teste adaptativo computadorizado utilizando dados do ENEM

Erika Tiemi Anabuki¹ 

Tufi Machado Soares² 

Rafaela Reis Azevedo de Oliveira³ 

Resumo

A presente pesquisa objetiva o desenvolvimento e o *design* experimental de um Teste Adaptativo Computadorizado (CAT), a partir da plataforma R e de seus modelos estatísticos e computacionais destinados à modelagem psicométrica e à simulação de testes baseados na Teoria de Resposta ao Item (TRI). O estudo utiliza dados reais da prova de Linguagens e Códigos (LC) do Exame Nacional do Ensino Médio (ENEM), referentes às edições de 2021 a 2023. Para a calibração dos parâmetros dos itens do teste, foi definida uma amostra de 3.000 respondentes aos itens da prova de LC do ENEM. Os parâmetros dos itens foram calibrados por meio do modelo logístico de três parâmetros. A partir dos parâmetros calibrados pela TRI, o CAT construído seleciona, de maneira adaptativa, itens adequados ao nível de habilidade/proficiência de cada participante, reduzindo o número total de questões em comparação aos testes tradicionais. Essa redução contribui para menor fadiga e diminui a probabilidade de respostas aleatórias. Para testar a viabilidade do CAT, foram realizados testes experimentais com 12 alunos concluintes do ensino médio. Os resultados indicam que o CAT é sensível às características das proficiências individuais, ao permitir a identificação de perfis de desempenho distintos entre os respondentes. As conclusões destacam o CAT como um instrumento de avaliação adaptativo, reforçando seu potencial para análises educacionais baseadas em dados quantitativos e evidências, bem como para subsidiar intervenções pedagógicas direcionadas e individualizadas.

Palavras-chave: avaliação educacional; teste adaptativo computadorizado; tecnologias digitais.

Design and deployment of a computerized adaptive test using ENEM exam data

¹ Pós-doutoranda pelo Programa de Pós-graduação em Educação da Universidade Federal de Juiz de Fora (PPGE-UFJF). Professora EBTT no Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Campus Leopoldina, Minas Gerais, Brasil. ORCID: <https://orcid.org/0000-0003-4351-3538>. E-mail: tiemi@cefetmg.br

² Pós-doutorado em Educação pelo Programa de Pós-graduação em Educação da Universidade Católica do Rio de Janeiro (PUC-RJ) e em Estatística pelo Programa de Pós-graduação em Estatística da Universidade Federal do Rio de Janeiro (UFRJ). Professor Titular do Programa de Pós-graduação em Educação e Programa de Pós-graduação em Estatística da Universidade Federal de Juiz de Fora (UFJF), Campus Juiz de Fora, Minas Gerais, Brasil. ORCID: <https://orcid.org/0000-0001-9665-9341>. E-mail: tufi@caed.ufjf.br

³ Doutorado em Educação pelo Programa de Pós-graduação em Educação da Universidade Federal de Juiz de Fora (PPGE-UFJF). Professora do Programa de Pós-graduação em Educação da Universidade Federal de Juiz de Fora (UFJF), Campus Juiz de Fora, Minas Gerais, Brasil. ORCID: <https://orcid.org/0000-0002-3517-0339>. E-mail: tufi@caed.ufjf.br

Abstract

The present study aims to develop and experimentally design a Computerized Adaptive Test (CAT) using the R platform and its statistical and computational frameworks for psychometric modeling and the simulation of tests grounded in Item Response Theory (IRT). The research draws on real data from the Language and Codes (LC) section of the Brazilian National High School Exam (ENEM), covering the 2021–2023 editions. For item parameter calibration, a sample of 3,000 respondents to the LC items was defined, and the items were calibrated using the three-parameter logistic model. Based on these IRT-calibrated parameters, the CAT adaptively selects items aligned with each participant's ability or proficiency level, thereby reducing the total number of questions when compared with traditional fixed-form assessments. This reduction contributes to lower respondent fatigue and decreases the likelihood of random guessing. To assess the feasibility of the CAT, pilot testing was conducted with 12 graduating high school students. The results indicate that the CAT is sensitive to individual proficiency patterns, enabling the identification of distinct performance profiles among participants. Overall, the findings highlight the CAT as a robust adaptive assessment instrument, reinforcing its potential for data-driven educational analyses and evidence-based decision-making, as well as for supporting targeted and individualized pedagogical interventions.

Keywords: educational assessment; computerized adaptive test; digital technologies.

Desarrollo e implementación de un test adaptativo computarizado con datos del ENEM

Resumen

El presente estudio tiene como objetivo desarrollar y evaluar experimentalmente un Test Adaptativo Computarizado (CAT) mediante la plataforma R y sus marcos estadísticos y computacionales, destinados a la modelización psicométrica y la simulación de pruebas basadas en la Teoría de Respuesta al Ítem (TRI). La investigación utiliza datos de la prueba de Lenguajes y Códigos (LC) del Examen Nacional de la Enseñanza Media (ENEM) de las ediciones 2021–2023. Para la calibración de los ítems, se definió una muestra de 3.000 participantes y se aplicó el modelo logístico de tres parámetros. El CAT resultante selecciona de manera adaptativa ítems acordes al nivel de habilidad o competencia de cada estudiante, reduciendo el número total de preguntas respecto a los exámenes tradicionales de formato fijo, lo que disminuye la fatiga y la probabilidad de respuestas aleatorias. Para evaluar su viabilidad, se realizaron pruebas piloto con 12 estudiantes del último año de secundaria. Los resultados evidencian que el CAT es sensible a las diferencias individuales de competencia, permitiendo identificar perfiles de desempeño distintos. En conjunto, los hallazgos subrayan la pertinencia pedagógica y la robustez técnica del CAT como instrumento de evaluación adaptativa, destacando su potencial para análisis educativos basados en evidencia y para apoyar intervenciones pedagógicas más precisas e individualizadas.

Palabras clave: evaluación educativa; test adaptativo computarizado; tecnologías digitales.

Introdução

As avaliações educacionais são etapas fundamentais no processo de ensino-aprendizagem, desempenhando funções não apenas para o diagnóstico da aquisição de conhecimentos e competências, mas também para a identificação, monitoramento e acompanhamento do desempenho escolar dos estudantes. A avaliação educacional transcende a mera atribuição de notas ou classificações, permitindo compreender os processos de aprendizagem, detectar dificuldades individuais, orientar decisões

pedagógicas e promover intervenções que favoreçam o desenvolvimento integral dos alunos (Luckesi, 1998).

No contexto das avaliações em larga escala, tal como o Sistema de Avaliação do Ensino Básico (SAEB) e o Exame Nacional do Ensino Médio (ENEM), as avaliações realizadas pelos estudantes também são utilizadas como instrumentos institucionais para formulação de políticas públicas e a definição de estratégias para o aprimoramento da qualidade educacional, ademais das funções de prestação de contas e mensuração das proficiências de estudantes em diferentes regiões e contextos socioeconômicos (Brasil, 2024).

As avaliações em larga escala realizadas no Brasil são caracterizadas pelo modelo tradicional, isto é, testes impressos com quantitativo de questões (itens) fixos e aplicação simultânea (Sousa, 2021; Tabak *et al.*, 2023). Entretanto, também no contexto da sala de aula tradicional, as avaliações, especialmente os testes de múltipla escolha (a exemplo de simulados), em sua maioria são caracterizados por testes impressos, com a correção dos itens e cálculo das proficiências realizadas posteriormente pelo professor.

Neste contexto, um dos principais desafios dos testes impressos tradicionais está relacionado à rigidez do instrumento, que aplica o mesmo quantitativo de questões/itens a todos os participantes, independentemente de seu nível de proficiência/habilidade, o que pode resultar em testes demasiadamente fáceis para alguns estudantes (aqueles com maiores proficiências) e excessivamente difíceis para outros (aqueles com menores proficiências), comprometendo assim a precisão da medida, além de aumentar a propensão a respostas ao acaso, isto é, o “chute” (Piton-Gonçalves, 2020; Sousa, 2021; Tabak *et al.*, 2023). Além disso, especialmente em aplicações de média e larga escala, a exemplo do SAEB e ENEM, os testes impressos demandam desafios quanto à sua correção, aplicação e transporte.

Os sistemas de avaliação, assim como os demais processos de ensino e aprendizagem nas instituições educacionais, vêm sendo profundamente impactados pelo avanço das Tecnologias de Informação e Comunicação- TICs (Ribeiro; Silva, 2022). Em especial, o desenvolvimento e a aplicação de ferramentas baseadas em Inteligência Artificial (IA) têm promovido transformações significativas, possibilitando práticas pedagógicas e avaliativas mais dinâmicas, personalizadas e eficientes

(Almeida; Valente, 2023; Baylé; Desmarais, 2020; Carvalho; Santos, 2021; Chang; Zhang, 2021; Eggen, 2022).

Conforme abordado por Carvalho e Santos (2021), a incorporação das TICs nas escolas torna as práticas avaliativas mais interativas, ampliando o acesso a dados sobre o desempenho dos estudantes e possibilitando diagnósticos mais imediatos. A mediação tecnológica, nesse contexto, favorece a identificação mais rápida das dificuldades de aprendizagem, apoiando intervenções pedagógicas oportunas. Contudo, também impõe a docentes e instituições o desafio de repensar critérios avaliativos, competências digitais e condições de equidade, de modo que a inovação tecnológica resulte em melhorias efetivas no processo educativo, e não apenas em alterações superficiais nos instrumentos de avaliação.

Quanto ao contexto educacional brasileiro, esses avanços convivem com desafios significativos, sobretudo devido às desigualdades no acesso às tecnologias digitais entre regiões, redes de ensino e grupos sociais, realidade amplamente discutida por Almeida e Valente (2023). Essa assimetria tecnológica limita o potencial de inovação e exige que políticas públicas e instituições de ensino adotem estratégias que garantam equidade e inclusão digital. Logo, a incorporação das TICs e ferramentas de IA nas avaliações e mediações pedagógicas demandam não apenas recursos tecnológicos, tais como computadores e *softwares*, mas sobretudo, um olhar mais profundo sobre o sentido pedagógico do avaliar, ademais de investimentos na formação docente, infraestrutura adequada e ações que reduzam desigualdades, para que as tecnologias efetivamente contribuam para a melhoria do processo educativo.

Nesse sentido, Luckesi (1998) defende a avaliação como um processo mediador, contínuo e diagnóstico, voltado ao acompanhamento da aprendizagem e à superação de práticas meramente classificatórias. De forma convergente, Hoffmann (2014) compreende a avaliação como um processo formativo e reflexivo, centrado na observação e na intervenção pedagógica, respeitando os diferentes tempos, características e trajetórias dos estudantes. Nesse contexto, as TICs podem potencializar essa mediação ao favorecer registros sistemáticos, *feedbacks* imediatos e o acompanhamento do desenvolvimento discente, desde que orientadas por intencionalidade pedagógica. Assim, articuladas às perspectivas desses autores, as ferramentas digitais configuram-se como aliadas na consolidação de práticas

avaliativas formativas, alinhadas às demandas do sistema educacional contemporâneo.

À luz das transformações digitais ocorridas na prática escolar, os Testes Adaptativos Computadorizados (do inglês *Computerized Adaptive Testing* – CAT) se destacam como instrumentos de avaliação automatizados, que são atualmente possíveis devido aos avanços computacionais ocorridos nas últimas décadas, e potencializados pela popularização e facilidades fornecidas pelas ferramentas de IA (Baylé; Desmarais, 2020; Eggen, 2022).

Resumidamente, conforme descrito por Chang e Zhang (2021), os CATs são testes baseados em algoritmos computacionais, fundamentado pela Teoria de Resposta ao Item (TRI), que é um modelo estatístico computacional utilizado para analisar o desempenho/proficiência de participantes em testes. Ao contrário dos testes tradicionais, que somam apenas o número de acertos, a TRI considera a dificuldade dos itens, a habilidade do respondente e a qualidade/capacidade de diferenciação de cada item do teste (se ele consegue diferenciar com qualidade as proficiências dos respondentes). Neste caso, fundamentado pela TRI, o CAT ajusta dinamicamente a seleção dos itens com base nas respostas anteriores do respondente, adaptando o teste em tempo real à proficiência estimada. Com isso, é possível aplicar um número menor de itens sem comprometer a precisão da avaliação, o que permite mensurações mais dinâmicas e acuradas da proficiência do examinado.

A ideia central é que o teste se adapte ao desempenho do aluno, tornando-se mais preciso e eficiente. Ou seja, em um CAT itens muito difíceis não são aplicados a participantes com baixa proficiência porque não contribuem para uma estimativa precisa de sua proficiência, isto é, quando um aluno tem proficiência baixa (observado pelas respostas aos itens fornecidos), itens difíceis possuem probabilidade quase nula de acerto e, portanto, não fornecem informações úteis para diferenciar seu desempenho como também para ajustar a estimativa de proficiência (Eggen, 2022).

Além disso, essa abordagem contribui para a redução da fadiga do respondente e da ocorrência de acertos ao acaso (“chutes”). Outro aspecto relevante é a possibilidade de analisar a proficiência dos participantes em uma escala de proficiência derivada da TRI, a qual pode ser apresentada em sua forma padrão, isto é, com média 0 e desvio-padrão 1, ou transformada em escalas pedagógicas mais

acessíveis, como as utilizadas em avaliações em larga escala, a exemplo do ENEM (Baylé; Desmarais, 2020; Chang; Zhang, 2021; Eggen, 2022).

Recentemente, métodos estatísticos computacionais de estimação de proficiências em CAT baseados em IA, como *machine learning*, processamento de linguagem natural (PLN) e modelos neurais, têm ampliado o potencial dos CATs ao serem capazes de aprender com os dados e analisar grandes volumes de dados educacionais, especialmente em avaliações em larga escala (Baylé; Desmarais, 2020; Chang; Zhang, 2021; Lin *et al.*, 2020). O SAEB e ENEM utilizam em suas correções os algoritmos baseados em TRI, no entanto, a aplicação das avaliações ainda ocorre na forma tradicional impressa (Brasil, 2023; Piton-Gonçalves, 2012; Sousa, 2021; Tabak *et al.*, 2023).

Tendo em vista a integração de ferramentas de IA no âmbito educacional, e suas contribuições para simplificações de processos computacionais, sobretudo pela possibilidade de utilização de uma linguagem de programação mais simples e natural, as pesquisas e utilização de CATs com dados reais educacionais estão se tornando mais comuns (Aluisio; Piton-Gonçalves, 2015; Baylé; Desmarais, 2020; Catalani, 2019; Eggen, 2022; Jaloto; Primi, 2025; Magalhães; Alves; César, 2022; Sousa, 2021; Spenassato *et al.*; 2016; Van der Linden; Glas, 2010). Tais ferramentas têm sido aplicadas no contexto do desenvolvimento e *design* de CATs para otimizar a seleção de itens, aprimorar os métodos de estimação das proficiências, prever padrões de desempenho, identificar trajetórias de aprendizagem e, ainda, detectar comportamentos atípicos ou fraudulentos durante a aplicação de testes.

No Brasil, embora o uso do CAT ainda esteja em fase de consolidação, observa-se um crescimento no número de estudos que aplicam CAT e a IA em contextos educacionais, sobretudo em avaliações diagnósticas, formativas e de larga escala (Sousa, 2021). Pesquisas recentes têm explorado o potencial de algoritmos adaptativos em plataformas de ensino-aprendizagem, especialmente com o uso de *softwares* desenvolvidos na plataforma de acesso gratuito R, e seus pacotes associados, aliados a um *framework* para interfaces interativas (Alves, 2018; Baylé; Desmarais, 2020; Catalani, 2019; Chang; Zhang, 2021; Costa, 2023; Eggen, 2022; Jaloto; Primi, 2024; Magalhães; Alves; César, 2022; Pereira, 2020; Pinton-Gonçalves, 2020; Sousa, 2021; Spenassato *et al.*, 2016; Tabak *et al.*, 2023).

Ao considerar o cenário brasileiro e dados do ENEM, o presente estudo relaciona-se notadamente com o trabalho de Jaloto e Primi (2024), que exploram a aplicação de CAT com foco na seleção de itens para o ENEM, destacando estratégias de otimização da precisão da avaliação em larga escala. Embora relevante, o estudo dos autores concentra-se principalmente na comparação de algoritmos, e estimativas dos parâmetros dos itens e simulação de respostas a partir de algoritmos da TRI, sem abordar a implementação completa de um CAT operacionalizado com respondentes reais.

De forma complementar, Spenassato *et al.* (2016) analisam os benefícios dos testes adaptativos em contextos educacionais, evidenciando ganhos em eficiência quanto a diminuição dos itens administrados e validade psicométrica, mas com foco limitado a simulações, sem a implementação em alguma plataforma para coleta de dados com respondentes reais.

Diante do contexto apresentado, este artigo tem como objetivo descrever o desenvolvimento e *design* experimental de um CAT utilizando a plataforma de desenvolvimento R e seus pacotes/biblioteca *mirt*, *mirtCat* e *shiny*, e abordagem estatística e computacional fundamentada na TRI. Os dados reais para composição do banco de itens e estimação dos parâmetros dos itens do teste derivam-se do ENEM, da área de Linguagens e Códigos (LC) e das edições dos anos de 2021, 2022 e 2023. Objetiva-se, assim, demonstrar a viabilidade técnica e pedagógica na construção de um instrumento de avaliação adaptativo com base em dados reais. Dessa forma, com o intuito de alcançar o objetivo proposto, a seção seguinte detalha as etapas metodológicas implementadas na presente pesquisa.

Metodologia

A literatura aborda como procedimentos metodológicos computacionais de um CAT os modelos baseados em redes neurais, aprendizado por máquina (*machine learning*) e reforço e, predominantemente nas publicações nacionais, os modelos de TRI uni ou multivariada, que são as abordagens mais difundidas e validadas no contexto das avaliações educacionais (Catalani, 2019; Chang; Zhang, 2021; Eggen, 2022; Fernandes, 2022; Jaloto; Primi, 2024; Lin *et al.*, 2020; Piton-Gonçalves, 2012; Sousa, 2021).

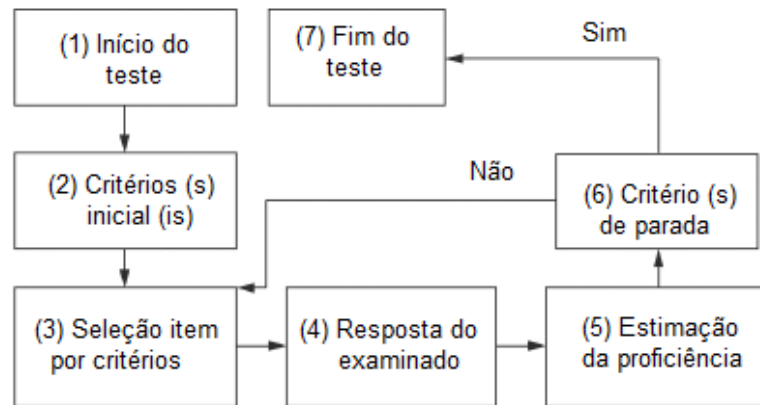
No contexto da TRI, que foi a metodologia utilizada no desenvolvimento do presente trabalho, ela se relacionada à modelagem estatística e computacional para as habilidades do respondente (também abordado como traço latente) e a definição de sua proficiência, além dos parâmetros associados aos itens, ou seja, a TRI relaciona-se diretamente com os aspectos psicométricos dos testes adaptativos (Klein, 2013; Pasquali; Primi, 2003).

Neste sentido, conforme explanado por Klein (2013), a TRI assume modelos de resposta ao item, que estabelecem uma relação entre a habilidade do respondente e a sua probabilidade de acerto. Os modelos são compostos, essencialmente, por parâmetros individuais (isto é, do respondente), parâmetros dos itens do teste e uma função que relaciona esses parâmetros com a probabilidade das possíveis respostas aos itens. Os procedimentos de inferência estatística são utilizados para obtenção de estimadores desses parâmetros a partir das respostas obtidas aos itens, que são definidos com base em critérios psicométricos da TRI.

Para melhor elucidar a metodologia de execução de um CAT, Aluisio e Piton-Gonçalves (2015) ilustram em um fluxograma o passo a passo resumido, mostrado na Figura 1, onde:

1. Inicia-se o teste (1) ao indivíduo e aplica-se um ou mais critérios iniciais para seleção do primeiro item administrado no teste (2).
2. Aplica-se um método de seleção de itens a partir de determinada abordagem metodológica (3). Exibem-se os itens de acordo com as respostas do examinado e seu grau de proficiência (4).
3. Estima-se a proficiência do respondente (5) e decide-se pela finalização ou não do teste de acordo com os parâmetros do(s) critério(s) de parada/finalização do teste (6). Se satisfeito os parâmetros, então fim de teste (7). Se não, o algoritmo volta ao passo 3.

Figura 1. Fluxograma com passo a passo metodológico de aplicação de um CAT.



Fonte: Adaptado de Aluisio e Piton-Gonçalves (2015).

Neste contexto, em conformidade com o exposto na Figura 1, os procedimentos metodológicos para aplicação de um CAT fundamentam-se, resumidamente, consoantes aos seguintes tópicos (Aluisio; Piton-Gonçalves, 2015; Alves, 2018; Baylé; Desmarais, 2020; Catalani, 2019; Eggen, 2022; Jaloto; Primi, 2024; Magalhães; Alves; César, 2022; Travitzki *et al.*, 2020; Van der Linden; Glass, 2010):

- A) **Banco de dados que contenha os itens do teste e outros dados associados, tais como as respostas corretas dos itens, e parâmetros psicométricos pré-calibrados pela TRI.** Dentre os diferentes modelos psicométricos que podem ser abordados pela TRI para estimação dos parâmetros dos itens, o modelo logístico de 3 parâmetros (implementado pelo ENEM, e o utilizado no presente trabalho) utiliza as métricas que informam o grau de dificuldade do item (refere-se ao nível de habilidade/proficiência necessário para acertar ou endossar o item), discriminação do item (refere-se à capacidade do item de discriminar entre indivíduos com diferentes habilidades), e o parâmetro relacionado à resposta ao acaso, isto é, o “chute” ao item. A calibração do banco de itens é realizada por algoritmos baseados em TRI, e necessita para isso de um banco robusto de respostas aos itens, isto é, de um quantitativo relevante de respondentes (Aluisio; Piton-Gonçalves, 2015; Eggen, 2022). Devido a isso, as publicações na área de CAT com aplicações de algoritmos de TRI utilizam como banco de dados as respostas de avaliações em larga escala, tais como o ENEM e SAEB, que possuem um extenso número de respondentes. No *software* R, o pacote *mirt* realiza a

calibração do banco de itens de um teste, a partir das respostas dos respondentes (Chalmers, 2025).

- B) **Crítérios para inicialização do teste e seleção do primeiro item ao respondente.** A seleção do primeiro item pode ocorrer aleatoriamente (método mais difundido), ou baseada em alguma informação prévia do respondente para selecionar itens mais fáceis ou difíceis (por exemplo, a partir da informação a priori de escores em algum teste anteriormente aplicado). No *software* R, a partir do pacote/biblioteca *mirtcat*, o *default* é o critério randômico (Chalmers, 2024).
- C) **Crítérios de seleção de itens.** Métodos estatísticos e computacionais que interpretem o nível da informação do item ou do teste, com o propósito de selecionar o item de forma adaptada ao respondente, e que maximize a informação de acordo com a habilidade estimada (Eggen, 2022). Os modelos estatístico mais utilizados como critérios no *software* R são o Máxima Informação (em inglês *Maximum Information*- MI), Esperança a Posteriori (em inglês, *Bayesian Expected a Posteriori* – EAP), método de Kullback-Leibler (KL), dentre outros (Chalmers, 2024, 2025).
- D) **Crítérios de estimação das proficiências/habilidades.** Em conformidade com Aluisio e Piton-Gonçalves (2015) e Baylé e Desmarais (2020), a estimação das proficiências podem ser realizadas por diferentes métodos e abordagens. Os três métodos mais difundidos na literatura são a Máxima Verossimilhança (em inglês, *Maximum Likelihood* – ML), o Máximo a Posteriori (em inglês, *Bayesian Maximum a Posteriori* – MAP) e a Esperança a Posteriori (em inglês, *Bayesian Expected a Posteriori* – EAP), a depender de algumas características dos dados do teste em questão, principalmente do tempo computacional de processamento exigido.
- E) **Crítérios de finalização/parada do teste.** A finalização do teste e estimação da proficiência final do respondente dependerá das características do teste e respondente, dos modelos e métodos estatísticos adotados, do estresse/fadiga do respondente, dentre outros fatores (Aluisio; Piton-Gonçalves, 2015; Eggen, 2022). Um CAT poderá utilizar como critério de finalização/parada o erro

padrão de estimação da proficiência (mais comumente utilizado), o tempo máximo de realização do teste, ou quando o número pré-determinado de itens é atingido, dentre outros critérios que o desenvolvedor do teste estipular no algoritmo. No presente trabalho foi definido como critério o erro padrão de estimação.

F) **Design do CAT e desenvolvimento de interface HTML.** Na plataforma R é utilizado o pacote/biblioteca *shiny*, que gera as interfaces HTML, permitindo em tempo real a interação com o respondente do teste. A interface permite a aplicação dinâmica e personalizada do teste, exibindo os itens conforme o desempenho do respondente, e cálculo automático da proficiência (Ryan *et al.*, 2025).

No contexto da presente pesquisa, esta é de natureza aplicada e tem como objetivo o desenvolvimento e *design* de um CAT baseado em dados reais do ENEM, utilizando a plataforma R como ferramenta computacional. Para isso, foram implementados os pacotes/bibliotecas *mirt*, *mirtCAT* e *shiny*, específicos para a modelagem psicométrica, calibração dos parâmetros do banco de dados e *design* do CAT, e que são amplamente adotados na literatura contemporânea sobre o tema, conforme abordado acima.

As etapas metodológicas abordadas na pesquisa, em conformidade com o exposto na referida seção, estão elencadas nas subseções a seguir.

Etapa 1- Preparação dos dados e calibração dos parâmetros dos itens

Nesta etapa foi realizada a extração e preparação do banco de dados do ENEM das edições dos anos de 2021, 2022 e 2023, a partir dos microdados públicos disponíveis no site do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), contendo respostas de candidatos a itens de múltipla escolha da área de Linguagens e Códigos (LC) (Brasil, 2024). Foi definido aleatoriamente o banco de dados da prova Azul. Em razão de limitações computacionais, para a etapa de cálculo e calibração dos parâmetros dos itens, fundamental para a construção do banco de itens do CAT, foi extraída uma amostra aleatória de respostas de 3.000 participantes (1.000 de cada edição), com o objetivo de assegurar a estimação estável dos parâmetros da TRI, a precisão das estimativas de proficiência e a seleção eficiente

dos itens ao longo da aplicação do CAT. Foram apenas considerados os respondentes que não apresentassem nenhuma resposta inconsistente (2 respostas ao mesmo tempo) ou em branco aos itens da prova de LC.

Foram selecionados todos os 40 itens que compõem a prova da área de LC de cada edição (foram retirados os 5 itens que compõem a disciplina de Línguas Estrangeiras- Inglês/Espanhol, portanto o banco de itens da área de LC, retirados os itens de Línguas, possuem o total de 40 itens), e eles foram calibrados a partir do modelo logístico de 3 parâmetros (3PL) pela abordagem da TRI. Portanto, o banco de itens do presente estudo compõem-se de 120 itens no total (40 itens de cada edição, de 2021 a 2023). A partir da análise do banco de itens, não houve questão repetida especificamente na área de LC nas três edições do ENEM utilizadas no presente trabalho.

A modelagem psicométrica foi conduzida a partir da implementação do pacote *mirt* do R, que possibilita a estimação dos parâmetros dos itens e das habilidades latentes por meio de métodos de máxima verossimilhança e estimação Bayesiana (Chalmers, 2025). O Inep passou a fornecer a partir das últimas edições do ENEM os parâmetros dos itens calculados por TRI nas planilhas dos microdados, no entanto, o pesquisador pode calcular os parâmetros, a partir do pacote *mirt*. Na presente pesquisa foi realizada a etapa de cálculo dos parâmetros dos itens a partir da TRI.

Na publicação de Jaloto e Primi (2024), utilizando os dados do ENEM de 2009 a 2019, a menor correlação (r) entre os parâmetros fornecidos e calculados pelos autores foi de 0,96 (estatisticamente significativas, com $p < 0,001$). No presente estudo o coeficiente de correlação foi de 0,812, na qual pode-se afirmar que esse valor, ainda que se utilizasse uma amostra mais representativa da população, não chegaria a 1 (correlação perfeita), uma vez que os dados da amostra utilizada diferem em escala da população original.

No presente trabalho, para dispor os parâmetros dos itens na mesma métrica utilizada pelo Inep, utilizou-se o método *sigma-mean*, e portanto, os dados possuem média 0 e desvio padrão de 1 (Jaloto; Primi, 2024).

Após a extração dos parâmetros dos itens, estes são salvos em uma planilha (cada coluna da planilha corresponde a um dos 3 parâmetros logísticos do modelo: dificuldade do item, discriminação e acerto ao acaso).

Para a inserção das questões do banco de itens na modelagem, de modo que possam ser apresentadas durante o teste, os pacotes do R permitem que essas questões sejam salvas em uma planilha, organizada de forma que cada coluna contenha: o enunciado da questão, as opções de respostas e a indicação da alternativa correta (esta última utilizada apenas pelo modelo como referência de gabarito, não aparecendo para o respondente durante o teste).

Assim, após o fornecimento dos arquivos com os parâmetros dos itens e das questões do teste, o algoritmo do pacote *mirtCat* e *shiny* estão com os dados de entrada definidos, possibilitando o desenvolvimento da etapa 2, a seguir.

Etapa 2- Modelagem e estimação

Nesta etapa são configurados, após a definição dos dados de entrada, os critérios de seleção dos itens e estimativa da proficiência dos respondentes. Os manuais de referência dos pacotes abordados pelo programa R explanam os tipos de critérios abordados em seus cálculos estatísticos e computacionais, dos quais dependem dos tipos de dados e modelos utilizados (Chalmers, 2024, 2025). Estes critérios podem ser alterados no próprio algoritmo dos pacotes. A presente pesquisa empregou para esta etapa o pacote *mirtCat*. Outro pacote fornecido pelo programa R é o *catR*, no entanto, para aplicações com TRI recomenda-se o pacote *mirtCat*.

Neste estudo adotaram-se como critérios de estimação o modelo de Máxima Informação (MI) para a seleção de itens, enquanto a proficiência dos examinados foi estimada pelo método de Esperança a Posteriori (EAP). O primeiro item aplicado a cada respondente foi selecionado de forma aleatória. O critério de finalização/parada do teste foi definido com base no erro padrão da medida da proficiência, estabelecido em 0,30 (Jaloto; Primi, 2024). Dessa forma, o CAT é concluído (critério de parada/finalização do teste) quando o erro padrão se torna inferior a 0,30 ou quando se atingem 40 itens, equivalente ao número máximo de itens do teste tradicional.

Etapa 3- Design e testes

Após a definição dos critérios do algoritmo para seleção adaptativa de itens e estimação das proficiências, com o uso do pacote *mirtCat*, o *design* experimental e implementação do teste adaptativo foi realizado na *framework* do pacote *shiny*. Este

pacote permite o desenvolvimento de interfaces *web* sem exigir amplo conhecimento de linguagem *HTML* ou *JavaScript*, ademais de ser interativo com o pacote *mirtCat*, ao conter uma função de servidor. Neste caso, a vantagem da utilização deste pacote *shiny* é a possibilidade de realizar as simulações do CAT no próprio servidor da plataforma R, não necessitando de um servidor externo (Chalmers, 2024, 2025; Ryan, 2025).

Para testar o funcionamento dos algoritmos implementados e a adequação da interface do CAT, nesta etapa foram realizados testes pilotos com 12 estudantes do 3º ano do ensino médio de uma escola pública. O tamanho reduzido da amostra nesta etapa decorreu do caráter exploratório desta fase do estudo, bem como de limitações operacionais e logísticas associadas à aplicação do CAT em ambiente computacional.

Com base nas etapas metodológicas previamente delineadas, foram coletados dados referentes aos testes experimentais, cujos resultados são apresentados na seção subsequente.

Resultados e Discussão

Os resultados apresentados referem-se ao processo de desenvolvimento e *design* experimental de um CAT proposto nesta pesquisa. Inicialmente são mostrados os resultados da calibração dos parâmetros dos itens e a implementação dos algoritmos adaptativos. Em seguida, apresentam-se os indicadores obtidos a partir dos testes experimentais, que incluem as métricas das proficiências estimadas dos participantes. Por fim, discute-se a efetividade da metodologia proposta em relação aos objetivos estabelecidos, evidenciando os avanços e desafios identificados ao longo do processo.

Neste contexto, a Figura 2 apresenta a imagem da tela com o *design* do CAT, e o enunciado de um item do banco de dados que foi selecionado durante um dos testes de CAT realizado, no caso, originalmente o item 6 da prova Azul de Linguagens e Códigos (LC) do ENEM do ano de 2023.

Figura 2 – Enunciado de um item do teste

Teste Adaptativo

Authors:
[Redacted Name]

Este teste é adaptativo. Cada questão se adapta ao seu nível de habilidade. Clique no botão abaixo para começar. Responda as questões de múltipla escolha selecionando apenas uma resposta. Após finalizar o teste, feche a aba.

NA

A sessão do Comitê Olímpico Internacional (COI) aprovou uma mudança histórica e inédita no lema olímpico, criado em 1894 pelo Barão Pierre de Coubertin para expressar os valores e a excelência do esporte. Mais de 120 anos depois, o lema tem sua primeira alteração para ressaltar a solidariedade e incluir a palavra "juntos": mais rápido, mais alto, mais forte — juntos. A mudança foi aprovada por unanimidade pelos membros do COI e celebrada pelo presidente da entidade. (Disponível em: <https://ge.globo.com>- Texto adaptado). De acordo com o texto, a alteração do lema olímpico teve como objetivo a:

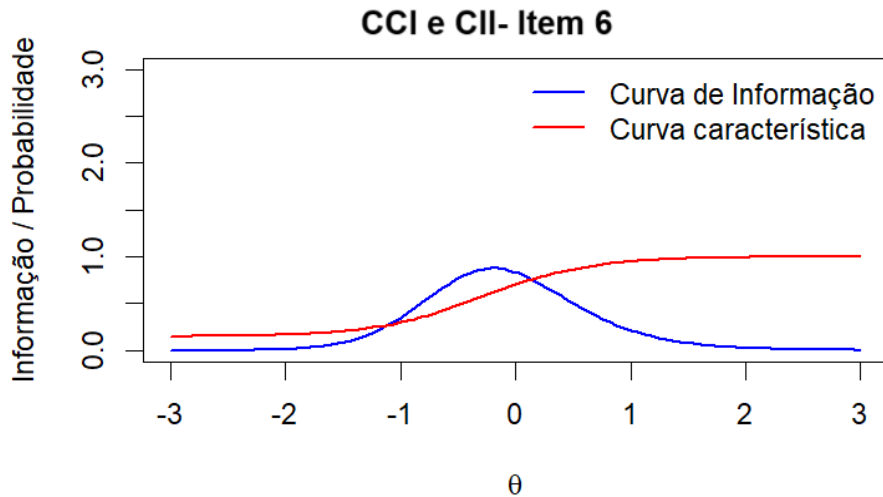
- unificação do lema anterior ao atual.
- aproximação entre o lema olímpico e o COI.
- junção do lema olímpico com os princípios esportivos.
- associação entre o lema olímpico e a cooperatividade.
- vinculação entre o lema olímpico e os eventos atléticos.

Fonte: Autores (2025)

Na abordagem da TRI, as curvas de informação do teste (CIT) representam a quantidade de informação que o teste e seus itens fornecem para diferentes níveis de habilidade dos respondentes. Cada item do teste possui sua curva de característica (CCI), que é uma representação gráfica que mostra a probabilidade de um respondente acertar ou endossar um item em função de sua habilidade. Também possui uma curva de informação individual (CII), que indica em quais níveis de habilidade o item é mais discriminativo e útil para estimar a proficiência em diferentes pontos da escala de habilidade/proficiência. Logo, a CIT, por sua vez, resulta da soma dessas informações individuais (CII), refletindo a precisão geral do instrumento ao longo da escala latente de proficiência. Essas curvas são fundamentais para avaliar a eficiência e a confiabilidade do teste, identificando as regiões onde a medição é mais precisa para estimação da proficiência, e auxiliando na seleção e no desenvolvimento dos itens.

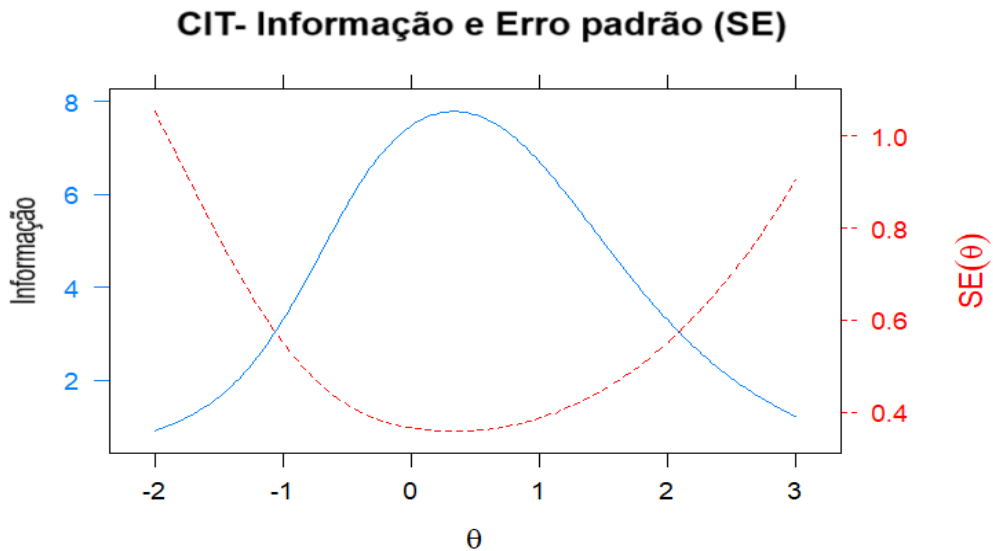
Na Figura 3 serão apresentadas a CCI e CII correspondente a um item aleatório do CAT (originalmente corresponde ao item 6 da prova de LC do Enem 2023). Enquanto isso, na Figura 4 ilustra-se a CIT, e portanto, permite-se visualizar a distribuição da informação tanto no nível do item quanto do teste como um todo.

Figura 3 – CCI e CII de um item do CAT aplicado



Fonte: Autores (2025)

Figura 4 – CIT do CAT aplicado



Fonte: Autores (2025)

Na CII da Figura 3, a curva apresenta picos mais altos entre os níveis de proficiência de -1 a 1, o que significa que o item é mais informativo e fornece maior precisão na estimativa da habilidade justamente nessa faixa (proficiência média corresponde ao 0 e desvio padrão 1, que é a escala padrão da TRI). Ou seja, o item é mais eficaz para diferenciar e avaliar os respondentes que possuem níveis de proficiência próximos a essa região média, oferecendo uma medição mais confiável nessa parte da escala. Enquanto isso, a CCI apresenta a curvatura acentuada próxima à medida de proficiência -1, o que indica que o item é mais sensível para avaliar respondentes com habilidade abaixo da média, ou seja, significa que a probabilidade

de acerto ao item aumenta para níveis de proficiência acima da medida de -1, no entanto, não é um item sensível para diferenciar respondentes com proficiências acima da média.

Já a CIT, mostrada na Figura 4, cuja curvatura é mais acentuada entre os níveis de proficiência -1 e 2 na escala, indica que o teste fornece a maior quantidade de informação e precisão nessa faixa de habilidades/proficiências. Isto é, esta região é onde a curva atinge seu valor máximo, e que corresponde onde o teste fornece mais informação e maior precisão na estimativa da habilidade. Isso significa que o instrumento é mais confiável para estimar a proficiência de respondentes com níveis moderados próximos à média (-1 e 2 na escala), oferecendo menor erro padrão de medição (SE) nessa região da escala.

No intuito de demonstrar a aplicação CAT sob estudo, foram realizados alguns testes experimentais. No âmbito do presente estudo, até o momento foram realizadas a aplicação do CAT com 12 respondentes. Para estes respondentes, de acordo com seus respectivos níveis de proficiência apresentados durante a aplicação do teste, no domínio do constructo avaliado pelo mesmo, isto é, proficiência na área de LC abordada pelo ENEM, o CAT selecionou adaptativamente na média 25 itens, variando de no mínimo 20 e no máximo 29 itens.

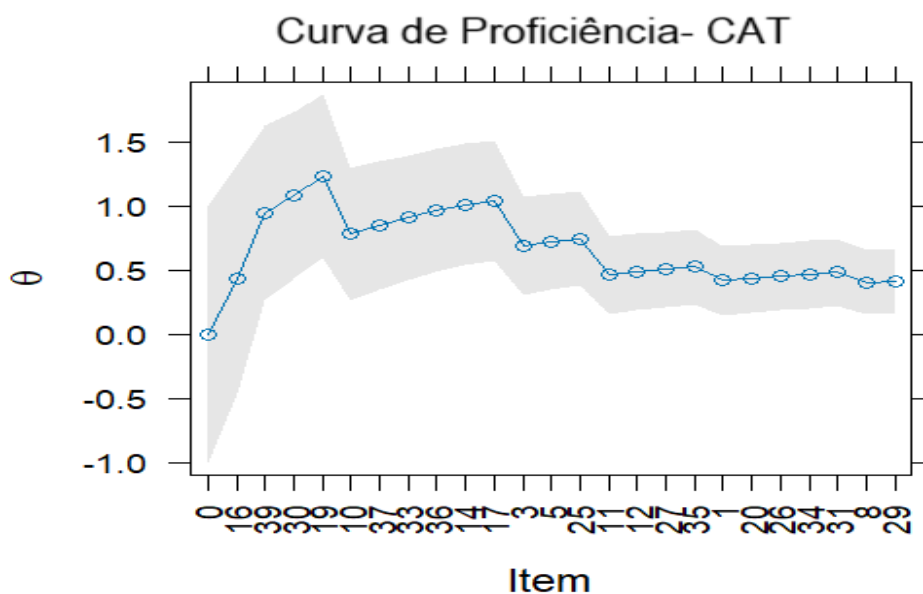
Essa seleção dinâmica de itens permitiu uma adaptação ao nível de proficiência dos respondentes, resultando em uma estimativa da proficiência no constructo avaliado, com uma redução significativa na quantidade de itens aplicados, e conseqüentemente diminuição da fadiga dos respondentes, uma vez que na prova tradicional, o quantitativo de itens da área de LC são 40 (retirados os 5 itens de Língua Estrangeira). O modelo ajustou continuamente a seleção dos itens visando maximizar a informação obtida, e redução do erro de medida, finalizando o teste quando o critério de parada/finalização do teste foi atingido. Nestes casos, houve redução de aproximadamente 37% (15 itens) no quantitativo de itens respondidos. Um aumento no número de itens fornecidos pelo CAT, por exemplo, fornecer sempre 40 itens, não diferenciou estatisticamente a estimação da proficiência em comparação com a redução adaptativa do número de itens.

No estudo de Jaloto e Primi (2024), os autores relataram que para 94,8% das aplicações simuladas de sua pesquisa, o teste do ENEM, em todas as áreas analisadas, poderia ter seus itens reduzidos para no máximo 20 itens, sem

comprometimento das estimativas de proficiências. A pesquisa de Spennassato *et al.* (2016) corrobora com os resultados obtidos na presente pesquisa, na qual os autores simularam a aplicação dos 45 itens de Matemática do ENEM da edição do ano de 2012 no formato de um CAT, e concluíram que um conjunto de 33 itens do CAT produziu resultados comparáveis com as proficiências reais obtidas a partir do teste tradicional.

A partir do exposto, na Figura 5 é apresentada a curva de proficiência de um dos respondentes do CAT aplicado na presente pesquisa. Para este respondente, de acordo com suas proficiências obtidas pelo algoritmo a partir de suas respostas aos itens, foram selecionados 24 itens, e sua proficiência final estimada foi de 0,41766. Esta curva da Figura 5 ilustra a estimativa do nível de habilidade do respondente ao longo das etapas do teste, permitindo visualizar como o modelo ajusta a avaliação conforme as respostas aos itens são fornecidas.

Figura 5 – Curva de proficiência em LC de um respondente ao CAT aplicado



Fonte: Autores (2025)

Conforme mostrado na Figura 5, a estimativa final de proficiência do respondente foi obtida através dos modelos adotados da TRI. Ao longo da aplicação, os itens foram selecionados de forma adaptativa, com base nas respostas anteriores e na estimativa provisória de proficiência (θ) atualizada a cada interação. Esse processo utilizou o critério de seleção que priorizou itens que fornecessem maior

informação (critério MI, conforme abordado na seção anterior) para o nível estimado do respondente, de modo a maximizar a precisão da medida. Ao final da aplicação, quando o critério de parada/finalização estabelecido foi alcançado, isto é, ocorreu a redução do erro padrão de medida abaixo de um limiar (0,30), a estimativa final da proficiência (θ) foi calculada utilizando o conjunto de respostas obtidas a partir dos 24 itens selecionados adaptativamente, resultando no valor reportado como proficiência final do respondente.

No contexto de um CAT, a média zero da proficiência estimada não representa ausência de conhecimento, mas sim um ponto de referência definido na escala de medida adotada pelo modelo, que é a escala padrão da TRI (média 0 e desvio padrão 1). Essa escala, que pode ser transformada para uma métrica pedagógica, é construída de forma que o valor zero corresponda ao nível médio de habilidade da população utilizada na calibração dos itens do teste. Assim, ao interpretar os resultados, é possível converter esse valor para uma escala pedagógica mais familiar, por exemplo, de 0 a 500 pontos e desvio padrão 100 (a exemplo do ENEM), permitindo assim que professores, gestores e estudantes compreendam melhor o significado da proficiência em termos de desempenho esperado e objetivos de aprendizagem.

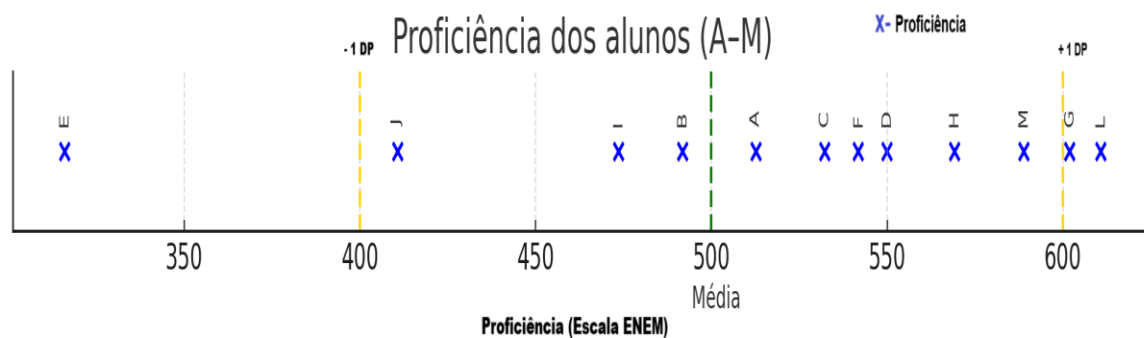
Neste sentido, a proficiência do respondente mostrada na Figura 5, a partir da escala do ENEM, seria de 541,766 pontos. Na escala de proficiência do ENEM (que possui média 500 e desvio padrão 100), o referido valor se localiza acima da média, aproximadamente 0,42 desvios-padrão acima de 500. Isso significa que, em uma interpretação pedagógica, esse desempenho se classifica como “acima da média”, ao superar a proficiência média da população de referência utilizada pelo ENEM.

Assim, o gráfico ilustrado na Figura 6 apresenta as proficiências estimadas dos 12 respondentes que participaram dos testes experimentais, identificados pelas letras de A a M. As proficiências estão representadas em uma escala padronizada, com média 500 e desvio padrão 100 (escala utilizada pelo ENEM). Cada ponto (x) no gráfico indica o valor da proficiência individual de cada respondente obtida no teste, permitindo, assim, visualizar as diferenças de desempenho entre eles.

Observa-se que alguns respondentes alcançaram proficiências acima da média, enquanto outros se situaram abaixo dela, evidenciando a variabilidade de resultados, mesmo em um grupo pequeno de participantes. Portanto, o gráfico apresentado na Figura 6, gerado a partir da aplicação do CAT, serve como uma

ferramenta pedagógica para compreender a distribuição das habilidades dos participantes e identificar padrões de desempenho individual.

Figura 6 – Gráfico com as proficiências dos respondentes estimadas pelo CAT aplicado



Fonte: Autores (2025)

Os resultados obtidos pela aplicação do CAT a 12 estudantes com itens da área de LC do ENEM permitiram estimar a proficiência dos respondentes com base nos modelos da TRI, ao mesmo tempo em que reduziram o número de itens administrados em relação ao exame tradicional. O comportamento adaptativo do algoritmo evidenciou a seleção de itens em função da habilidade estimada (θ), evitando a administração de itens excessivamente difíceis a estudantes com menor proficiência e ajustando progressivamente a dificuldade às respostas apresentadas.

O tamanho reduzido da amostra de respondentes justifica-se pelo caráter exploratório desta etapa da pesquisa, aliado a limitações operacionais e logísticas inerentes à aplicação do teste adaptativo em ambiente computacional. Nessa perspectiva, os resultados obtidos destinam-se prioritariamente à validação técnica e metodológica do sistema, não permitindo generalizações.

Isto posto, sob a perspectiva da avaliação mediadora proposta por Hoffmann (2014), o CAT, enquanto ferramenta mediada por TICs, configura-se como um recurso pedagógico capaz de acompanhar o processo de aprendizagem de forma contínua e diagnóstica, fornecendo informações relevantes para possíveis intervenções pedagógicas. Contudo, ressalta-se que limitações inerentes aos CATs (por exemplo o uso de ferramentas computacionais de alto desempenho para suas modelagens e o tamanho do banco de dados e amostras) restringem a generalização dos resultados,

indicando a necessidade de estudos futuros com amostras de respondentes mais amplas e diversificadas.

Considerações finais

Os resultados do referido trabalho indicam que o desenvolvimento e aplicação do CAT com itens da área de Linguagens e Códigos do ENEM possibilitou estimar de forma adaptativa a proficiência dos respondentes. A análise do desempenho do CAT demonstrou que os algoritmos adaptativos implementados selecionaram os itens adequados ao nível de habilidade de cada indivíduo, reduzindo o número de questões aplicadas, em comparação com o teste tradicional em papel. Dessa forma, a presente pesquisa evidencia que o CAT proporciona uma avaliação sensível às diferenças individuais dos participantes, cumprindo os objetivos de mensuração da proficiência de maneira personalizada, ao mesmo tempo em que minimiza a fadiga durante o teste e reduz a ocorrência de respostas aleatórias.

A coerência das estimativas do CAT foi definida com base em critérios psicométricos da TRI, considerando os parâmetros de discriminação/dificuldade e a função de informação dos itens (que são as métricas utilizada pelo ENEM), como também a redução do erro padrão da proficiência e estabilidade das estimativas ao longo do teste.

Para além da mensuração da proficiência, os achados contribuem para o debate sobre as avaliações educacionais com abordagens de TICs e IA. Neste contexto, o CAT mostrou-se capaz de identificar perfis de desempenho distintos entre os participantes, fornecendo subsídios para a compreensão dos níveis de aprendizagem e para a proposição de intervenções pedagógicas mais individualizadas e eficazes, fundamentadas em dados e evidências quantitativas.

No entanto, o presente estudo apresentou algumas limitações, especialmente quanto ao conjunto amostral utilizado, o que afetou na calibração dos parâmetros dos itens e conseqüentemente na precisão da medida da proficiência. Outra limitação se refere ao banco de itens, que pode ser considerado restrito em termos quantitativo e de qualidade nos níveis de informação dos itens (um banco de item extenso oferece maior qualidade na distinção das proficiências dos respondentes), limitando, portanto, a capacidade do CAT de distinguir precisamente a variação de proficiências dos participantes.

Além disso, salienta-se que a aplicação prática do CAT foi realizada em um contexto específico (amostra de 12 alunos do 3º ano do ensino médio), o que pode comprometer a generalização dos resultados para outras populações ou instituições de ensino.

Como etapas futuras para aprimoramento da referida pesquisa, tem-se a ampliação do banco de itens, incluindo um maior quantitativo de questões e itens com perfis de informação variados ao longo da escala de proficiência. Ademais da aplicação e testagem prática do CAT em contextos diversificados, a fim de aperfeiçoar a validade das estimativas de proficiência e avaliar o impacto do teste adaptativo sobre a motivação e o engajamento dos estudantes. Além disso, pesquisas futuras serão realizadas quanto à melhoria do *layout* e do *design* do CAT, visando uma experiência de usuário mais intuitiva, bem como na disponibilização do sistema em um servidor, a exemplo do *Moodle*, permitindo que seja acessado de forma independente e pública por estudantes, educadores e pesquisadores.

Referências

ALMEIDA, M. E. B.; VALENTE, J. A. (Org.). **Tecnologia na escola: a entrada da inteligência artificial na educação**. São Paulo: Loyola, 2023.

ALUISIO, S. M.; PITON-GONÇALVES, J. Teste adaptativo computadorizado multidimensional com propósitos educacionais: princípios e métodos. **Ensaio: Avaliação e Políticas Públicas em Educação**, Rio de Janeiro, v. 23, n. 87, p. 389-414, 2015. Disponível em: <https://www.scielo.br/pdf/ensaio/v23n87/0104-4036-ensaio-23-87-389.pdf>. Acesso em: 11 nov. 2024.

ALVES, L. M. A. **Um estudo sobre testes adaptativos computadorizados baseados na Teoria da Resposta ao Item**. 2018. 94 f. Dissertação (Mestrado em Estatística) – Universidade Federal do Ceará, Fortaleza, 2018. Disponível em: <https://repositorio.ufc.br/handle/riufc/33278>. Acesso em: 12 nov. 2024.

BAYLÉ, S.; DESMARAIS, M. C. Artificial intelligence in computerized adaptive testing: review and new prospects. **International Journal of Artificial Intelligence in Education**, [S.l.], v. 30, n. 4, p. 589–622, 2020. Disponível em: <https://link.springer.com/article/10.1007/s40593-020-00218-3>. Acesso em: 11 dez. 2024.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP. **Exame Nacional do Ensino Médio**. Documentação técnica do Enem e Microdados. Brasília: INEP, 2024.

CARVALHO, S. M. P. de; SANTOS, M. A. B. dos. Tecnologias digitais, mocinhas ou vilãs? olhares sobre o impacto na cognição dos estudantes. **Educitec - Revista de Estudos e Pesquisas sobre Ensino Tecnológico**, Manaus, v. 7, p. e126121, 2021. Disponível em: <https://doi.org/10.31417/educitec.v7.1261>. Acesso em: 11 dez. 2025.

CATALANI, E. M. T. **Teste adaptativo informatizado da Provinha Brasil: a construção de um instrumento de apoio para professores(as) e gestores(as) de escolas.** 2019. 282 f. Tese (Doutorado em Educação) – Faculdade de Educação, Universidade de São Paulo, São Paulo, 2019.

CHANG, H.; ZHANG, J. Recent advances in computerized adaptive testing with applications in educational assessment. **Educational Measurement: Issues and Practice**, [S.l.], v. 40, n. 2, p. 3–14, 2021. Disponível em: <https://onlinelibrary.wiley.com/journal/17453992>. Acesso em: 12 dez. 2024.

CHALMERS, R. P. *mirt: Multidimensional Item Response Theory*. R package version 2.3.0, 2025. Disponível em: <https://cran.r-project.org/package=mirt>. Acesso em: 11 mar. 2025.

CHALMERS, R. P. *mirtCAT: Computerized Adaptive Testing with the 'mirt' Package*. R package version 1.6.0, 2024. Disponível em: <https://cran.r-project.org/package=mirtCAT>. Acesso em: 11 mar. 2025.

COSTA, T. F. F. **Testes adaptativos informatizados (TAI) e desafios da avaliação da proficiência em leitura nos anos iniciais do Ensino Fundamental: um modelo de análise de alternativas de itens de múltipla escolha como contribuição para o sucesso escolar.** 2023. Dissertação (Mestrado em Educação) – Universidade de São Paulo, São Paulo, 2023.

EGGEN, T. J. H. M. Psychometric developments in computerized adaptive testing: a review. **Psychometrika**, [S.l.], v. 87, p. 637–659, 2022. Disponível em: <https://doi.org/10.1007/s11336-022-09852-w>. Acesso em: 10 nov. 2024.

FERNANDES, P. G. **Testes adaptativos computadorizados como um processo de decisão markoviano: equilíbrio ótimo entre eficiência e precisão.** 2022. 201 f. Tese (Doutorado em Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022. Disponível em: https://bdtd.ibict.br/vufind/Record/USP_dc0aacd41eb9cfa241fad78b9e79d2b9. Acesso em: 16 fev. 2025.

HOFFMANN, J. **Avaliação mediadora: uma prática em construção da pré-escola à universidade.** 33. ed. Porto Alegre: Mediação, 2014.

JALOTO, A.; PRIMI, R. Enem de próxima geração com menos itens e alta confiabilidade usando CAT. **Estudos em Avaliação Educacional**, São Paulo, v. 35, e10142, 2024. Disponível em: https://doi.org/10.18222/eae.v35.10142_port. Acesso em: 11 fev. 2025.

KLEIN, R. Alguns aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências. **Estudos em Avaliação Educacional**, São Paulo, v. 21, n. 78, p. 35–56, 2013. Disponível em: http://www.scielo.br/pdf/ensaio/v21n78/aop_0213.pdf. Acesso em: 11 jan. 2025.

LIN, C. *et al.* Application of machine learning algorithms to online adaptive testing with real-time feedback. **British Journal of Educational Technology**, [S.l.], v. 51, n. 6, p. 2193–2211, 2020. Disponível em: <https://onlinelibrary.wiley.com/journal/14678535>. Acesso em: 14 dez. 2024.

LUCKESI, C. C. **Avaliação da aprendizagem escolar: estudos e proposições.** São Paulo: Cortez, 1998.

MAGALHÃES, T. R.; ALVES, A. M. S.; CÉSAR, C. M. Desenvolvimento de um teste adaptativo informatizado com pacotes R. **Revista Brasileira de Informática na Educação**, [S.l.], v. 30, n. 1, p. 162–183, 2022. Disponível em: <https://sol.sbc.org.br/journals/index.php/rbie/article/view/21198>. Acesso em: 19 fev. 2025.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria da resposta ao item – TRI. **Avaliação Psicológica**, [S.l.], v. 2, n. 2, p. 99-110, 2003. Disponível em: https://arquivos.ufrj.br/arquivos/2023038199fe6037813060b8467bb780aTRI_pasquali.pdf. Acesso em: 11 nov. 2024.

PEREIRA, J. **Testes adaptativos computadorizados: uma abordagem prática utilizando o software R**. 2020. 85 f. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2020. Disponível em: <https://lume.ufrgs.br/handle/10183/214913>. Acesso em: 19 jan. 2025.

PITON-GONÇALVES, J. **Desafios e perspectivas da implementação computacional de testes adaptativos multidimensionais para avaliações educacionais**. 2012. 153 f. Tese (Doutorado) – Universidade de São Paulo, São Carlos, 2012. Acesso em: 13 dez. 2024.

PITON-GONÇALVES, J. Testes adaptativos para o Enade: uma aplicação metodológica. **Revista Meta: Avaliação**, [S.l.], v. 12, n. 36, p. 37–56, 2020. Disponível em: <https://revistas.cesgranrio.org.br/index.php/metaavaliacao/article/view/2735>. Acesso em: 14 fev. 2025.

RIBEIRO, D. C.; SILVA, M. P. Influência dos aspectos formativos e geracionais no uso das tecnologias digitais na prática pedagógica: o que dizem os professores?. **Educitec - Revista de Estudos e Pesquisas sobre Ensino Tecnológico**, Manaus, v. 8, p. e202022, 2022. Disponível em: <https://doi.org/10.31417/educitec.v8.2020>. Acesso em: 10 dez. 2025.

RYAN, W. *et al.* **Shiny: Web Application Framework for R**. R package version 1.7.8, 2025. Disponível em: <https://cran.r-project.org/package=shiny>. Acesso em: 14 mar. 2025.

SOUSA, D. **Comparação entre testes adaptativos computadorizados e testes tradicionais em larga escala no Brasil**. 2021. 123 f. Dissertação (Mestrado em Avaliação de Políticas Públicas) – Universidade Federal do Ceará, Fortaleza, 2021. Disponível em: <https://repositorio.ufc.br/handle/riufc/57987>. Acesso em: 20 mar. 2025.

SPENASSATO, D. *et al.* Testes adaptativos computadorizados aplicados em avaliações educacionais. **Revista Brasileira de Informática na Educação**, [S.l.], v. 24, n. 2, p. 1-12, 2016. Disponível em: <http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/6416>. Acesso em: 15 mar. 2025.

TABAK, G. C. *et al.* Teste adaptativo multiestágio para o ENEM. **Revista Brasileira de Informática na Educação**, [S.l.], v. 31, p. 1–18, 2023. Disponível em: <https://journals-sol.sbc.org.br/index.php/rbie/article/view/2529>. Acesso em: 11 dez. 2024.

TRAVITZKI, R. *et al.* Teste adaptativo informatizado da Provinha Brasil-Leitura: Resultados e perspectivas. **Estudos em Avaliação Educacional**, São Paulo, v. 31, n. 78, p. 525-553, 2020. Disponível em: <https://publicacoes.fcc.org.br/eae/article/view/7216>. Acesso em: 21 jan. 2025.

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

Recebido: 29/08/2025

Aprovado: 18/02/2026

Publicado: 26/02/2026

Como citar (ABNT): ANABUKI, E. T.; SOARES, T. M.; OLIVEIRA, R. R. A. de. Design e implementação de Teste Adaptativo Computadorizado utilizando dados do ENEM. **Educitec - Revista de Estudos e Pesquisas sobre Ensino Tecnológico**, Manaus, v. 12, e277126, 2026.

Contribuição de autoria:

Erika Tiemi Anabuki: Conceituação, Investigação, Análise Formal, Conceituação, Curadoria de Dados, Metodologia, Software, Validação, Visualização, Escrita (revisão e edição).

Tufi Machado Soares: Conceituação, Investigação, Metodologia, Recursos, Supervisão, Escrita (revisão e edição).

Rafaela Reis Azevedo de Oliveira: Conceituação, Investigação, Metodologia, Recursos, Supervisão, Escrita (revisão e edição).

Editor responsável: Iandra Maria Weirich da Silva Coelho

Direito autoral: Este artigo está licenciado sob os termos da Licença Creative Commons-Atribuição 4.0 Internacional.

