Alexandre Jaloto

Um Enem mais curto, preciso e seguro: desenvolvimento de uma aplicação em formato de Testagem Adaptativa Computadorizada

Apoio:

Campinas

2023

Alexandre Jaloto

Um Enem mais curto, preciso e seguro: desenvolvimento de uma aplicação em formato de Testagem Adaptativa Computadorizada

Tese apresentada ao Programa de Pós-Graduação *Stricto Sensu* em Psicologia da Universidade São Francisco, Área de Concentração – Avaliação Psicológica, para obtenção do título de doutor

Orientador: Prof. Dr. Ricardo Primi

Campinas

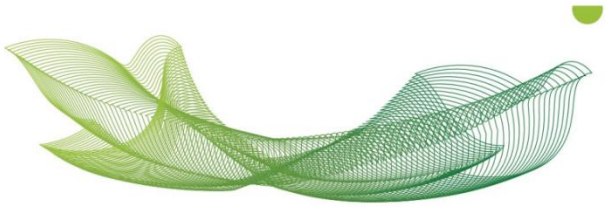
2023

157.93 Jaloto, Alexandre.
J27e Um Enem mais curto, preciso e seguro: desenvolvimento de uma aplicação em formato de Testagem Adaptativa Computadorizada / Alexandre Jaloto. – Campinas, 2023.
156 p.

Tese (Doutorado) – Programa de Pós-Graduação *Stricto Sensu* em Psicologia da Universidade São Francisco.
Orientação de: Ricardo Primi.

1. Psicometria. 2. Avaliação educacional. 3. Testagem adaptativa informatizada. I. Primi, Ricardo. II. Título.

Sistema de Bibliotecas da Universidade São Francisco - USF
Ficha catalográfica elaborada por: Tatiana Santana Matias - CRB-08/8303



Educando
para a paz

PROGRAMA DE PÓS-GRADUAÇÃO STRICTO SENSU EM PSICOLOGIA

Alexandre Marques Jaloto Rego defendeu a tese **"UM ENEM MAIS CURTO, PRECISO E SEGURO: DESENVOLVIMENTO DE UMA APLICAÇÃO EM FORMATO DE TESTAGEM ADAPTATIVA COMPUTADORIZADA"** **aprovado** pelo Programa de Pós-Graduação Stricto Sensu em Psicologia da Universidade São Francisco em 28 de fevereiro de 2023 pela Banca Examinadora constituída por:

Prof. Dr. Ricardo Primi
Orientador e Presidente

Prof. Dr. Alexandre José de Souza Peres
Examinador

Profa. Dra. Denise Reis Costa
Examinadora

Prof. Dr. Evandro Morais Peixoto
Examinador

Prof. Dr. Felipe Valentini
Examinador

Profa. Dra. Mariana Cúri
Examinadora

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Dedico este trabalho às pessoas que existem e são valiosas, nos termos do atual Ministro dos Direitos Humanos e Cidadania, Silvio Almeida.

Agradecimentos

Aviso de *spoiler*: é um agradecimento *nerd*.

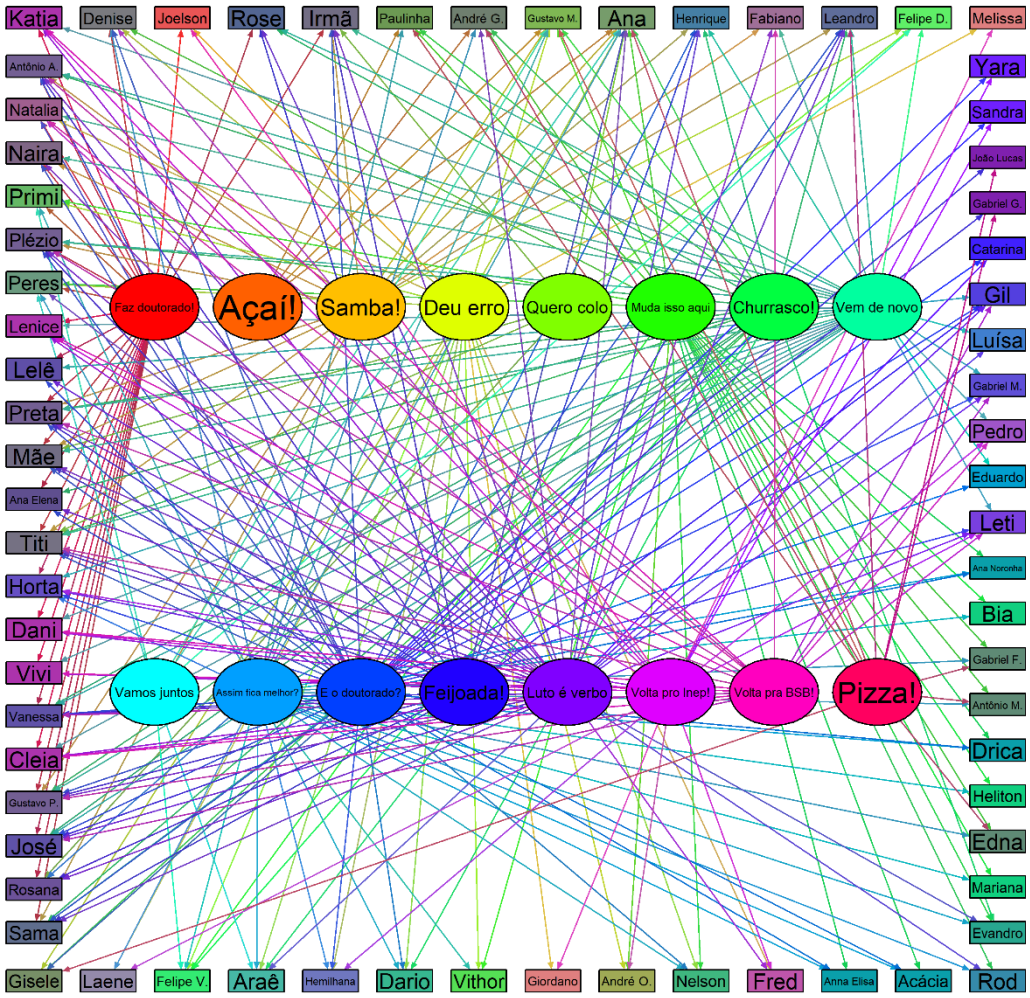
Pra mim, esta parte da tese é a mais gostosa e a mais difícil de se fazer. A mais gostosa porque é quando lembramos de todo o percurso, todas as alegrias, dificuldades, apoios, conselhos, ganhos e perdas. É a mais difícil porque certamente deixamos de citar uma pessoa que deveria ser mencionada. Não está sendo diferente agora.

Meus agradecimentos serão pelo que vivi do doutorado. Isso inclui a fase anterior, quando ganhei muito apoio para a seleção (Faz doutorado!). Os quatro anos do curso foram repletos de memórias boas, foram encontros e reencontros presenciais e à distância. Churrasco, Açaí, Samba, Feijoada e Pizza fizeram esses quatro anos de doutorado mais leves. E por que não pensar que os encontros prévios também me fortaleceram? Obrigado por esses momentos leves. Claro que me fortaleci não só no lazer, mas também no processo de escrita. As contribuições para a tese foram fundamentais, e essas memórias (Muda isso aqui, Assim fica melhor?) são ótimas. Recebi inúmeras leituras de texto e um bocado de ideias inovadoras nesses anos. Obrigado por me ajudarem a avançar academicamente. Aliás, até mesmo quem às vezes só chegava com um "E o doutorado?" do nada também fez parte dessa jornada. Obrigado!

Claro que nem tudo são flores. E os erros no R? Obrigado a você que ouviu "Deu erro", seja durante o curso, seja antes disso. Tive seu apoio e de quem disse "Vamos juntos" no percurso. Essas memórias também caminham comigo. Outras memórias importantes são as dos reencontros nas diferentes cidades. É um misto de "Vem de novo" com "Volta pra BSB". Saber do acolhimento em diferentes lugares também me fortaleceu na jornada. Ainda, ter no horizonte a possibilidade de aplicar com vocês no Inep o que desenvolvi nesses anos foi um motivador. Por isso, "Volta pro Inep" também é uma memória boa. Aliás, é importante destacar a importância do afastamento para a conclusão desta tese nesta qualidade. O retorno que espero dar é significativamente maior do que o investimento que o Estado fez neste trabalho.

Também vivi, durante esses quatro anos, momentos difíceis em que falei “Quero colo”; e recebi. Não só pelo doutorado em si, mas pelo momento que todos nós vivemos nesses anos. A pandemia começou em 2020 e ainda não acabou. Somado a isso, tivemos um governo que negou a ciência e tentou silenciar a diversidade, as diferenças, a vida de muitos. Por isso, também quero agradecer às pessoas que diariamente me lembram que Luto é verbo, militar é verbo, e temer, jamais.

Agora a parte *nerd*. Para representar meu agradecimento, fiz um diagrama que mostra essas memórias que citei acima e as pessoas a quem as associo. As elipses representam aquilo que não é visto diretamente, ou seja, as memórias que tenho guardadas. Elas se manifestam por meio dos retângulos, que são as pessoas a quem agradeço por esta tese. Obrigado por nesses quatro anos concretizarem esses momentos que me marcaram. Obrigado por me darem as palavras e os gestos nesses anos. Sou grato por ser acolhido por tantas famílias, por tantos espaços, por tantas ideias.



Resumo

O Exame Nacional do Ensino Médio (Enem) é composto por uma redação e quatro provas de 45 itens de múltipla escolha: Ciências Humanas e suas Tecnologias (CH); Ciências da Natureza e suas Tecnologias (CN); Linguagens, Códigos e suas Tecnologias (LC); e Matemática e suas Tecnologias (MT). O exame é utilizado como processo de seleção para ingresso em cursos de ensino superior. Esse uso impõe desafios para o exame no seu formato, como: produzir notas precisas para uma população diversa, minimizar o efeito da posição do item sobre o desempenho e construir testes equivalentes. É possível avançar nesses desafios ao aplicar o Enem em formato de Testagem Adaptativa Computadorizada (em inglês *Computerized Adaptive Testing*, CAT). Portanto, o objetivo deste trabalho foi desenvolver uma CAT do Enem mais eficiente, precisa e segura do que o seu formato atual. Dividimos a tese em dois Artigos e dois Produtos. O Artigo 1 comparou distribuições amostrais na calibração de itens no modelo logístico de três parâmetros da Teoria de Resposta ao Item. Utilizamos informações das quatro provas do Enem 2020 para simular as respostas de 5.040 participantes sorteados a partir de três tipos de desenho amostral: aleatório, retangular e deslocado. Não houve diferença significativa entre os desenhos para o parâmetro de discriminação. A amostra deslocada recuperou os parâmetros de dificuldade em CH melhor do que a retangular e em CN, melhor do que a aleatória. As amostras deslocada e aleatória recuperaram os parâmetros de pseudochute melhor do que a retangular nas quatro provas. Os resultados não apontam para a prevalência de um tipo de amostra para calibrar os itens do Enem 2020. O Produto 1 consistiu em um pacote estatístico para simulação de CAT em ambiente R. O Artigo 2 avaliou o método de controle de exposição progressivo restrito (PR) com diferentes parâmetros de aceleração em uma CAT em termos de eficiência, precisão e segurança. Manipulamos o método de seleção de itens (Aleatório, Máxima Informação de Fisher – MIF – e PR com dois parâmetros de aceleração) e o critério de parada (Tamanho fixo de 20 e 45 itens, Erro padrão de 0,30 e Redução do erro de 0,015 com erro padrão de 0,30) e simulamos a aplicação de 16 condições de CAT para cada prova. Por último, simulamos a aplicação do Enem 2020 em formato linear e comparamos com a CAT de tamanho fixo de 20 itens. O tamanho da prova foi maior com MIF. Nas CATs de tamanho fixo e com critério de parada do erro padrão, MIF e PR (com ambos os parâmetros de aceleração) tiveram resultados parecidos para a precisão. Com o critério de redução do erro, o PR performou pior. A segurança aumentou conforme o parâmetro de aceleração aumentou. A versão adaptativa do Enem teve precisão maior do que a versão linear. O Produto 2 da tese foi a publicação de um aplicativo da CAT Enem com o algoritmo determinado no Artigo 2. Concluímos que é possível reduzir o tamanho do Enem e melhorar sua precisão e segurança com uma CAT.

Palavras-chave: psicometria; avaliação educacional em larga escala; testagem adaptativa informatizada

Abstract

The Brazilian High School Exam (Enem) consists of a written essay and four tests with 45multiple-choice items: Human Sciences and their Technologies (HS); Natural Sciences and their Technologies (NS); Languages, Codes and their Technologies (LC); and Mathematics and their Technologies (MT). The exam is used as a selection process for entry into higher education courses. This use poses challenges for the exam in its format, such as: producing precise scores for a diverse population, minimizing the effect of item position on performance, and constructing equivalent tests. It is possible to advance in these challenges by applying Enem in a Computerized Adaptive Testing (CAT) format. Therefore, the objective of this work was to develop a more efficient, precise, and secure Enem CAT than its current format. We divided the thesis into two Articles and two Products. Article 1 compared sample distributions in the calibration of items in the three-parameter logistic model of Item Response Theory. We used information from the four Enem 2020 tests to simulate the responses of 5,040 participants drawn from three types of sampling designs: random, rectangular, and shifted. There was no significant difference between the designs for the discrimination parameter. The shifted sample recovered the difficulty parameters in HS better than the rectangular sample and in NS better than the random sample. The shifted and random samples recovered the pseudo-guessing parameters better than the rectangular sample in all four tests. The results do not point to the prevalence of one type of sample to calibrate the Enem 2020 items. Product 1 consisted of a statistical package for simulating CAT in an R environment. Article 2 evaluated the method of progressive restricted exposure (PR) control with different acceleration parameters in a CAT in terms of efficiency, precision, and security. We manipulated the item selection method (Random, Maximum Fisher Information - MFI - and PR with two acceleration parameters) and the stopping criterion (Fixed length of 20 and 45 items, Standard Error of 0.30 and Error Reduction of 0.015 with Standard Error of 0.30) and simulated the application of 16 CAT conditions for each test. Finally, we simulated the application of Enem 2020 in a linear format and compared it to the 20-item fixed-length CAT. The test size was larger with MFI. In the fixed-length CATs and with the stopping criterion of standard error, MFI and PR (with both acceleration parameters) had similar results for precision. With the error reduction criterion, PR performed worse. Security increased as the acceleration parameter increased. The adaptive version of Enem had higher precision than the linear version. Product 2 of the thesis was the publication of an Enem CAT application with the algorithm determined in Article 2. We concluded that it is possible to reduce the size of Enem and improve its precision and security with a CAT.

Keywords: psychometrics; large-scale educational assessment; computerized adaptive testing

Resumen

El Examen Brasileño de Enseñanza Media (Enem) consta de una redacción y cuatro pruebas de 45 preguntas de opción múltiple: Ciencias Humanas y sus Tecnologías (CH); Ciencias de la Naturaleza y sus Tecnologías (CN); Lenguajes, Códigos y sus Tecnologías (LC); y Matemáticas y sus Tecnologías (MT). El examen se utiliza como proceso de selección para ingresar a cursos de educación superior. Este uso implica desafíos para el examen en su formato, como producir notas precisas para una población diversa, minimizar el efecto de la posición del ítem sobre el desempeño y construir pruebas equivalentes. Es posible avanzar en estos desafíos al aplicar el Enem en formato de Prueba Adaptativa Computarizada (Computerized Adaptive Testing, CAT en inglés). Por lo tanto, el objetivo de este trabajo fue desarrollar una CAT del Enem más eficiente, precisa y segura que su formato actual. Dividimos la tesis en dos artículos y dos productos. El Artículo 1 comparó distribuciones muestrales en la calibración de ítems en el modelo logístico de tres parámetros de la Teoría de Respuesta al Ítem. Utilizamos información de las cuatro pruebas del Enem 2020 para simular las respuestas de 5,040 participantes sorteados a partir de tres tipos de diseño muestral: aleatorio, rectangular y desplazado. No hubo diferencia significativa entre los diseños para el parámetro de discriminación. La muestra desplazada recuperó los parámetros de dificultad en CH mejor que la rectangular y en CN, mejor que la aleatoria. Las muestras desplazada y aleatoria recuperaron los parámetros de pseudochute mejor que la rectangular en las cuatro pruebas. Los resultados no apuntan a la prevalencia de un tipo de muestra para calibrar los ítems del Enem 2020. El Producto 1 consistió en un paquete estadístico para simulación de CAT en ambiente R. El Artículo 2 evaluó el método de control de exposición progresivo restringido (PR) con diferentes parámetros de aceleración en una CAT en términos de eficiencia, precisión y seguridad. Manipulamos el método de selección de ítems (Aleatorio, Máxima Información de Fisher – MIF – y PR con dos parámetros de aceleración) y el criterio de parada (Tamaño fijo de 20 y 45 ítems, Error estándar de 0,30 y Reducción del error de 0,015 con error estándar de 0,30) y simulamos la aplicación de 16 condiciones de CAT para cada prueba. Por último, simulamos la aplicación del Enem 2020 en formato lineal y comparamos con la CAT de tamaño fijo de 20 ítems. El tamaño de la prueba fue mayor con MIF. En las CATs de tamaño fijo y con criterio de parada del error estándar, MIF y PR (con ambos parámetros de aceleración) tuvieron resultados similares para la precisión. Con el criterio de reducción del error, el PR rindió peor. La seguridad aumentó conforme el parámetro de aceleración aumentó. La versión adaptativa del Enem tuvo una precisión mayor que la versión lineal.

Palabras clave: psicometría; evaluación educativa a gran escala; pruebas adaptativas informatizadas

Sumário

Lista de figuras

Lista de tabelas

Apresentação

Esta tese é um dos produtos que entrego decorrentes do curso de doutorado. Ela conta a história da produção de uma versão do Exame Nacional do Ensino Médio (Enem) em formato de Testagem Adaptativa Computadorizada (em inglês *Computerized Adaptive Testing*, CAT). O Enem é composto por quatro provas de 45 itens e uma redação, cuja aplicação ocorre em dois dias com duração de pelo menos cinco horas cada. Para além dos desafios psicométricos impostos ao Enem (como demonstrar a validade preditiva do exame, produzir notas precisas para uma população diversa, minimizar o efeito da posição do item sobre o desempenho e construir testes equivalentes), sua complexidade operacional é grande. Anualmente, pelo menos quatro milhões de pessoas se inscreveram no Enem entre 2009 e 2020, o que demandou a impressão, transporte e manuseio de centenas de milhões de páginas de um exame de alto impacto que requer a devida segurança. Posso afirmar que uma grande motivação desta tese é caminhar para avanços no Enem que facilitem a lida com sua grandiosidade em termos psicométricos e logísticos.

Desde meu ingresso no Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) em 2009, venho notando a carência de pesquisas com medidas educacionais no Brasil. Em especial, percebi uma lacuna de pesquisas com psicometria que utilizassem testes de larga escala como o Enem e que apontassem para possíveis avanços que dessem conta dos desafios do exame. Após oito anos de envolvimento direto com as atividades relacionadas ao Enem, dos quais dois dedicados às análises psicométricas, vi na CAT uma alternativa viável para aprimorar o Enem. Por isso, iniciei os estudos sobre aplicabilidade da CAT no maior teste padronizado brasileiro. O embrião desta tese surgiu em 2017, antes mesmo de eu saber que era uma ideia de tese, quando fiz minha primeira simulação em CAT com o pacote mirtCAT (Chalmers, 2016). Essa simulação me rendeu uma apresentação em congresso internacional

(Jaloto, 2018), além de conhecimento adquirido e o início de um relacionamento com o mundo da CAT.

Em 2020, no curso do doutorado, dei continuidade aos estudos de CAT aplicados ao Enem com a profundidade que o tema demanda, agora em parceria com o professor Ricardo Primi. Esses estudos geraram três novos trabalhos em congresso (Jaloto & Primi, 2021a, 2021b, 2022), sendo um premiado pela Associação Internacional de CAT (Iacat, em inglês *International Association for Computerized Adaptive Testing*) e um manuscrito que está em fase de preprint (Jaloto & Primi, 2023c). Esta tese, que é o produto maior do meu curso de doutorado, apresenta um recorte de todo o esforço e desejo de ampliar e compartilhar o conhecimento em CAT na área educacional, especialmente no Brasil. Eis outra motivação para este trabalho.

O objetivo geral desta tese foi desenvolver uma CAT do Enem mais eficiente, precisa e segura do que o formato atual do exame. A pergunta que respondemos foi: é possível reduzir o tamanho do Enem e melhorar sua precisão e segurança (em termos de exposição de itens) com uma CAT? Para atingir nosso objetivo, avaliamos a qualidade do banco de itens da CAT (formado por itens aplicados do Enem), desenvolvemos um pacote de simulação de CAT, determinamos o algoritmo da testagem e por último a publicamos na plataforma shinyapps.io. Portanto, esta tese se insere em três das cinco etapas da diretriz de Thompson e Weiss (2011) para o desenvolvimento de uma CAT: Pré-testagem, calibração e linkagem; Determinação da especificação final da CAT; e Publicação da CAT.

Dividimos a tese em dois artigos e dois produtos. O Artigo 1 objetivou comparar diferentes distribuições amostrais na calibração de itens no modelo logístico de três parâmetros da Teoria de Resposta ao Item. Nesse Artigo, avaliamos a calibração dos itens do Enem, o que contribuiu para avaliar a qualidade do banco de itens da CAT. O Produto 1 da tese foi o simCAT (Jaloto & Primi, 2023a), um pacote estatístico aberto desenvolvido em ambiente R (R Core

Team, 2019) voltado para simulação de CAT. As simulações realizadas nesta tese são apresentadas no Artigo 2, cujo objetivo foi avaliar o método de controle de exposição progressivo restrito com diferentes parâmetros de aceleração em uma CAT com base na eficiência (em termos de tamanho do teste), precisão e segurança. O Produto 2 da tese foi a publicação da CAT Enem (Jaloto & Primi, 2023b) em shiny (Chang et al., 2021) com as especificações determinadas a partir das simulações do Artigo 2. Uma vez que uma das motivações desta tese é compartilhar o conhecimento em CAT, e dado nosso compromisso com a ciência aberta, todos os códigos utilizados estão disponibilizados em repositórios do github. Desejamos uma boa leitura e esperamos que este trabalho possa contribuir para o avanço na área de medidas educacionais, especialmente em CAT.

Introdução

Desde 2009, o Exame Nacional do Ensino Médio (Enem), que é composto por uma redação e quatro provas de 45 itens de múltipla escolha, tem sido utilizado como parte ou processo único de seleção para ingresso em cursos de graduação de instituições de ensino superior. Esse uso do Enem impõe diversos desafios para o exame no seu formato, tais como: (1) apresentar capacidade preditiva do desempenho no ensino superior (validade preditiva); (2) selecionar uma amostra de itens que representem o domínio do conteúdo necessário para ingresso no ensino superior (validade de conteúdo); (3) minimizar o efeito da posição do item sobre o desempenho do participante; (4) adequar a confiabilidade para um intervalo da escala de cerca de sete unidades de desvio padrão, dada a ampla variação da nota de corte dos cursos; (5) evitar o compartilhamento de itens de uma aplicação (cola ou trapaça). Uma possibilidade de avançar em alguns desses desafios (3, 4 e 5) é aplicar o Enem em formato de Testagem Adaptativa Computadorizada (em inglês *Computerized Adaptive Testing*, CAT). Portanto, o objetivo deste trabalho foi desenvolver uma CAT do Enem mais eficiente, precisa e segura do que o formato atual do exame (linear, aplicado no formato de lápis e papel). Este estudo apresenta uma reflexão sobre a viabilidade técnica de se implementar uma CAT no Enem, haja vista seu potencial para aprimorar a logística de aplicação do exame e contemplar os desafios impostos. Ainda, o estudo avança no conhecimento sobre a CAT em contexto educacional ao simular aplicações com bancos de mais de 700 itens, além de desenvolver um pacote aberto para simulação de CAT e de publicar a versão do Enem em formato de CAT.

Enem

O Enem foi aplicado pela primeira vez em 1998 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), autarquia do Ministério da Educação (MEC, então Ministério da Educação e do Desporto). O exame surgiu como um modelo de avaliação que tinha como referência principal a articulação entre a Educação Básica e a cidadania (Inep,

2009). De caráter voluntário, o Enem tinha como objetivos (i) conferir ao cidadão um parâmetro para autoavaliação; (ii) criar referência nacional para os egressos de qualquer das modalidades do ensino médio; (iii) fornecer subsídios às diferentes modalidades de acesso à educação superior; e (iv) constituir-se em modalidade de acesso a cursos profissionalizantes pós-médio (Portaria n. 438, 1998).

Em 2009, com o objetivo de consistir no processo seletivo único para ingresso nos cursos de graduação de instituições federais, a estrutura do Enem foi alterada. Até 2008, o exame era composto por um teste de 63 itens interdisciplinares e após sua reformulação, sua estrutura tomou emprestado os fundamentos do Exame Nacional para Certificação de Competências de Jovens e Adultos (Encceja) Ensino Médio (MEC, 2009). Mais informações sobre o Encceja estão disponíveis no portal do Inep (2023a). Com a mudança, o Enem passou a se organizar em termos de áreas de conhecimento e passou a contar com a aplicação, além da redação, de quatro provas diferentes, a saber: Ciências Humanas e suas Tecnologias (CH); Ciências da Natureza e suas Tecnologias (CN); Linguagens, Códigos e suas Tecnologias (LC); e Matemática e suas Tecnologias (MT). Cada prova possui 45 itens, mede 30 habilidades agrupadas em um número variável de competências entre as áreas (Inep, 2009) e produz uma medida unidimensional por meio do modelo logístico de três parâmetros da Teoria de Resposta ao Item (3PL; Inep, 2021). Em LC, cinco das 30 habilidades se referem a uma língua estrangeira. Desde 2010, o participante escolhe entre inglês e espanhol e responde cinco itens dos 45 de LC nessa língua.

Desde a mudança do Enem em 2009, cada edição passou a contar com no mínimo duas aplicações. Em geral, a primeira é a que conta com maior número de participantes e é chamada de regular. A segunda normalmente é destinada a pessoas privadas de liberdade (PPL) e a participantes que tenham passado por alguma intempestividade na primeira aplicação, por exemplo acidentes naturais ou problemas logísticos do exame. Entre 2009 e 2020, cada edição

do Enem contou com no mínimo quatro milhões de inscritos. Nesse período, as edições de 2016 e 2020 contaram com três aplicações. Essas informações estão disponíveis nos microdados do Enem disponibilizados no portal do Inep (2023b).

Nesta tese, utilizamos a denominação *teste* para fazer menção ao instrumento utilizado para estimar a proficiência do sujeito, *prova* para nos referir ao conjunto de testes da área do conhecimento avaliada no Enem (CH, CN, LC e MT) e *aplicação* para mencionar o evento em que o sujeito é submetido ao teste. Adicionalmente, utilizamos *proficiência* para nos referir à variável latente medida na prova do Enem e *nota* para nos referir à magnitude da variável atribuída ao sujeito, ou seja, a representação numérica da proficiência. A seção do documento teórico do Enem que norteia a elaboração dos itens é denominada matriz de referência (Inep, 2009). A matriz de referência de cada prova traz uma lista de 30 descritores de habilidades agrupadas em competências. Este trabalho não pretende problematizar o conceito de competências e habilidades, mas focar na possibilidade do uso da CAT no Enem. Para uma discussão sobre a aplicação desses conceitos no Enem, sugerimos uma consulta a Primi et al. (2001).

Implicações de usar o Enem para seleção

Existem vários desafios associados ao uso do Enem para seleção em termos de confiabilidade e validade. Em primeiro lugar, é essencial demonstrar a validade do teste para predizer o desempenho no ensino superior. No entanto, são raros os estudos de validade preditiva de testes educacionais no Brasil (e.g., Moreira, 2017; Primi, 2006a, 2006b; Souza et al., 2013), e a escassez de estudos que exploram esse aspecto do Enem é ainda maior (e.g., Ferreira-Rodrigues, 2015). A segunda questão é a avaliação da equivalência das notas da redação, uma vez que os juízes diferem em seus níveis de leniência e severidade (Primi et al., 2019), o que não é levado em consideração na versão atual do exame. Por fim, o erro de medida é uma preocupação ao usar o teste com candidatos com níveis de proficiência muito diferentes.

Não sabemos quão grande é o erro em cada nível da escala, com alguns cursos tendo pontos de corte em níveis mais baixos da escala e outros em níveis mais altos. O Inep transforma e padroniza as notas do Enem para terem média 500 e desvio padrão 100, utilizando como referência as estatísticas da edição de 2009 (Inep, 2012).

Dada a grande variação nos pontos de corte de ingresso no ensino superior, os itens devem abranger um amplo espectro de dificuldade para se obter notas suficientemente confiáveis para decisões de seleção. Como exemplo, nos microdados da edição 2020 do Sistema de Seleção Unificada (SiSU 2020), que utilizou o Enem 2019, é possível verificar que a nota de corte para ingresso no curso de Ciências Sociais da Universidade Federal do Rio de Janeiro (UFRJ) para vagas reservadas para alunos negros de baixa renda com deficiência oriundos de escolas públicas foi 394,10. Já a menor nota para ingresso no curso de Medicina da UFRJ por meio de vagas de ampla concorrência foi 790,98. Nesse ano, as pontuações de corte variaram de 227,78 (Engenharia de Aquicultura na Universidade Federal do Paraná) a 928,30 (Medicina na Universidade Federal do Maranhão). Além disso, para se qualificar para bolsas e financiamentos públicos, como o Programa Universidade Para Todos (Prouni) e o Fundo de Financiamento Estudantil (Fies), é necessária uma média mínima de 450. Com base nesse intervalo de pontos de corte (227,78 a 928,30), fica evidente que as provas exigem confiabilidade adequada em uma faixa de mais de sete unidades de desvio padrão. Essas informações e outras informações sobre o histórico das edições do Enem e do SiSU estão disponíveis nos portais do Inep (2023c) e do MEC (2023).

O Inep busca atingir os objetivos do Enem por meio de testes com número fixo de 45 itens. No entanto, essa situação pode resultar em um dilema para quem os elabora. Quando se trata de um teste como o Enem, com número fixo de itens e que se propõe a medir uma população diversa em termos de magnitude da variável latente, depara-se com o dilema de *bandwidth-fidelity* (McBride, 1976). Um teste de pico fornece precisão satisfatória em um

pequeno intervalo da escala, o que aumenta a confiabilidade para as medidas dos sujeitos localizados nessa região. Para os demais sujeitos, em geral os itens serão muito difíceis ou muito fáceis e pouco contribuirão para a precisão de suas notas. Já um teste retangular fornece uma precisão relativamente igual em todas as regiões da escala, porém baixa. Assim, nesse caso precisa-se negociar entre confiabilidade e largura do espectro da escala. A CAT pode ser uma maneira de superar esse dilema, pois objetiva fornecer notas com precisão adequada para os diferentes níveis do espectro da escala.

Adicionalmente, a fadiga também é uma preocupação. Sabe-se que em testes educacionais em larga escala de alto impacto, a posição de uma pergunta pode afetar suas propriedades psicométricas (Domingue et al., 2020). No caso do Enem 2016, a posição dos itens de MT esteve associada ao desempenho dos participantes (Barichello et al., 2022). Os itens tiveram porcentagens de acerto diferentes de acordo com a posição em cada um dos quatro cadernos, que possuíam os mesmos 45 itens, porém em ordens distintas. A correlação entre a diferença de posição do item nos cadernos e a diferença de acerto nesse item foi de 0,33. Esses achados sugerem que a fadiga pode interferir no desempenho dos alunos.

Quando há diferenças individuais de desempenho e essas diferenças não se relacionam com o construto medido pelo Enem, isso pode resultar em uma questão de equidade para o teste. Alunos com níveis semelhantes de proficiência podem pontuar de forma diferente como resultado da fadiga, em vez da competência principal. A aplicação de uma CAT poderia mitigar o problema do tamanho de um teste e, consequentemente, reduzir os efeitos potenciais da fadiga. A promessa de uma aplicação em CAT é manter a confiabilidade de um teste com menos itens, ou até mesmo melhorá-la (Veldkamp & Matteucci, 2013; Weiss, 2011).

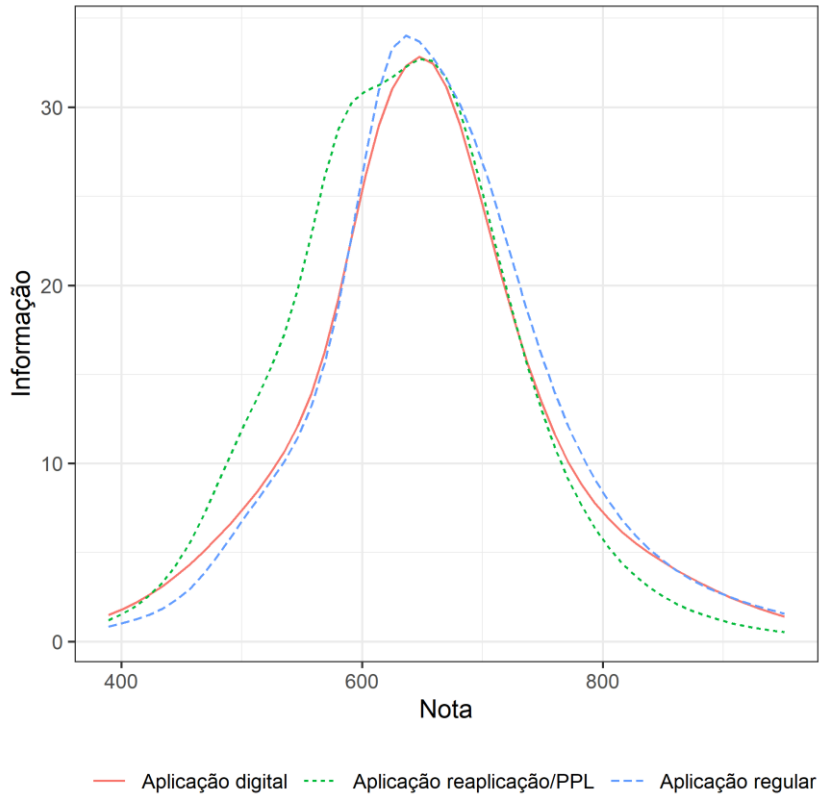
Outra questão relevante para o Enem se refere à montagem dos testes. Como indicado, cada edição do Enem possui pelo menos duas aplicações. Os testes de cada aplicação da mesma prova estão linkados (von Davier, 2011), ou seja, os itens estão posicionados na mesma métrica.

Isso sugere que os resultados das aplicações sejam comparáveis e as notas tenham as mesmas interpretações. No entanto, para que isso possa acontecer e os testes estejam de fato equalizados, eles precisam atender a especificações estatísticas semelhantes (Kolen & Brennan, 2014). Isso inclui garantir que o erro padrão condicionado à nota seja igual em ambos os testes. Para tal, é necessário que eles tenham curvas de informação semelhantes, o que nem sempre ocorre.

A título de ilustração, a Figura 1 apresenta as curvas de informação dos testes das três aplicações de CH do Enem 2020, a edição mais recente utilizada neste trabalho. Essas curvas foram construídas a partir dos parâmetros divulgados nos microdados do Enem. Nota-se que o pico de informação da Aplicação Regular é maior do que as demais, e a curva da Reaplicação/PPL possui mais informação na região mais fácil da escala do que as demais curvas. Como consequência, a incerteza (erro padrão) da nota calculada nas provas será diferente, ainda que os valores pontuais de medida sejam semelhantes.

Figura 1

Curvas de informação dos três testes de Ciências Humanas do Enem 2020



Outra possível consequência da montagem de provas diferentes é a diferença na quantidade de acertos necessários para atingir a mesma nota em aplicações diferentes. No modelo 3PL da Teoria de Resposta ao Item (TRI), acertar a mesma quantidade de itens de um teste resulta em notas diferentes, caso os itens sejam diferentes. Por exemplo, se o Sujeito A acertar somente os cinco itens mais fáceis de um teste, sua nota será maior do que a do Sujeito B, caso este último acerte somente os cinco mais difíceis desse teste. Isso tem relação com a forma de calcular a proficiência, que envolve a verossimilhança. Para o modelo, é mais verossímil que o Sujeito B tenha uma nota mais baixa caso erre os itens fáceis, ainda que acerte os difíceis.

A mesma lógica vale para testes cujos itens cubram amplitudes diferentes da escala, por exemplo se a proporção de itens fáceis for diferente nos testes. Isso tem relação com a menor diferença mensurável (*Least Measureble Difference*, LMD; Wright & Stone, 1979), que consiste no mais perto que dois sujeitos podem estar posicionados na escala, sem possuir a mesma proficiência. Ou seja, diz respeito à menor distância entre duas proficiências medidas em um teste. Em termos práticos, a LMD se refere ao menor aumento na proficiência ao se acertar um item. Quanto mais itens o teste possui em uma determinada região da escala (por exemplo, entre 400 e 450), menor será a LMD nessa amplitude. Como consequência, o sujeito precisará acertar mais itens para ter uma nota que supere essa amplitude (no caso, uma proficiência maior que 450). Se dois testes possuem quantidades diferentes de itens fáceis, eles terão diferentes LMD na região mais fácil da escala. Como o sujeito precisa acertar os itens fáceis para ter uma nota mais alta, e nesse caso a quantidade de itens fáceis é diferente nos testes, então a quantidade mínima de acertos para se alcançar determinado nível da escala varia de teste para teste.

A título de ilustração, a Tabela 1 apresenta a quantidade de itens mais fáceis de cada aplicação da prova de CH do Enem 2020 que precisam ser acertados para atingir alguns níveis

na escala do Enem. Para atingir o ponto de corte para concorrer ao Prouni e ao Fies (450), na Aplicação reaplicação/PPL e na Aplicação digital é preciso acertar os seis itens mais fáceis, e na Aplicação regular, os cinco mais fáceis. Já para atingir uma nota próxima de 660, na Aplicação reaplicação/PPL são necessários cinco acertos a mais do que na Aplicação regular, considerando os itens mais fáceis. Ainda, a diferença entre a nota máxima das aplicações chega a 40 pontos (0,4 unidade de desvio padrão).

Tabela 1

Quantidade de acertos necessários nos testes de Ciências Humanas do Enem 2020 para alcançar determinados níveis da escala, e a nota correspondente

Nível	Aplicação	Acertos	Nota*
450	Regular	5	450,2
	Reaplicação / PPL	6	467,9
	Digital	6	453,2
660	Regular	29	664,8
	Reaplicação / PPL	34	662,8
	Digital	31	667,1
Máximo	Regular	45	862,6
	Reaplicação / PPL	45	822,2
	Digital	45	856,4

Nota. *A nota do teste foi calculada considerando que o participante acertou os *n* itens mais fáceis, onde *n* é a quantidade de acerto

Diante do explicitado nos últimos cinco parágrafos, nota-se que existem pelo menos duas limitações no Enem devido à montagem dos testes. Uma é a possível fragilidade da equalização, dadas as diferenças nas curvas de informação. A outra é a diferença na quantidade de itens necessários para se obter a mesma nota em testes de aplicações diferentes. Cabe ressaltar que não se sabe o quanto a necessidade de acertar mais itens se relaciona com a fadiga do participante do Enem e consequentemente com a sua nota. Porém, necessariamente o participante precisará se dedicar a mais itens em um teste com tempo limitado para conseguir a mesma nota que ele conseguiria em outro, a depender do seu nível de proficiência.

A CAT pode ser utilizada para avançar nessas limitações. Uma vez que a administração dos itens é feita de acordo com as respostas anteriores, o algoritmo montará um teste único para cada sujeito. Essa personalização do teste possibilita que a precisão das notas seja similar ao longo de toda a escala, caso o banco de itens seja suficiente. Em um teste linear como o Enem, é comum que sua curva de informação tenha um pico em determinada região da escala, o que gera precisões melhores nessa região. Em uma aplicação adaptativa, é como se a curva de informação se aproximasse de uma reta horizontal, pois a precisão seria semelhante em todas as regiões da escala. Com isso, é possível aplicar testes diferentes para sujeitos diferentes com a mesma precisão e com a mesma quantidade de itens. Neste trabalho avaliamos a eficiência (em termos de tamanho do teste), a precisão (em termos de erro padrão de medida, diferença entre notas reais e simuladas e correlação entre essas notas) e a segurança (em termos de exposição de itens) de diferentes desenhos de CAT.

Testagem Adaptativa Computadorizada

A CAT vem sendo utilizada como um formato de aplicação de testes internacionais como o Exame de Registro de Graduação (GRE, em inglês *Graduate Record Examination*) e o Sistema de Informação de Medição de Resultados Relatados pelo Paciente (Promis, em inglês *Patient-Reported Outcomes Measurement Information System*). No contexto nacional, alguns trabalhos apontam para a potencialidade de sua aplicação em testes educacionais, como em exames de proficiência em língua estrangeira (Costa et al., 2009; Cúri & Silva, 2019; Karino et al., 2009), na Provinha Brasil (Travitzki et al., 2021) e no próprio Enem (Jaloto & Primi, 2023c; Spenassato et al., 2016).

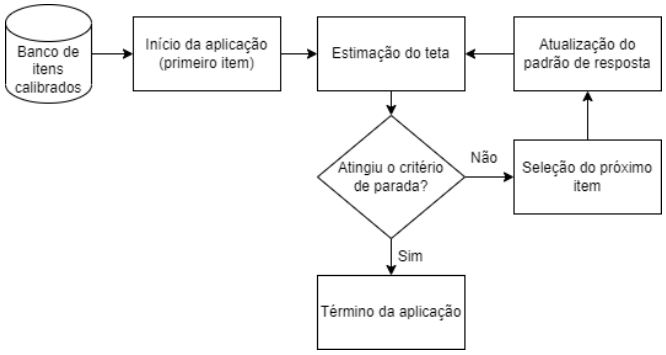
Em uma testagem adaptativa, os itens são administrados ao participante de acordo com sua resposta ao item anterior (Weiss & Kingsbury, 1984). A arquitetura de uma CAT é composta por cinco componentes (Peres, 2019; Thompson & Weiss, 2011; Weiss & Kingsbury, 1984):

1. banco de itens
2. início da aplicação
3. seleção dos itens
4. estimação do teta
5. critério de parada

A Figura 2 apresenta o funcionamento da uma aplicação em CAT. A partir de um banco de itens calibrados (componente 1), a CAT inicia com a seleção e administração de um item (componente 2). Após a aplicação desse item, o programa calcula o valor de teta (i.e., variável latente) do sujeito (componente 4). Em seguida, o algoritmo verifica se o critério de parada (componente 5) foi atingido. Caso negativo, o programa seleciona o próximo item a ser aplicado (componente 3), atualiza o padrão de respostas do sujeito (agora com dois itens) e calcula novamente o valor de teta (componente 4). Esse ciclo se mantém até que o critério de parada (componente 5) seja atingido. Quando isso ocorre, a aplicação encerra.

Figura 2

Arquitetura de uma Testagem Adaptativa Computadorizada



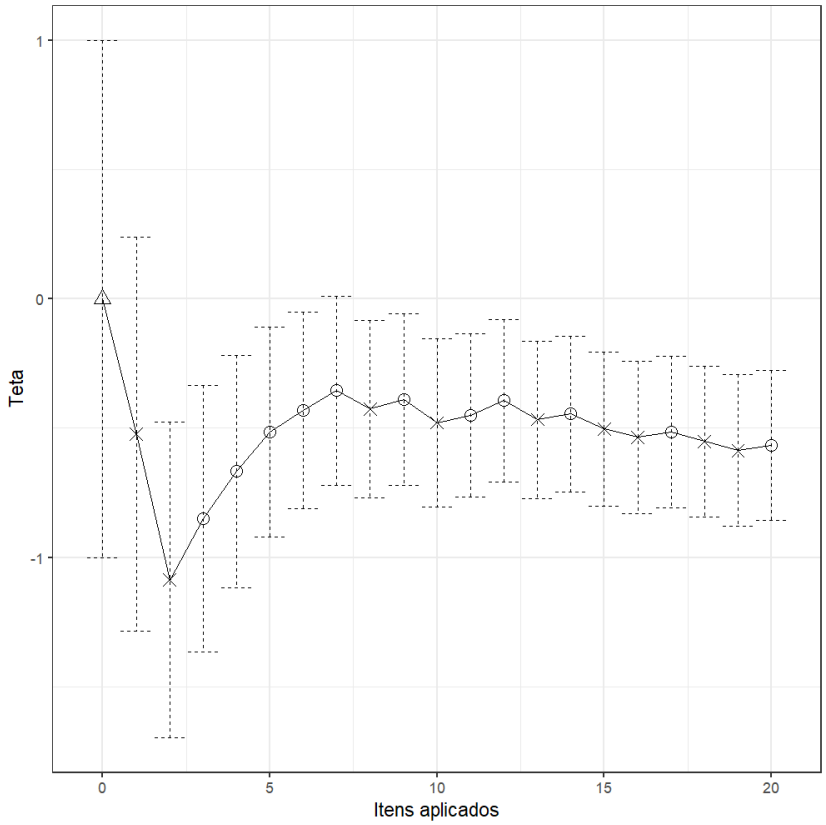
Fonte: elaboração própria

A Figura 3 ilustra a aplicação de uma CAT a um sujeito fictício. Em cada ponto, o acerto é representado por um círculo e o erro é representado por um "X". O triângulo representa o início da aplicação, quando nenhum item foi administrado. O segmento vertical tracejado tem o comprimento do dobro do valor do erro padrão para o teta provisório. A aplicação inicia

selecionando o item mais informativo para o teta 0, valor determinado arbitrariamente. Nesse exemplo, o sujeito erra esse item, o programa atualiza o seu padrão de respostas, calcula novamente seu teta provisório (−0,52) e seleciona o item mais informativo para esse ponto. O sujeito erra o segundo item e o programa seleciona o item para o teta provisório (−1,09), que é acertado. Conforme a aplicação progride, o erro padrão diminui, bem como a diferença entre os tetas provisórios e a redução do erro a cada item administrado. Nesta ilustração, o critério de parada é o tamanho da prova (20 itens).

Figura 3

Ilustração de aplicação de um teste em formato de CAT



Nota. Em cada ponto do gráfico, o acerto é representado por um círculo e o erro é representado por um "X". O comprimento do segmento tracejado corresponde ao dobro do valor do erro padrão do teta provisório.

A seguir, descrevemos os cinco componentes de uma CAT com base em Magis et al. (2017), Peres (2019), Thompson e Weiss (2011) e Weiss e Kingsbury (1984).

Banco de itens

O banco de itens de uma CAT é formado por itens calibrados, ou seja, com parâmetros psicométricos conhecidos. Em geral o modelo utilizado em uma CAT é a TRI, mas alguns testes utilizam a Teoria Clássica dos Testes (TCT) para classificar sujeitos (e.g., Rudner & Guo, 2011). A opção pela TRI favorece estimações pontuais de teta pois fornece precisão para cada ponto da escala. Outra vantagem da TRI é que ela permite posicionar os itens e os sujeitos na mesma métrica, por isso é possível administrar itens que estejam posicionados próximos ao teta do sujeito, o que otimiza a aplicação (Labarrère et al., 2011; Weiss, 1982).

Em uma CAT que utiliza a TRI, todos os itens devem ser calibrados e linkados (von Davier, 2011), para que os parâmetros dos itens e dos sujeitos sejam comparáveis. No caso do Enem, suas provas possuem itens pré-testados que fazem a ligação entre as diferentes aplicações e edições do exame. Portanto, a linkagem entre os testes de cada aplicação e de cada edição é feita a partir dos itens comuns aos pré-testes. Por isso, mesmo que as aplicações não tenham itens comuns entre si, e os sujeitos não sejam os mesmos, as notas são comparáveis porque os itens pré-testados fazem a ligação entre os testes. Os parâmetros desses itens estão posicionados na escala do Enem, e na aplicação oficial eles estão fixos (restritos ao valor prévio). Com isso, caso haja algum item novo no exame, ele é calibrado na mesma escala do Enem (Inep, 2012, 2021).

As diretrizes para o desenvolvimento do banco devem ser elaboradas a partir de simulações. Os estudos devem subsidiar decisões sobre a quantidade de itens do banco, sua distribuição ao longo da escala, distribuição de conteúdo e taxas de exposição. Os estudos de simulação devem ser desenhados de modo a contemplar diversas situações, como comparar bancos que cubram um largo espectro da escala com bancos que cubram um curto espectro, e itens de alta discriminação com itens de baixa.

Uma consideração importante para os estudos de simulação de banco é que a curva de informação do teste precisa coincidir com os objetivos do teste. Por exemplo, se o teste for

utilizado para certificação, é necessário que ele tenha mais informação na região do ponto de corte que determina a aprovação. No caso do Enem, em que os pontos de corte para ingressar nos cursos estão em um largo espectro da escala, é necessário que a informação seja alta em um grande intervalo. Por isso, o banco precisa contar com muitos itens de diferentes dificuldades.

O banco de itens deste trabalho foi composto pelos itens aplicados nas edições do Enem de 2009 a 2020 e os parâmetros dos itens foram obtidos a partir dos microdados disponibilizados pelo Inep. O Inep divulgou os parâmetros oficiais dos itens do Enem em 2022, o que facilitou os estudos sobre a qualidade da calibração desses itens a partir de então. No entanto, ainda não existem informações públicas sobre a qualidade da calibração dos itens do Enem ou sobre a adequação do desenho amostral adotado pelo Inep. Por isso, neste trabalho também avaliamos o desenho da amostra de calibração dos itens do Enem, com o intuito de refletir sobre a qualidade dos parâmetros dos itens do banco. Portanto, na seção seguinte problematizamos alguns desenhos amostrais que têm sido utilizados em calibrações de itens pela TRI.

Amostra de calibração. Não há um consenso sobre o tamanho amostral mínimo necessário para calibrar os itens adequadamente, e esse valor varia de acordo com o modelo de TRI, a quantidade de itens e a distribuição dos tetas da amostra. Porém, alguns estudos indicam que a partir de 750 sujeitos as calibrações de itens no modelo 3PL retornam valores de parâmetros aceitáveis (Nunes & Primi, 2005; Şahin & Anıl, 2017; Şahin & Weiss, 2015). Além do tamanho amostral, a distribuição dos valores de teta também deve ser considerada na calibração, porém o efeito dessa distribuição sobre a calibração tem sido pouco explorado (e.g., Paek et al., 2021; Wingersky & Lord, 1984).

A distribuição retangular (uniforme) se mostrou mais adequada do que a distribuição em forma de sino para calibrar itens do Teste de Inglês como Língua Estrangeira (TOEFL, em inglês *Test of English as a Foreign Language*) no modelo 3PL. Os erros dos parâmetros dos

itens foram menores para uma amostra retangular do que para uma amostra em forma de sino com o dobro de tamanho (Wingersky & Lord, 1984). A distribuição retangular performa melhor em calibrações com amostras menores devido à quantidade proporcionalmente maior de sujeitos de baixo teta em relação a amostras com distribuição normal (Han, 2012).

Já quando se comparou amostras aleatórias com amostras enviesadas, a primeira distribuição performou melhor em calibrações no modelo logístico de dois parâmetros da TRI (2PL). Nessa comparação, as amostras enviesadas eram compostas exclusivamente por sujeitos de baixo teta ou alto teta. Nas calibrações do modelo de Rasch, a distribuição das amostras afetou pouco a qualidade dos parâmetros e o tamanho amostral foi um fator mais importante para se garantir a invariância dos parâmetros (Paek et al., 2021).

A calibração dos itens no Enem é feita a partir de uma amostra estratificada pelo número de acertos (Inep, 2021). Existem algumas diferenças entre esse desenho e os demais utilizados nos estudos de calibração citados. Uma diferença está na distribuição dos tetas que não é aleatória e nem retangular, mas enviesada para conter uma proporção maior de sujeitos com tetas mais altos. No entanto, apesar de a amostra ser enviesada, o desenho garante que haja sujeitos posicionados em todas as regiões da escala. A outra diferença é que a estratificação dos sujeitos é feita a partir da soma de acertos e não dos tetas da TRI, o que não garante que a distribuição final de teta siga as proporções pensadas inicialmente. Por isso, neste estudo investigamos se o desenho amostral utilizado para calibrar os itens do Enem retorna parâmetros aceitáveis em comparação com outros desenhos.

Início da aplicação

O primeiro item a ser administrado pode ser selecionado a partir de um determinado teta ou ser aleatório. O teta pode ser determinado com base em uma distribuição prévia, por exemplo a média observada em uma aplicação passada. No caso do Enem, é possível obter a média das notas da edição anterior e utilizar esse valor como ponto de partida. Caso o método de seleção

do item seja o de Máxima Informação de Fisher (MIF), o programa seleciona o item que possui maior informação para esse teta. Outra possibilidade é utilizar o valor de referência da escala, que por padrão é 0.

Caso o único critério para o início da aplicação seja a máxima informação para um determinado teta, o primeiro item será o mesmo para todos os participantes e ele terá 100% de exposição. Em testes de alto impacto (*high stakes*) como o Enem, é desejável incluir métodos de controle de exposição no algoritmo da CAT. Uma possibilidade é sortear um valor de teta inicial, por exemplo de um intervalo entre -1,0 e 1,0. Outra possibilidade é selecionar os cinco itens com mais informação para o valor de teta e sortear um deles. No extremo, o primeiro item pode ser totalmente aleatório. Os métodos de controle de exposição dos itens serão abordados mais adiante.

Seleção dos itens

De forma geral, o algoritmo de seleção do item objetiva selecionar o item mais adequado para o teta provisório do sujeito e que atenda às restrições do teste (balanceamento de conteúdo e controle exposição).

Métodos de seleção de item

A seleção do item tipicamente se baseia na sua informação psicométrica, o que otimiza a aplicação. Por exemplo, isso evita que itens muito fáceis sejam apresentados a sujeitos com teta alto, ou que itens difíceis sejam apresentados a sujeitos com teta baixo. Ainda, em uma prova de certificação, evita a administração de itens com pouca informação para o ponto de corte. A Tabela 2 apresenta alguns métodos de seleção de itens comumente utilizados em CAT.

Tabela 2

Exemplos de métodos de seleção de itens em uma Testagem Adaptativa Computadorizada

Método	Descrição	Referência
Máxima Informação de Fisher	Seleciona o item mais informativo para o teta provisório do sujeito.	(Weiss, 1982)

bOpt	Seleciona o item com a dificuldade mais próxima do teta provisório do sujeito.	Urry (1970)
Máxima informação ponderada pela verossimilhança (maximum likelihood weighted information)	Pondera a máxima informação a partir da função de verossimilhança do padrão de resposta provisório.	Veerkamp e Berger (1997)
Máxima informação ponderada pela posterior (maximum posterior weighted information)	Pondera a máxima informação a partir da distribuição posterior do teta provisório.	van der Linden (1998)
Kullback-Leibler ponderada pela verossimilhança	Seleciona o item com a maior informação de Kullback-Leibler ponderada pela função de verossimilhança. É selecionado o item com a maior capacidade de discriminar qualquer par de teta.	Chang e Ying (1996)
Kullback-Leibler ponderada pela posterior	Pondera a informação de Kullback-Leibler pela distribuição posterior do teta provisório. É selecionado o item com a maior capacidade de discriminar qualquer par de teta.	Chang e Ying (1996)
Mínima variância posterior predita (minimum expected posterior variance)	Verifica a variância do teta esperada para cada item e seleciona aquele que retorna a menor.	Owen (1975)
Aleatório	Seleciona o item aleatoriamente.	

Para o conhecimento de outros métodos de seleção de item, recomendamos a leitura de Magis et al. (2017). A seguir, detalhamos o funcionamento dos métodos utilizados neste trabalho. O método MIF consiste em selecionar o item mais informativo para o teta provisório, dentre os itens disponíveis para aplicação. O item j^* selecionado para administração será aquele que:

$$j^* = arg \max_{j \in S} I_j(\hat{\theta}_t) \quad (1)$$

onde S representa o banco de itens disponíveis para aplicação, I_j é a informação do item j e $\hat{\theta}_t$ corresponde ao teta provisório, estimado após a administração de t itens. A expressão $arg\max_{j \in S}$ indica que será selecionado o item j disponível no banco S que retorne o maior resultado da

função, ou seja, o maior valor de informação para o teta provisório. O uso do método de MIF permite reduzir o tamanho de um teste e ainda manter ou aumentar a precisão média em relação à versão linear do teste (Jatobá et al., 2020; Kalender & Berberoglu, 2017; Tsaousis et al., 2021). Além disso, o MIF otimiza o teste em relação a outros métodos de seleção de itens em CAT (Sulak & Kelecioğlu, 2019).

O método de seleção aleatória também foi utilizado neste trabalho. Nesse método, um item é selecionado aleatoriamente do conjunto de itens disponíveis. O método é utilizado em estudos de simulação como referência para quantificar o quanto outros métodos aprimoram a estimação e a precisão dos tetas para além de uma simples seleção aleatória de itens (e.g., Veldkamp et al., 2010).

Restrições à seleção de itens

Além de ser capaz de apresentar o item mais adequado para o sujeito dado seu teta provisório, o algoritmo de seleção deve contemplar restrições do teste, como balanceamento de conteúdo e controle de exposição. O balanceamento de conteúdo é necessário para evitar que um teste contenha itens de somente um conteúdo e que a amostra de itens represente o domínio avaliado. De posse das especificações do teste, se estabelece a proporção em que cada conteúdo deve estar presente na aplicação. No caso do Enem, o balanceamento de conteúdo garante que as habilidades da matriz de referência sejam apresentadas em proporções semelhantes. Já o controle de exposição é importante em testes de alto impacto como o Enem, que demandam alta segurança. Sem esse controle, os itens mais discriminativos são aplicados a grande parte da população e podem se tornar conhecidos, o que pode ser uma ameaça à validade das interpretações e dos usos dos resultados do teste. Além disso, a subexposição dos itens menos discriminativos pode acarretar desperdício financeiro, pois esses itens serão subaproveitados. Por isso, os métodos de controle de exposição objetivam evitar a superexposição ou a subexposição de itens, o que aumenta a segurança do teste.

Balanceamento de conteúdo. Um método para balancear o conteúdo em um teste é a CAT restrita (*constrained CAT*, CCAT; Kingsbury & Zara, 1989). Esse método consiste em selecionar um item dentre aqueles classificados com o conteúdo mais distante de atingir a proporção desejada. Primeiro o algoritmo verifica o conteúdo mais distante da proporção estabelecida. Em seguida, o programa seleciona o subgrupo de itens que corresponde a esse conteúdo. Por último, administra o item desse subgrupo segundo o método de seleção estabelecido.

Uma limitação do método CCAT é que a ordem dos conteúdos apresentados pode ser previsível. O método CAT restrita modificado (*modified constrained CAT*, MCCAT; Leung et al., 2000) contorna a previsibilidade do ordenamento dos conteúdos do CCAT, pois em vez de selecionar somente o subgrupo mais distante, o algoritmo seleciona todos os subgrupos que ainda não alcançaram a proporção desejada.

O método do modelo multinominal modificado (*modified multinomial model*, MMM; Chen & Ankenmann, 2004) também contorna a previsibilidade do ordenamento dos conteúdos do CCAT, pois insere um componente aleatório na seleção do subgrupo de itens. Com base nas proporções desejadas de conteúdo, o algoritmo constrói uma distribuição cumulativa de soma um. Em seguida, sorteia-se um número aleatório com distribuição uniforme entre zero e um. Esse número corresponde a uma área na distribuição cumulativa. É do conteúdo localizado nessa área que o item será selecionado.

Em uma simulação com aplicações fixas de 16, 28 e 40 itens, com o método de MIF e controle de exposição de Sympson-Hetter (que será explicado mais adiante), os três métodos de balanceamento de conteúdo apresentaram resultados semelhantes em termos de precisão. Nos testes de 16 e 28 itens, a superexposição dos itens com o CCAT foi ligeiramente maior e a dos demais, semelhante. Em testes de 40 itens, a superexposição de itens com o MMM foi ligeiramente menor do que a dos demais (Leung et al., 2003). Apesar dessa ligeira superioridade

do método MMM em testes de 40 itens, neste trabalho utilizamos o MCCAT, pois o MMM não garante que a proporção de conteúdo seja seguida, pois é um método probabilístico. Além disso, o controle de exposição que adotamos (progressivo restrito) impede a superação da taxa de exposição desejada, diferentemente do Sympson-Hetter.

Controle de exposição. É possível agrupar os métodos de controle de exposição em quatro categorias: condicionais, randomizados, estratificados e combinados (Georgiadou et al., 2007; Leroux et al., 2013). Os métodos condicionais em geral incluem um parâmetro de controle de exposição r, que indica a probabilidade de administrar o item dado que ele foi selecionado. Os métodos randomizados incluem uma randomização e em geral selecionam aleatoriamente um item de um subgrupo próximo do nível de maior informação. Os métodos estratificados agrupam os itens de acordo com suas propriedades psicométricas e selecionam o item a partir desses estratos. Os métodos combinados agregam estratégias diferentes de controle de exposição. A Tabela 3 apresenta alguns métodos de controle de exposição de itens em uma CAT.

Tabela 3

Exemplos de métodos de controle de exposição de itens em uma Testagem Adaptativa Computadorizada

Método	Descrição	Referência
	Métodos condicionais	
Sympson-Hetter	Cada item recebe um parâmetro de controle de exposição entre zero e um. Quando ele é selecionado, sorteia-se um número aleatório de uma distribuição uniforme entre zero e um. Se esse número for menor ou igual ao parâmetro estabelecido, o item é administrado. Caso contrário, um novo item é selecionado.	Hetter e Sympson (1997)
Máxima informação restrita	No início da aplicação, somente os itens com taxa de exposição menor do que a desejada ficam disponíveis para seleção. Os itens	Revuelta e Ponsoda (1998)

Método	Descrição	Referência
	remanescentes são selecionados pelo método de MIF.	
	Métodos randomizados	
Randomizado (randomesque)	Sorteia um item dentre os n itens mais informativos para o teta provisório. A quantidade n de itens que compõem esse grupo é fixa em todas as rodadas da CAT.	Kingsbury e Zara (1989)
5-4-3-2-1	O primeiro item é sorteado dentre os cinco mais informativos para o teta inicial. O segundo é sorteado dentre os quatro mais informativos para o teta provisório. O terceiro, dentre os três e o quarto, dentre os dois. A partir do quinto item, é administrado o mais informativo.	McBride e Martin (2014)
Progressivo	Seleciona o item que apresenta a maior soma entre dois componentes: um aleatório e um informativo. No início da aplicação, o peso do componente aleatório é de 100% e o do informativo, 0%. O peso do componente aleatório reduz ao longo da aplicação e o do informativo, aumenta.	Revuelta e Ponsoda (1998)
	Métodos estratificados	
a-estratificado	Os itens são agrupados de acordo com o parâmetro de discriminação. No início da aplicação, são selecionados itens com discriminação menor. No final, os itens com discriminação maior.	Chang e Ying (1999)
a-estratificado com b-blocos	Os itens são ordenados de acordo com sua dificuldade e em seguida agrupados de acordo com o parâmetro de discriminação. O algoritmo seleciona o grupo de itens com dificuldade próxima ao teta provisório. Em seguida, seleciona o item de maneira semelhante ao método a-estratificado.	Chang et al. (2001)
	Métodos combinados	
Progressivo restrito	No início da aplicação, aplica a regra do método de máxima informação restrita. Em seguida, aplica o método progressivo.	Revuelta e Ponsoda (1998)

Método	Descrição	Referência
Combinação de a-estratificado com Sympson-Hetter	Inclui o método Sympson-Hetter no método a-estratificado.	Leung et al. (2002)

Neste trabalho, utilizamos o método progressivo restrito (PR), que é um método que combina o de máxima informação restrita e o progressivo. Por isso, detalhamos esses três métodos a seguir.

O método de máxima informação restrita (Revuelta & Ponsoda, 1998) foi proposto como um método computacionalmente mais simples que o método Sympson-Hetter (Hetter & Sympson, 1997), que demanda simulações para determinação do parâmetro r de cada item. Na máxima informação restrita, nenhum item pode ser exposto a mais do que uma determinada proporção das aplicações. No início da aplicação, somente os itens com taxa de exposição menores do que r^{max} estarão disponíveis. Os itens remanescentes são então selecionados pelo método de MIF. Em termos de notação, o método pode ser descrito da seguinte maneira:

$$r_j = \begin{cases} 1 & se\ A_j < r^{max} \\ 0 & se\ A_j \geq r^{max} \end{cases} \quad (2)$$

onde r_j é a probabilidade do item j ser administrado dado que foi selecionado pelo método MIF, A_j é a taxa de exposição do item j e r^{max} é a taxa de exposição máxima estabelecida para os itens. Além de impor uma exposição máxima para todos os itens, esse método também reduz a sobreposição de itens sem comprometer a precisão da medida (Barrada et al., 2009; Huebner, 2012; Revuelta & Ponsoda, 1998). A sobreposição corresponde à proporção de itens iguais aplicados a dois participantes selecionados aleatoriamente. Apesar de reduzir a superexposição e a sobreposição de itens, o método da máxima informação restrita não reduz a subexposição de itens (Revuelta et al., 1998).

Outra desvantagem desse método é a previsibilidade da seleção dos itens (Barrada et al., 2009), pois o primeiro item será o mesmo a cada n respondentes, onde n depende do valor de r. Por exemplo, suponha que a taxa máxima de exposição esteja fixada em $r^{max} = 0{,}30$. O item inicial do primeiro respondente terá taxa de exposição $A_j = 1{,}00$ quando o segundo

respondente iniciar o teste. Caso não houvesse restrição de exposição, o primeiro item do teste seria igual para todos os sujeitos, pois o método de MIF seleciona sempre o item mais informativo para o teta provisório, que no início da aplicação seria o mesmo para todos. No entanto, esse item não será aplicado para o segundo sujeito, pois a taxa de exposição supera r^{max}. Para o terceiro respondente, a taxa de exposição desse item será $A_j = 0{,}50$, para o quarto, $A_j = 0{,}33$ e para o quinto, $A_j = 0{,}25$, quando será aplicado novamente como primeiro item. Nessa situação, esse item será aplicado a cada quatro participantes, o que leva sua exposição a $A_j = 1{,}00$ para esse grupo de sujeitos. Uma maneira de contornar essas limitações desse método é incluir um componente aleatório na seleção do item.

O uso do método progressivo (Revuelta & Ponsoda, 1998) inclui um componente aleatório no processo de escolha do item. Por isso, ele reduz a superexposição, a subexposição e a sobreposição de itens, com pouco impacto na precisão em relação à ausência de controle (Barrada et al., 2008; Huang et al., 2012; Yasuda et al., 2022). O item selecionado será aquele que apresentar a maior soma entre dois componentes, sendo um aleatório e outro informativo. Esses componentes recebem um peso que vai se alterando ao longo da aplicação. No início da aplicação, o peso dado ao componente informativo é zero, e para o componente aleatório esse peso vale um. Conforme a aplicação evolui, o peso do componente informativo aumenta e o do componente aleatório diminui. Em termos de notação, no método progressivo o item selecionado j^* será aquele que:

$$j^* = arg\max_{j \in S} \left| (1 - W)R_j + WI_j\left(\hat{\theta}_t\right) \right| \quad (3)$$

onde R_j é um número aleatório com distribuição uniforme sorteado de um intervalo entre zero e o valor da maior informação dentre os itens disponíveis no banco, $\left[0; max_{j \in S} I_j\left(\hat{\theta}_t\right)\right]$, e W é o peso dado aos componentes aleatório (R_j) e informativo $\left[I_j\left(\hat{\theta}_t\right)\right]$ da equação. A expressão $arg\max_{j \in S}$ indica que será selecionado o item j disponível no banco S que retorne o maior

resultado da função, ou seja, que apresente a maior soma entre os componentes aleatório e informativo. O peso W foi concebido da seguinte maneira (Revuelta & Ponsoda, 1998):

$$W = \frac{t}{M} \tag{4}$$

onde M é o tamanho máximo do teste. Posteriormente, Barrada et al. (2008) propuseram outra forma de calcular esse peso:

$$W = \begin{cases} 0, & \text{se } q = 1 \\ \frac{\sum_{b=2}^{t}(b-1)^k}{\sum_{b=2}^{N}(b-1)^k}, & \text{se } q \neq 1 \end{cases} \tag{5}$$

onde N é o tamanho do teste, k é um parâmetro que permite controlar a velocidade com que W se distancia de 0 ao longo da aplicação e q é a posição do item na aplicação. Valores negativos de k aceleram o aumento de W e o componente aleatório perde importância mais rapidamente. Quanto maior o parâmetro de aceleração k, mais lentamente W aumenta, e mais lentamente o componente aleatório perde importância. Por isso, valores mais altos de k aumentam a taxa de uso dos itens menos discriminativos (Barrada et al., 2008). Quanto maior a importância da informação na seleção dos itens, mais os itens com maior discriminação são selecionados. As Equações 4 e 5 implicam completa aleatoriedade no início do teste, pois quando $W = 0$, o que resta na Equação 3 é o componente aleatório. Conforme W aumenta, a importância do componente informativo aumenta e a do componente aleatório diminui.

Algumas aplicações possuem tamanho variado, por exemplo quando o critério de parada se relaciona com a precisão da medida. Nesses casos, W pode ser calculado da seguinte maneira (McClarty et al., 2006)

$$W = \frac{EP_{parada}}{EP_t} \tag{6}$$

onde EP_{parada} é o valor de erro padrão utilizado como critério de parada e EP_t é o erro padrão para o teta provisório. Magis e Barrada (2017) modificaram essa forma de calcular W e

incorporaram o parâmetro de aceleração no pacote catR (Magis & Raîche, 2012) da seguinte maneira:

$$W = \begin{cases} 0, & \text{se } q = 1 \\ max\left[\frac{I(\hat{\theta}_t)}{I_{parada}}, \frac{q}{M-1}\right]^k, & \text{se } q \neq 1 \end{cases} \quad (7)$$

onde I_{parada} é a informação necessária para atingir o valor de parada do erro padrão. O pacote catR implementado no R utiliza as Equações 3, 5 e 7. Neste estudo também utilizamos essas equações, pois avaliamos o efeito do aumento do valor do parâmetro de aceleração sobre a eficiência, precisão e segurança do teste. Para calcular a informação para um determinado teta, esse pacote utiliza a seguinte equação:

$$I = \frac{1}{EP(\theta)^2} \quad (8)$$

Essa equação se refere ao cálculo da informação para métodos não bayesianos, como o da máxima verossimilhança (*maximum likelihood*, ML). Como neste trabalho utilizamos um método bayesiano para estimar o teta, adotamos a seguinte equação para calcular a informação (Nicewander & Thomasson, 1999):

$$I = \frac{1}{EP(\theta)^2} - 1 \quad (9)$$

O método PR (Revuelta & Ponsoda, 1998) adota a taxa máxima de exposição (método da máxima informação restrita) e inclui um componente aleatório na seleção do item (método progressivo). Dessa forma, somente os itens com taxa de exposição abaixo da estabelecida estarão disponíveis para administração, e eles serão selecionados pelo método progressivo. Por combinar esses dois métodos, ele evita a superexposição e reduz a sobreposição e a subexposição. Além disso, esse método pouco piora a precisão da medida quando comparada a situações sem controle de exposição (Barrada et al., 2008; Lee & Dodd, 2012; Leroux & Dodd, 2016; Leroux et al., 2013, 2019).

Estimação do teta

A TRI é comumente utilizada para calcular o teta do participante em uma CAT (e.g., Kalender & Berberoglu, 2017; Mizumoto et al., 2019; Tsaousis et al., 2021). Uma das vantagens da TRI em relação à TCT é que ela permite posicionar os itens e os sujeitos em uma mesma escala. Desse modo, é possível avaliar o quão informativo é um item para um sujeito em cada etapa da aplicação. Isso contribui para otimizar a aplicação, pois o algoritmo tende a selecionar os itens que contribuirão mais para a estimação do teta do sujeito. Uma possibilidade de calcular o teta pela TRI é por meio do método de ML. A limitação desse método é a impossibilidade de estimar um teta quando o sujeito possui 100% de acerto ou de erro. Já métodos Bayesianos, como o *expected a posteriori* (EAP) e o *maximum a posteriori* (MAP), superam essa limitação. No entanto, eles são enviesados pois as estimativas de teta têm influência da distribuição prévia utilizada e são menos enviesadas quando essa distribuição possui valores próximos da distribuição da amostra (Cúri & Silva, 2019). Apesar disso, esses métodos se mostram mais precisos em aplicações de CAT do que o ML (Barrada et al., 2009; Seo & Choi, 2018). Neste trabalho utilizamos o método EAP.

Critério de parada

As CATs podem ser de tamanho fixo ou variável. Quando de tamanho fixo, o critério de parada é a quantidade de itens aplicados ao sujeito. Quando são utilizados outros critérios de parada, a CAT possui tamanho variável, e a quantidade de itens necessários é adaptável. Alguns critérios avaliam a estimativa do teta, outros, o erro de medida e outros, o banco de itens. A Tabela 4 apresenta alguns critérios de parada, sua descrição e exemplos de aplicação.

Tabela 4

Exemplos de critérios de parada em uma Testagem Adaptativa Computadorizada

Critério	Descrição	Exemplo
Tamanho fixo	O teste encerra quando uma determinada quantidade de itens é aplicada.	Cúri e Silva (2019)

Erro padrão	O teste encerra quando o erro padrão do teta provisório é menor ou igual a um determinado valor.	Leroux et al. (2019)
Diferença absoluta do teta	O teste encerra quando a diferença absoluta entre o teta provisório e o teta anterior é menor de que um determinado valor.	Babcock e Weiss (2012)
Mínima informação do banco	O teste encerra quando nenhum item disponível possui informação psicométrica para o teta provisório maior do que determinado valor.	Stafford et al. (2019)
Redução do erro	O teste encerra quando a redução do erro após a administração de um item é menor do que um determinado valor.	Kallen et al. (2018)
Redução predita do erro padrão (Predicted standard error reduction)	O teste encerra quando o item selecionado pelo método de Mínima variância posterior predita reduz o erro padrão em um valor menor do que o determinado.	Morris et al. (2020)

Quando o critério de parada é o erro padrão, a testagem encerra se o erro padrão de medida atinge um valor considerado satisfatório. Uma possibilidade para se estabelecer o ponto de corte desse critério é relacioná-lo com a confiabilidade na abordagem clássica dos testes, pois quando o desvio padrão da métrica é 1, temos (Nicewander & Thomasson, 1999):

$$EP(\theta) = \sqrt{1 - \rho(\hat{\theta})} \tag{10}$$

onde $\rho(\hat{\theta})$ é a confiabilidade para um dado teta, e $EP(\theta)$ é o erro padrão de medida. Por exemplo, alguns estudos utilizam um erro padrão de 0,30 (Aytuğ Koşan et al., 2019; Choi et al., 2011; Kalender & Berberoglu, 2017), que equivale a uma confiabilidade de 0,91, pois:

$$0{,}30 = \sqrt{1 - \rho(\hat{\theta})} \therefore \rho(\hat{\theta}) = 0{,}91 \tag{11}$$

Quando o banco de itens possui uma distribuição relativamente uniforme em termos de informação psicométrica, esse critério pode promover uma precisão similar ao longo da escala. No entanto, caso a curva de informação do banco seja do tipo pico, para pessoas localizadas em regiões de pouca informação as aplicações podem contar com a administração de itens que contribuirão pouco para a redução do erro. Como consequência, elas se tornam

desnecessariamente extensas e em alguns casos a precisão desejada não é alcançada (Wang et al., 2019).

Uma possibilidade de otimizar uma CAT de banco não uniforme é usar o critério da redução predita do erro padrão (*Predicted Standard Error Reduction*, PSER; Choi et al., 2011), que considera a informação psicométrica do banco disponível. Esse critério encerra a aplicação quando não há itens disponíveis que melhorem substancialmente a precisão do teta. Quando o critério de parada é o erro padrão, os sujeitos cujo teta não se localiza na região informativa da escala terão aplicações mais longas desnecessariamente, pois ainda que novos itens sejam apresentados a mudança no erro será pouca. Por mais que a precisão do teta provisório seja insatisfatória, a administração de novos itens não contribuirá para reduzir o erro de medida. Por isso, o critério PSER encerra a aplicação em vez de apresentar novos itens desnecessariamente (Choi et al., 2011; Morris et al., 2020).

O parâmetro do algoritmo responsável por esse término é o *hypo*, que indica o valor da redução predita no erro que determina o fim da aplicação a qualquer momento. Caso a redução seja menor do que o valor de *hypo*, a aplicação encerra. O PSER possui um segundo parâmetro (*hyper*), responsável por fazer a aplicação continuar mesmo após o erro padrão desejado ser alcançado. Caso o erro de medida seja menor do que o desejado, a aplicação continua enquanto a redução predita for maior do que *hyper*, o que melhora a precisão para os casos em que o teta está posicionado em uma região com alta informação psicométrica. Assim, a aplicação encerra em uma das duas condições:

$$\Delta EP_{pred}(\theta) < hypo$$
$$EP(\theta) < EP_{parada} \wedge \Delta EP_{pred}(\theta) < hyper$$

onde $\Delta EP_{pred}(\theta)$ é a diferença predita no erro padrão e EP_{parada} é o valor de erro padrão usado como critério de parada. Por exemplo, se o valor de parada do erro for 0,30, o parâmetro *hypo* for 0,01 e o parâmetro *hyper* for 0,05, a aplicação encerra a qualquer momento se $\Delta EP_{pred}(\theta) < 0{,}01$. Adicionalmente, caso o erro padrão atinja 0,30, a aplicação encerra se $\Delta EP_{pred}(\theta) <$

0,05. O parâmetro *hyper* permite que a aplicação aumente a precisão em relação ao valor desejado, caso isso ainda seja possível. Já o parâmetro *hypo* evita que a aplicação continue desnecessariamente.

O uso do critério PSER aumenta a eficiência da CAT sem comprometer a precisão (Choi et al., 2011; Morris et al., 2020), porém neste trabalho não o utilizamos. Optamos pelo critério da redução do erro (Luijten et al., 2021), que é mais recente e menos utilizado. Esse critério funciona de maneira semelhante ao parâmetro *hypo* do PSER, ou seja, a aplicação encerra quando a redução do erro é menor do que a estabelecida. A diferença deste critério para o PSER é que ele utiliza o valor do erro observado em vez do erro predito, ou seja, verifica a redução do erro após a administração de um item. Isso torna a aplicação menos custosa em termos computacionais, pois não é necessário verificar o erro predito para cada item do banco. Ainda, sua implementação é menos complexa.

Algumas CATs contam com combinações de critério, o que pode aumentar a eficiência da aplicação. Um exemplo é combinar o critério do erro padrão com a redução do erro, como feito por Kallen et al. (2018). Essa combinação se assemelha ao caso específico do critério PSER que tem o parâmetro *hyper* de valor infinito. Nesse caso, além da CAT encerrar quando a redução do erro for menor do que *hypo*, ela encerra quando o erro padrão atinge o valor desejado, pois qualquer valor de redução após a próxima administração será menor que o *hyper* (infinito).

Independente dos critérios de parada descritos, a CAT também pode incluir um mínimo e um máximo de itens aplicados. No que diz respeito a estabelecer um número máximo de itens adequadamente, isso evita que a aplicação se torne desnecessariamente longa com a administração de itens que pouco contribuem para a precisão (Stafford et al., 2019; Wang et al., 2022). Já em aplicações muito curtas, a diferença entre o teta verdadeiro e o teta simulado aumenta. Isso pode ser evitado ao se adotar um número mínimo de itens razoável, como nove

(Stafford et al., 2019) ou dez (Babcock & Weiss, 2012). Neste trabalho, avaliamos CATs de tamanho fixo (20 e 45 itens) e de tamanho variável (erro padrão de 0,30; e combinação do erro padrão de 0,30 com a redução do erro de 0,015). Nas CATs de tamanho variável, adotamos o tamanho mínimo de 15 itens e o máximo de 60 itens.

Desenvolvimento de uma CAT

Para desenvolver uma CAT e implementar os cinco componentes citados acima, Thompson e Weiss (2011) propuseram uma diretriz com cinco etapas, que vão desde se questionar se a CAT é adequada para a avaliação em questão, até a sua publicação. A Tabela 5 apresenta as cinco etapas e as principais ações necessárias para concluir cada etapa.

Tabela 5

Matriz com as etapas para desenvolver uma CAT

	Etapa	Ações principais
1	Estudos de viabilidade, aplicabilidade e planejamento	Simulações de Monte Carlo; avaliação do estudo de caso
2	Desenvolvimento do conteúdo do banco de itens ou utilização de um banco existente	Elaboração e revisão de itens
3	Pré-testagem e calibração do banco de itens	Pré-teste e análise de itens
4	Determinação da especificação final da CAT	Simulações híbridas ou *post-hoc*
5	Publicação da CAT	Publicação e distribuição; desenvolvimento de programa

Fonte: traduzida de Thompson e Weiss (2011)

As diretrizes propostas indicam que o primeiro passo para implementar uma CAT é verificar se ela é viável e faz sentido para um determinado programa de testagem ou avaliação. Nessa etapa, deve-se avaliar a real necessidade e viabilidade de se transformar um teste linear em adaptativo. Alguns estudos podem subsidiar a tomada de decisão, como simulações de Monte Carlo para investigar questões relacionadas ao tamanho do teste, à precisão da medida, à exposição dos itens e ao tamanho do banco de itens.

De posse das informações iniciais, o segundo passo para desenvolver uma CAT é constituir um banco de itens. As simulações feitas na etapa anterior devem ser aproveitadas

para prover diretrizes para o banco. As considerações necessárias para constituir o banco de itens foram apresentadas na seção Banco de itens desta tese. Após a elaboração dos itens, eles precisam ser pré-testados para que seus parâmetros da TRI sejam conhecidos. Para a aplicação de uma CAT, é necessário que todos os itens estejam calibrados e linkados (von Davier, 2011). O passo seguinte para desenvolver uma CAT é especificar o seu algoritmo, ou seja, montar a arquitetura representada na Figura 2 com os componentes descritos nas seções anteriores. Com a CAT arquitetada, é preciso que ela seja aplicada por meio de um programa em computador.

Neste trabalho, desenvolvemos uma CAT para o Enem. Mais especificamente, nossos estudos se inserem nos três últimos passos do desenvolvimento de uma CAT. O primeiro artigo desta tese avalia a calibração dos itens do Enem, o que contribui para avaliar a qualidade do banco de itens da CAT. O segundo artigo desta tese realiza uma simulação robusta para propor um algoritmo para a CAT do Enem. Outro produto desta tese é a publicação da aplicação da CAT arquitetada de acordo com o algoritmo proposto.

Adicionalmente, produzimos um pacote aberto para realizar simulações em CAT com especificações não incluídas em outros programas estatísticos (Jaloto & Primi, 2023a). O pacote deste trabalho se baseia no pacote catR (Magis & Raîche, 2012) e avança em alguns pontos. Uma das singularidades do pacote desenvolvido é a possibilidade de utilizar os métodos MCCAT e MMM para balancear o conteúdo do teste. Outra característica do pacote é a possibilidade de calcular o parâmetro W_t do método progressivo pelas Equações 4, 5, 6 e 7. Outra inclusão do pacote é o critério de parada da redução do erro. Alteramos também a forma de calcular a informação de um item para a Equação 9, que considera a estimação do teta por um método Bayesiano. No que diz respeito à estimação do teta, utilizamos a função *fscores* do pacote mirt (Chalmers, 2012), que calcula mais rapidamente do que o algoritmo da função *eapEst* do pacote catR. Por último, excluímos as linhas de comando do pacote catR que anulavam a semente aleatória após selecionar um item, o que garante a reprodutibilidade do

estudo quando o método de seleção contém um componente aleatório (por exemplo, com o controle de exposição progressivo ou com o balanceamento de CCAT).

Simulação em CAT

Estudos de simulação devem subsidiar as decisões sobre os componentes da CAT. As simulações podem ser do tipo Monte Carlo, *post-hoc* ou híbrido. As simulações de Monte Carlo usam somente dados gerados aleatoriamente, que seguem determinadas distribuições. No contexto de desenvolvimento de uma CAT, elas são utilizadas quando não há respostas reais aos itens. As simulações *post-hoc* ocorrem quando são utilizados somente dados reais. Já as simulações híbridas contam com dados reais na medida do possível e dados simulados como complemento.

As simulações de Monte Carlo se baseiam no fato de que a TRI fornece a probabilidade de um sujeito acertar determinado item em um teste, dados os parâmetros do sujeito e do item. Essa probabilidade no modelo 3PL é calculada por:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \quad (12)$$

onde θ é o teta do sujeito, e é a constante neperiana e a_i, b_i e c_i são os parâmetros de discriminação, dificuldade e pseudochute do item, respectivamente. Uma simulação de Monte Carlo em uma CAT consiste nas seguintes etapas (Nydick & Weiss, 2009):

1. Gerar os parâmetros dos sujeitos. Esses parâmetros são gerados aleatoriamente e seguem distribuições determinadas. Por exemplo, pode-se adotar uma distribuição normal com média 0 e desvio padrão 1. Sabendo-se dos parâmetros de distribuições de populações em outras aplicações, eles também podem ser utilizados. Por exemplo, no caso do Enem é possível verificar a distribuição dos sujeitos em cada edição do exame e utilizar suas informações estatísticas. Neste trabalho, utilizamos as notas dos participantes do Enem 2020.

2. Gerar os parâmetros dos itens. Os parâmetros são gerados aleatoriamente seguindo distribuições determinadas. Por exemplo, se o parâmetro de discriminação seguir uma distribuição log-normal com média 0,5 e desvio padrão 1, seus valores serão sempre positivos. Caso haja informações prévias sobre distribuições de parâmetros em aplicações anteriores, elas podem ser utilizadas neste momento. No caso do Enem, é possível utilizar os parâmetros dos itens já aplicados e dos pré-testados ainda não publicados. Ainda, é possível complementar o banco de itens com a geração de novos parâmetros. Neste trabalho, utilizamos os parâmetros dos itens aplicados nas edições de 2009 a 2020.
3. Gerar a matriz de respostas com base em um modelo de TRI. Para cada combinação de sujeito e item, é feita uma simulação para se obter uma resposta dicotômica (0, para erro ou 1, para acerto). Para simular a resposta de um sujeito a um item, primeiro calcula-se a probabilidade de acerto ao item. Em seguida, gera-se um número aleatório com distribuição uniforme entre zero e um. Se esse número gerado for menor ou igual à probabilidade de acertar o item, atribui-se acerto (1) ao sujeito. Nesse caso, é como se o sujeito tivesse superado o item. Caso o número aleatório seja maior do que a probabilidade de acerto, é como se o item tivesse superado o sujeito, por isso atribui-se o erro (0). Essa operação é repetida para todos os sujeitos e itens. Por exemplo, para um item com discriminação 1,5, dificuldade 1,0 e pseudochute 0,20, um sujeito com teta 1,0 possui probabilidade de 0,60 de acerto. Caso o número sorteado seja menor ou igual a 0,60, atribui-se acerto ao sujeito. Se o número sorteado for maior do que 0,60, atribui-se erro.
4. Implementar a CAT a partir da matriz de respostas. Uma vez definidas as especificações dos componentes da CAT, as respostas geradas serão utilizadas para

simular uma situação de aplicação. A resposta a cada item será utilizada conforme ele for selecionado pelo algoritmo e ao final estima-se o teta do sujeito.

Simulações *post-hoc* utilizam dados empíricos. Diferentemente da simulação do tipo Monte Carlo, nessa simulação a matriz de respostas é composta somente por respostas reais aos itens. No caso de uma simulação de CAT das aplicações do Enem disponíveis no portal do Inep, isso só é possível se o banco de itens for composto por uma única aplicação (45 itens). Um exemplo é a simulação *post-hoc* de Spenassato et al. (2016) com itens de MT do Enem 2012. As notas obtidas na CAT de tamanho fixo de 33 itens tiveram correlação de 0,998 com a nota original calculada com 45 itens.

Nas simulações híbridas, são utilizados dados empíricos sempre que possível e os dados faltantes são substituídos com simulações de Monte Carlo. Ou seja, nesse caso a matriz de respostas é composta por respostas reais aos itens e dados simulados. Uma possível aplicação dessa simulação é em situações de um pré-teste, em que os participantes respondem cadernos diferentes por conta do tamanho do teste, ou se inclui o procedimento de Blocos Incompletos Balanceados (Bekman, 2001).

No caso do Enem, é possível realizar simulações híbridas utilizando as informações das provas aplicadas, porém optamos por simular as respostas em sua totalidade. Isso porque este é um trabalho que inicia uma discussão sobre aplicação do Enem em formato adaptativo, no que diz respeito ao pioneirismo em termos de magnitude do banco de itens e complexidade de análise. Como não conhecemos a magnitude do efeito de interferências externas nas respostas no Enem (como fadiga e motivação), em um ambiente simulado podemos ter uma linha de base em que os dados seguem os postulados do modelo de TRI adotado no exame. Nesse caso, garantimos que a probabilidade de acerto dos itens é dada exclusivamente pelos parâmetros (dos itens e do sujeito) relacionados a somente uma dimensão. Ou seja, os padrões de respostas

são explicados pela variável latente do modelo e não são influenciados por dimensões adicionais como a fadiga e a motivação.

As simulações em CAT são importantes para determinar suas especificações, no que se refere tanto ao algoritmo da CAT (início da aplicação, seleção dos itens, estimação do teta e critério de parada), quanto ao banco de itens que de fato será utilizado para cada sujeito. Existem diversos programas e pacotes comerciais e gratuitos que se prestam a essa tarefa, como catIrt (Nydick, 2014), catR (Magis & Raîche, 2012), CATSim (Weiss & Guyer, 2012), Firestar (Choi, 2020) e mirtCAT (Chalmers, 2016). Apesar de terem similaridades, cada pacote possui diferentes possibilidades de especificação de simulação, bem como diferentes limitações. Para atender às especificações do desenho deste trabalho, construímos o pacote simCAT com base no pacote catR.

Pesquisas com CAT

O método MIF tem se mostrado uma opção para aumentar a eficiência da aplicação em termos de quantidade de itens administrados, sem impactar fortemente na precisão. A simulação de Sulak e Kelecioğlu (2019), do tipo Monte Carlo, contou com um banco de 250 itens. Quando a aplicação tinha tamanho fixo de 40 itens, o erro padrão médio foi 0,18. Com critério de parada de erro padrão, as quantidades médias de itens aplicados foram 30,07 (erro menor que 0,20) e 7,07 (erro menor que 0,40).

Simulações do tipo *post-hoc* também demonstram a eficiência do método MIF. A simulação de Bulut e Kan (2012) contou com dois bancos de 40 e um de 80 itens e o critério de parada foi o erro padrão (0,25, 0,30 e 0,40). A média de itens aplicados na CAT variou de nove a 22 e a correlação entre o teta real e o teta simulado, de 0,93 a 0,98. Mizumoto et al. (2019) utilizaram uma CAT de tamanho fixo para reduzir três testes de língua inglesa de 115, 73 e 56 para 20, 15 e 10 itens, respectivamente. Os testes foram aplicados a 760 estudantes

universitários japoneses com o método de seleção de MIF. O erro padrão ficou abaixo de 0,33 para proporções que variaram de 69,6% a 74,3% dos participantes do estudo.

Com o objetivo de verificar a segurança de um teste, outros estudos compararam diferentes métodos de controle de exposição. A simulação de Lee e Dodd (2012) contou com três bancos de 120 itens (um com itens fáceis, outro com médios e outro com difíceis). O método de seleção de itens foi o de MIF e o tamanho da aplicação foi fixado em 20 itens. As autoras compararam dois métodos de controle de exposição de itens (randomizado de seis itens e PR com taxa de 0,30) e verificaram pouca ou nenhuma piora na raiz do erro quadrático médio (REQM) em relação à ausência de controle. Nas condições com o método PR todos os itens do banco foram utilizados. Diferentemente, a proporção de itens não utilizados com o método randomizado variou de 11,7% a 30,2%. Essa proporção chegou a 32,8% sem controle de exposição. Ainda, a taxa máxima de exposição dos itens com o método PR foi de 0,30, ao passo que com o método randomizado esse valor chegou a 0,78 e sem controle de exposição, a 1,00. A maior média de sobreposição de itens com o método PR foi 5,22, ao passo que com o método randomizado e sem controle de exposição esses valores chegaram a 9,95 e 10,13 respectivamente.

A simulação de Leroux e Dodd (2016) incluiu o balanceamento de conteúdo pelo CCAT e comparou três métodos de controle de exposição (randomizado com seis itens, Sympson-Hetter com taxa de 0,30 e PR com taxa de 0,30), além de dois critérios de parada (tamanho fixo de 20 itens; e erro padrão de 0,30 com máximo de 20 itens). A simulação foi do tipo Monte Carlo e os bancos tinham 157 e 245 itens. Os valores médios da REQM foram ligeiramente piores para o método PR, apesar de baixos. Esses valores chegaram a no máximo 0,34, enquanto nas condições sem controle de exposição eles chegaram a 0,31. No entanto, esse método proporcionou maior segurança para o teste, pois suas taxas médias de sobreposição foram as

mais baixas. Ainda, somente nas aplicações com esse método todos os itens do banco foram utilizados.

A simulação do tipo Monte Carlo de Barrada et al. (2008) comparou diferentes valores para o parâmetro de aceleração k do método PR (-1, 0, 1 e 2) por meio de simulações com bancos de 500 itens e aplicações de tamanhos fixos de 25 e 40 itens. No geral, os valores de REQM foram ligeiramente menores quando $k = 0$ e ligeiramente maiores quando $k = 2$. No entanto, quando $k = 2$ a quantidade de itens utilizados em no máximo 2% das aplicações reduziu, o que indica que os itens menos apresentados nas outras condições foram mais administrados. Neste trabalho, comparamos diferentes valores do parâmetro k em CATs de tamanhos variáveis e fixos.

Mais recentemente Leroux et al. (2019) simularam CATs com bancos de 43, 86 e 172 itens. A precisão foi similar entre os métodos de controle de exposição (PR com taxa de 0,30 e randomizado de três e seis itens), porém a segurança do teste foi maior com o PR. Nas aplicações com esse método, todos os itens do banco foram administrados e a sobreposição de itens foi menor. Em termos de eficiência da CAT (tamanho da aplicação), as aplicações com o critério de parada de PSER foram ligeiramente melhores do que as com o de erro padrão de 0,32. No entanto, a precisão das CATs com critério de parada PSER foi ligeiramente pior, com REQM em geral 0,01 maior (no máximo 0,32).

O aumento na eficiência da CAT com o critério de PSER é esperado por conta do parâmetro *hypo*, que evita a administração desnecessária de itens. A simulação de Morris et al. (2020) verificou que utilizar hypo = 0,015 e hyper = 0,025 pode produzir resultados tão precisos quanto uma CAT com critério de parada de erro padrão de 0,30 e uma aplicação mais eficiente em termos de tamanho. O estudo mostrou que esse valor de *hypo* aumenta a eficiência da CAT sem reduzir a precisão quando o banco não possui uma distribuição uniforme, ou seja, quando há um pico de informação. Isso reforça o potencial de utilizar a redução do erro como critério

de parada para aumentar a eficiência de uma CAT. Neste trabalho avaliamos o critério da redução do erro, que funciona de maneira semelhante ao parâmetro *hypo* do PSER, porém é mais simples computacionalmente. Trata-se de um avanço, pois são raros os estudos que investigam sua eficiência e seu impacto na precisão.

Kallen et al. (2018) verificaram a possibilidade de reduzir uma CAT da escala de Saúde Global do *Patient-Reported Outcomes Measurement Information System* (PROMIS), que possui como critério de parada o erro padrão (3,0), com máximo de 12 itens. A escala possui média 50 e desvio padrão 10. A alteração na CAT foi uma combinação no critério de parada: a resposta aos dois primeiros itens associada a boa saúde, o erro padrão de 3,0, a redução do erro em 0,1 ou o máximo de 12 itens. A combinação alterada reduziu o tamanho médio da aplicação para todos os oito testes investigados e essa diferença chegou a 3,3 itens. Apesar de não ser possível associar toda a redução à inclusão do critério de redução do erro, as proporções de aplicações encerradas por esse critério variaram de 1,9% a 13,3%. Ainda, a REQM entre as CATs originais e as alteradas foi 2,24 considerando todos os domínios avaliados, o que indica que a redução da aplicação não teve grande impacto nos tetas. Se a escala fosse transformada para uma de média 0 e desvio padrão 1, essa REQM seria de 0,022. A composição dos bancos variou de 16 a 121 itens. Neste trabalho, verificamos a relação do critério de parada da redução do erro com a eficiência e a precisão do teste, utilizando bancos com mais de 700 itens.

O presente estudo

Nas seções anteriores, defendemos a possibilidade do uso de uma CAT para aprimorar o Enem, dado o extenso tamanho de suas provas e a necessidade de se produzir uma nota precisa em um largo intervalo da escala. Uma das etapas do desenvolvimento de uma CAT é a construção de um banco de itens calibrados. Sabe-se que a calibração dos itens é afetada pelo desenho amostral, que inclui o tamanho e a distribuição dos tetas. No entanto, o efeito da distribuição de teta sobre a calibração dos itens é pouco estudado. Adicionalmente, não se sabe

se o desenho amostral utilizado na calibração dos itens do Enem é capaz de recuperar adequadamente os parâmetros dos itens.

Outra etapa no desenvolvimento de uma CAT é a determinação das especificações do seu algoritmo. As simulações de CAT descritas nos parágrafos anteriores indicam que o controle de exposição pelo método PR aumenta a segurança do banco (diminui a superexposição, a subexposição e a sobreposição) sem comprometer a precisão da medida. Apesar de o aumento do parâmetro de aceleração desse método ter um impacto positivo na segurança de uma CAT de tamanho fixo, ele tem sido pouco estudado (e.g., Barrada et al., 2008), e seus efeitos em CATs de tamanho variável são desconhecidos.

Por isso, o objetivo deste trabalho foi desenvolver uma CAT do Enem mais eficiente, precisa e segura do que o formato atual do exame. A pergunta que pretendemos responder neste trabalho foi: é possível reduzir o tamanho do Enem e melhorar sua precisão e segurança com uma CAT? Dividimos o trabalho de acordo com seus quatro objetivos específicos, portanto em quatro partes. O primeiro objetivo específico, que foi o objetivo do primeiro artigo, foi comparar diferentes distribuições amostrais na calibração de itens no modelo 3PL. Elaboramos duas perguntas de pesquisa, apresentadas seguidas das hipóteses de estudo (H):

1. Qual desenho amostral recupera melhor os parâmetros dos itens em uma calibração? (H1) A amostra com distribuição retangular de acertos retorna parâmetros mais semelhantes aos parâmetros reais.
2. O desenho amostral utilizado para calibrar os itens do Enem 2020 (amostra deslocada) retorna parâmetros aceitáveis em comparação com outros desenhos? (H2) A amostra estratificada pelo número de acertos (deslocada) retorna parâmetros aceitáveis.

O segundo objetivo específico desta tese foi desenvolver um pacote de simulação de CAT. Por isso, a segunda parte desta tese é a apresentação do pacote simCAT, um pacote estatístico aberto desenvolvido em ambiente R e disponibilizado em repositório do github.

O segundo artigo desta tese contemplou o terceiro objetivo específico, que foi avaliar o controle de exposição PR com diferentes parâmetros de aceleração em uma CAT em termos de eficiência, precisão e segurança. As três perguntas de pesquisa e as hipóteses do estudo foram as seguintes:

3. Como a eficiência, a precisão e a segurança do teste variam ao se adotar o controle de exposição PR? (H3) A eficiência e a precisão do método PR serão semelhantes à do MIF; (H4) a segurança do PR será maior do que a do MIF.
4. Como a eficiência, a precisão e a segurança variam em função do parâmetro de aceleração do método PR? (H5) A eficiência e a precisão do método PR serão semelhantes com todos os parâmetros de aceleração; (H6) quanto maior o parâmetro de aceleração, maior a segurança.
5. Qual é o impacto sobre a precisão do Enem ao reduzir seu tamanho por meio de uma CAT com controle de exposição? (H7) A precisão de uma CAT com PR será maior do que a do teste linear.

O último objetivo desta tese foi publicar a CAT Enem com as especificações selecionadas a partir de simulações. Por isso, a quarta parte desta tese é a apresentação da CAT Enem, uma plataforma de aplicação aberta desenvolvida em ambiente R e disponibilizada em repositório do github.

Artigo 1: Efeito da distribuição amostral na calibração de itens no modelo 3PL

Alexandre Jaloto

Ricardo Primi

Resumo

A distribuição amostral de sujeitos tem sido considerada na calibração de itens, porém seu efeito foi pouco explorado. Adicionalmente, os estudos utilizaram tetas previamente conhecidas, o que não ocorre em testes com itens não calibrados. Por isso, investigamos a adequação de diferentes distribuições amostrais para a calibração dos itens do Exame Nacional do Ensino Médio (Enem). Utilizamos três desenhos amostrais (aleatório, retangular e deslocado) para simular a calibração do Enem 2020 com o método de máxima verossimilhança marginal. Não houve diferença relevante para a discriminação. Para a dificuldade, a amostra deslocada foi melhor em duas comparações. Para o pseudochute, a amostra retangular foi pior. Os resultados não apontam para a prevalência de um tipo de amostra para a calibração.
Palavras-chave: sorteio amostral; psicometria; avaliação educacional em larga escala; teoria de resposta ao item

Abstract

Sampling distribution of subjects has been considered in the calibration of items, but its effect has been little explored. Additionally, the studies used previously known thetas, which does not occur in tests with uncalibrated items. For this reason, we investigated the adequacy of different sample distributions for the calibration of the Brazilian High School Examination (Enem) items. We used three sample designs (random, rectangular and shifted) to simulate the Enem 2020 calibration with the marginal maximum likelihood method. There was no relevant difference for discrimination. For difficulty, the shifted sample was better in two comparisons. For the pseudo-guess, the rectangular sample was worse. The results do not point to the prevalence of one type of sample for calibration.
Keywords: sample draw; psychometrics; large-scale educational assessment; item response theory

Resumen

La distribución muestral de sujetos ha sido considerada en la calibración de ítems, pero su efecto ha sido poco explorado. Además, los estudios utilizaron tetas previamente conocidos, lo que no ocurre en las pruebas con ítems no calibrados. Por esta razón, investigamos la adecuación de diferentes distribuciones muestrales para la calibración de los ítems del Examen Brasileño de Educación Secundaria (Enem). Utilizamos tres diseños de muestra (aleatorio, rectangular y desplazado) para simular la calibración de Enem 2020 con el método de máxima verosimilitud marginal. No hubo diferencia relevante para la discriminación. Para la dificultad, la muestra desplazada fue mejor en dos comparaciones. Para la pseudo-azar, la muestra rectangular fue peor. Los resultados no apuntan a la prevalencia de un tipo de muestra para la calibración.
Palabras clave: sorteo muestral; psicometría; evaluación educativa a gran escala; teoría de la respuesta al ítem

Alguns aspectos são importantes de se considerar na calibração de itens com a Teoria de Resposta ao Item (TRI), como o tamanho amostral de sujeitos e a distribuição do teta (variável latente) na amostra. O tamanho amostral de cerca de 750 sujeitos tem sido razoável para calibrar itens em alguns contextos (e.g., Nunes & Primi, 2005; Şahin & Anıl, 2017; Şahin & Weiss, 2015). No entanto, o efeito da distribuição do teta na calibração tem sido pouco explorado (e.g., Paek et al., 2021; Wingersky & Lord, 1984). Adicionalmente, as investigações sobre calibração utilizam sujeitos com valores de teta já calculados pela TRI, o que não ocorre em situações de aplicação de testes educacionais com itens não calibrados. Diversos testes educacionais no Brasil e em outros países utilizam o arcabouço da TRI para estimar o teta dos indivíduos. Como exemplo, temos os testes do Sistema de Avaliação da Educação Básica (Saeb), do Exame Nacional do Ensino Médio (Enem), do Programa Internacional de Avaliação de Estudantes (Pisa, em inglês *Programme for International Student Assessment*) e da Avaliação Nacional do Progresso da Educação (Naep, em inglês *National Assessment of Educational Progress*). Sabe-se que uma calibração inadequada pode impactar a precisão da estimação do teta em regiões da escala com baixa informação psicométrica (Cheng & Yuan, 2010; Şahin & Weiss, 2015). Por isso, neste trabalho investigamos a adequação de diferentes distribuições amostrais para calibrar os itens do Enem, que utiliza o modelo logístico de três parâmetros da TRI (3PL). No Enem, um cálculo inadequado da nota pode causar uma seleção injusta de candidatos a vagas de cursos de ensino superior.

Não há um consenso sobre o tamanho amostral mínimo necessário para calibrar os itens adequadamente, e esse valor varia de acordo com o modelo de TRI, a quantidade de itens e a distribuição do teta da amostra. Nunes e Primi (2005) utilizaram dados de uma prova de matemática de 25 itens aplicada a alunos do 5º ano do ensino fundamental da região nordeste do Brasil. Os autores verificaram que amostras com no mínimo 200 pessoas retornavam

estimativas estáveis dos parâmetros dos itens e das pessoas. Akour e AL-Omari (2013) também utilizaram dados de uma prova de matemática, mas aplicada a alunos da 8ª série da Jordânia. Os autores recomendam um tamanho mínimo de 500 sujeitos para calibrar 30 itens. O estudo utilizou amostras com distribuição retangular de teta. Şahin e Weiss (2015) verificaram que em uma testagem adaptativa com banco de no mínimo 100 itens, uma amostra de 150 sujeitos para calibração foi suficiente para estimar o teta adequadamente. As amostras foram retangulares e as respostas aos itens foram simuladas. Şahin e Anıl (2017) utilizaram respostas de 6.288 sujeitos a um teste de língua inglesa para calibrar dez, 20 e 30 itens. As amostras foram estratificadas por instituição de ensino. Os autores verificaram que para o modelo 3PL, 750 sujeitos foram necessários para calibrar dez ou 20 itens. Para calibrar 30 itens, 350 sujeitos foram necessários.

As distribuições de teta utilizadas nesses trabalhos foram distintas, porém nenhum deles comparou como elas afetavam a qualidade da calibração, e poucos estudos investigaram esse tema. Wingersky e Lord (1984) utilizaram dados do TOEFL e verificaram que uma amostra com distribuição retangular de teta forneceu parâmetros com erros menores do que se a amostra tivesse o dobro de tamanho, mas distribuição em forma de sino. Os autores utilizaram o modelo 3PL. Na simulação de Paek et al. (2021), a distribuição de teta pouco afetou a calibração no modelo de Rasch. Os autores compararam a calibração de amostras aleatórias e enviesadas (formadas somente por sujeitos de baixo teta ou de alto teta) e verificaram que o tamanho amostral era um fator mais importante para garantir a invariância dos parâmetros do que sua distribuição. Já no modelo logístico de dois parâmetros (2PL), amostras enviesadas tiveram performance inferior às amostras aleatórias de mesmo tamanho. Os autores hipotetizaram que o parâmetro de discriminação tornou a estimação do parâmetro de dificuldade do item mais sensível ao viés da amostra. O estudo não verificou a adequação de amostras enviesadas para

calibrar itens no modelo 3PL. Além disso, as amostras enviesadas eram compostas exclusivamente por sujeitos da parte superior ou inferior da escala.

Esses estudos que investigaram o efeito da distribuição amostral sobre a calibração utilizaram o valor de teta da TRI para estratificar a população e sortear as amostras. No entanto, é comum em calibração de itens de testes educacionais a ausência desses valores para o sorteio da amostra, uma vez que para esse cálculo é necessário possuir os parâmetros dos itens. Por exemplo, no Enem a calibração dos itens não calibrados é feita a partir de uma amostra dos próprios participantes, antes de suas notas serem calculadas (Inep, 2021).

Desde 2009, o Enem tem sido um dos principais meios de acesso ao ensino superior no Brasil por meio do Sistema de Seleção Unificada (SiSU). O exame é composto por uma redação e quatro provas de 45 itens de múltipla escolha, a saber: Ciências Humanas (CH); Ciências da Natureza (CN); Linguagens e Códigos (LC); e Matemática (MT). Cada prova produz uma medida unidimensional por meio do modelo 3PL (MEC, 2009).

A amostra para calibração dos itens do Enem é estratificada pelo número de acertos e contém 100.000 participantes. Os três estratos são definidos pelos pontos do percentil 25 e percentil 95 de acertos dos participantes. Para o primeiro e o terceiro estratos são sorteados 25.000 participantes e para o segundo, 50.000. Para a prova de LC, metade de cada estrato é composta por participantes de cada língua estrangeira (inglês e espanhol).

Dentre os estudos citados nos parágrafos anteriores, o tamanho de 750 sujeitos foi o maior valor apontado como o mínimo para calibrar itens de um teste. Ainda, amostras com distribuição retangular forneceram boas calibrações e amostras enviesadas retornaram bons parâmetros no modelo Rasch, apesar de terem sido mais limitadas no modelo 2PL. Não se sabe sobre a adequação do uso de amostras enviesadas na calibração de itens no modelo 3PL, utilizado no Enem. Cabe destacar que a estratificação da população por teta foi feita a partir do

valor estimado na TRI, o que só é possível quando já se tem os parâmetros dos itens, como em situação de simulação.

Dadas as lacunas apontadas, o objetivo deste trabalho foi comparar diferentes distribuições amostrais na calibração de itens no modelo 3PL. Elaboramos duas perguntas de pesquisa, apresentadas seguidas das hipóteses de estudo:

1. Qual desenho amostral recupera melhor os parâmetros dos itens em uma calibração? (H1) A amostra com distribuição retangular de acertos retorna parâmetros mais semelhantes aos parâmetros reais.
2. O desenho amostral utilizado para calibrar os itens do Enem 2020 (amostra deslocada) retorna parâmetros aceitáveis em comparação com outros desenhos? (H2) A amostra estratificada pelo número de acertos (deslocada) retorna parâmetros aceitáveis.

Neste trabalho, avançamos ao verificar a adequação de uma amostra enviesada na calibração de itens no modelo 3PL. Ainda, incluímos nessa amostra sujeitos de todas as regiões da escala, em vez de exclusivamente de uma região. Outro avanço é que neste estudo estratificamos a população de acordo com a soma de acertos e não com o teta estimado pela TRI. Nesse ponto, nos aproximamos de uma situação real de calibração dos itens de um teste educacional como o Enem. Por último, este estudo pretende preencher uma lacuna de produção científica sobre o efeito da distribuição amostral na calibração dos itens, tema pouco abordado apesar de se saber que a diferença nessa distribuição tem relação com a qualidade dos parâmetros dos itens.

Método

Desenho da simulação

Neste estudo avaliamos a calibração por meio da manipulação do desenho da amostra (aleatória, retangular ou deslocada). A simulação foi feita nas quatro áreas do Enem, portanto

tivemos $3 \times 4 = 12$ condições (Tabela 6). Cada condição foi replicada 100 vezes, portanto tivemos $12 \times 100 = 1200$ casos. Os procedimentos foram realizados no ambiente R de programação (R Core Team, 2019) e os comandos utilizados estão disponíveis no repositório http://www.github.com/alexandrejaloto/tese_artigo1.

Tabela 6

Condições da simulação de calibração

Área	Tipo de amostra		
	Aleatória	Retangular	Deslocada
CH	*	*	*
CN	*	*	*
LC	*	*	*
MT	*	*	*

Dados

Os dados utilizados neste estudo são secundários e foram extraídos em fevereiro de 2022 dos microdados do Enem, disponibilizados no portal do Inep (2023b). Foi gerado um banco de dados para cada área. Para compor cada um, selecionamos os seguintes participantes: os que responderam pelo menos uma questão da respectiva prova (a resposta em branco é representada por "." na variável TX_RESPOSTA); e os que responderam prova sem adaptação (as provas adaptadas, por exemplo tamanho ampliado, possuem código específico na variável CO_PROVA).

Tipo de sorteio da amostra. As amostras de calibração foram sorteadas do banco de dados simulado. O procedimento da simulação está descrito mais adiante. Foram três tipos de desenho amostral, todos com tamanho de 5.040 sujeitos. Seriam 5.000, porém por conta de arredondamentos do segundo e terceiro tipos de desenho, esse número foi ampliado para todos os desenhos. Para fins de replicação, cada desenho amostral foi sorteado 100 vezes. O desenho amostral aleatório simples objetivou gerar amostras com a mesma distribuição da população do Enem 2020. Foram sorteados 5.040 sujeitos aleatoriamente. Na área de LC, sorteamos 2.520

sujeitos que responderam língua espanhola e 2.520 sujeitos que responderam língua inglesa. Chamamos essa amostra de aleatória.

O desenho uniformemente estratificado objetivou garantir que a amostra contivesse sujeitos em todas as faixas de acerto em proporções semelhantes. Para isso, primeiro dividimos a população em nove estratos que correspondiam a faixas de acerto com intervalo de cinco pontos (5, 10, 15, 20, 25, 30, 40 e 45). Em seguida sorteamos o mesmo número de pessoas de cada estrato. Cada estrato foi composto por 560 sujeitos. Sempre que um estrato tivesse menos do que 560 sujeitos no arquivo de banco, ele foi complementado com o estrato imediatamente superior, caso possível. Caso não fosse possível, o complemento era feito com o superior seguinte, até que o quantitativo fosse atingido. No caso de o último estrato não possuir 560 sujeitos, o complemento foi realizado com o estrato imediatamente inferior. Se não fosse possível, o complemento foi feito com o inferior seguinte. Na área de LC, sorteamos 280 sujeitos que responderam língua espanhola e 280 sujeitos que responderam língua inglesa de cada estrato. Chamamos essa amostra de retangular.

O desenho amostral estratificado também gerou uma amostra estratificada, porém os pontos de corte dos estratos foram diferentes. Estabelecemos três estratos de acordo com a soma de acertos, separados pelos percentis 25 e 95, por meio da função *quantile*. Sorteamos 1.260 participantes dos estratos inferior e superior, e 2.520 participantes do estrato intermediário. Esse desenho garante que a amostra tenha alta proporção de participantes com teta alto. Ele foi utilizado neste trabalho porque é o mesmo que o Inep utiliza para calibrar os itens do Enem (Inep, 2021). Na área de LC, sorteamos metade dos valores indicados em cada estrato para cada língua estrangeira. Chamamos essa amostra de deslocada.

Transformação das escalas

Neste estudo, precisamos realizar transformações nas escalas dos parâmetros dos itens e das notas dos sujeitos, por dois motivos. Primeiro, porque as notas dos participantes e os

parâmetros dos itens divulgados nos microdados do Enem estão posicionados em escalas que possuem referências diferentes. A referência da escala das notas do Enem são os concluintes regulares de escola pública do Enem 2009 (amostra normativa). Essa escala, que é a oficial do Enem, possui média 500 e desvio padrão 100. A referência da escala dos parâmetros divulgados é a amostra utilizada para a calibração dos itens no Enem 2009 (amostra da primeira calibração). Essa escala possui média 0 e desvio padrão 1. E segundo, porque cada replicação realizada neste estudo teve uma referência na calibração, que foi a amostra sorteada, e gerou uma escala com média 0 e desvio padrão 1. Ou seja, neste estudo trabalhamos com escalas de três grupos de referências diferentes (amostra normativa, amostra de calibração do Enem 2009 e cada amostra sorteada nas replicações da simulação).

A consequência da diferença das referências das métricas é que a comparabilidade dos parâmetros fica comprometida. Isso ocorre porque há um deslocamento da escala. É como se comparássemos a mesma temperatura medida em Celsius e em Fahrenheit. Para fazer essa comparação, é preciso que ambos os valores estejam na mesma escala, por exemplo ao transformar a medida em Celsius para Fahrenheit. No nosso caso, é como se estimássemos as notas com os itens posicionados na métrica dos microdados divulgados e posteriormente com os itens posicionados na métrica oficial do Enem. Apesar da correlação entre os dois conjuntos de notas ser 1,0, a média da diferença absoluta (MDA) e a raiz do erro quadrático médio (REQM) não serão zero. Por isso, para comparar os parâmetros calibrados neste trabalho com os divulgados, foi preciso posicionar os parâmetros dos itens e as notas na mesma escala.

Para garantir a comparabilidade, posicionamos os itens e os sujeitos na métrica oficial do Enem padronizada com média 0 e desvio padrão 1. Ou seja, a referência da escala foi a amostra normativa, porém padronizada. Poderíamos utilizar a escala oficial do Enem, porém transformar a escala para média 0 e desvio padrão 1 facilita as análises, pois as distribuições prévias da calibração e da estimação da proficiência se baseiam em uma escala padronizada.

De forma resumida, as mudanças nas escalas envolveram a sua transformação para a escala oficial do Enem (média 500 e desvio padrão 100) e em seguida a padronização (média 0 e desvio padrão 1). Nos próximos parágrafos, descrevemos as equações utilizadas nessas transformações, que tiveram como base o método de equalização média-sigma (Hambleton et al., 1991).

A título de explicação, primeiramente utilizaremos X e x para nos referir a parâmetros posicionados em uma escala original, e Y e y para nos referir a parâmetros posicionados na escala transformada de interesse (escala oficial do Enem, com média 500 e desvio padrão 100). Nesse caso, X_i corresponde à nota do sujeito i na escala X e Y_i corresponde à sua nota na escala Y. Apesar de numericamente diferentes, teoricamente essas notas representam a mesma magnitude no traço medido (por exemplo, proficiência em Matemática). Por isso, a padronização dessas notas deveria resultar em valores iguais, e podemos expressar essa igualdade assim:

$$\frac{Y_i - \bar{Y}}{DP_y} = \frac{X_i - \bar{X}}{DP_x} \tag{13}$$

Onde $\bar{Y}$ e DP_y representam a média e o desvio padrão das notas da amostra na escala Y, e $\bar{X}$ e DP_x representam a média e o desvio padrão das notas da amostra na escala X. Nessa equação, consideramos que as notas padronizadas das duas métricas são iguais, pois provêm da mesma amostra. Porém, na prática elas não são exatamente iguais, pois em uma calibração é pouco provável que a distribuição de teta da amostra selecionada tenha média exatamente zero e desvio padrão exatamente 1, apesar de os programas (como o pacote mirt) assumirem esses valores para identificação da métrica. Se isolarmos Y_i, teremos (Muñiz, 1997):

$$Y_i = \frac{DP_y}{DP_x} X_i + \left[\bar{Y} - \frac{DP_y}{DP_x} \bar{X}\right] \tag{14}$$

As constantes *k* e *d*

$$k = \frac{DP_y}{DP_x} \tag{15}$$

$$d = \bar{Y} - k\bar{X} \tag{16}$$

extraídas da Equação 14 representam as constantes de transformação (escala e origem respectivamente) das notas dos sujeitos e dos parâmetros dos itens da escala X para a escala Y. Portanto, para transformar a nota da escala X para a escala Y, utilizamos a seguinte equação:

$$Y_i = kX_i + d \tag{17}$$

Como as notas dos sujeitos e os parâmetros dos itens estão na mesma métrica, podemos usar essas constantes para transformarmos os parâmetros b e a do item j da métrica X para a métrica Y. Nesse caso, usamos as seguintes equações lineares:

$$b_{jy} = kb_{jx} + d \tag{18}$$

$$a_{jy} = \frac{a_{jx}}{k} \tag{19}$$

Onde a_{jy} e b_{jy} são os parâmetros a e b do item j na métrica Y, e a_{jx} e b_{jx} são os valores desses parâmetros na escala X.

Para transformar as notas e os parâmetros dos itens da escala oficial do Enem para a escala padronizada do Enem, aplicamos as seguintes equações:

$$\theta_{i\,01} = \frac{Y_i - 500}{100} \tag{20}$$

$$b_{j\,01} = \frac{b_{jy} - 500}{100} \tag{21}$$

$$a_{j\,01} = a_{jy} * 100 \tag{22}$$

Onde $\theta_{i\,01}$, $b_{j\,01}$ e $a_{j\,01}$ são a nota do sujeito i e os parâmetros a e b do item j na métrica oficial do Enem padronizada (média 0 e desvio padrão 1), e Y_i, b_{jy} e a_{jy} são a nota e os parâmetros na métrica oficial do Enem (média 500 e desvio padrão 100).

À exceção das notas divulgadas, que estavam na escala oficial do Enem, as transformações das escalas passaram pelos seguintes passos:

1. Igualdade entre a padronização das notas da amostra na escala a ser transformada e na escala oficial do Enem
2. Obtenção das constantes de transformação
3. Transformação para a escala oficial do Enem (média 500 e desvio padrão 100)
4. Transformação para a escala oficial do Enem padronizada (média 0 e desvio padrão 1)

Transformação dos parâmetros oficiais

Como mencionado, os parâmetros divulgados nos microdados encontram-se em uma métrica com média 0 e desvio padrão 1 cuja referência é a amostra de calibração do Enem 2009. Para levá-los à métrica oficial do Enem, cuja referência é a amostra normativa, foi preciso obter as constantes de transformação *k* e *d*, que não eram 100 e 500.

Para calcular os valores de *k* e *d*, inicialmente estimamos a proficiência dos primeiros 300.000 sujeitos do banco dos microdados em cada área. A estimação foi feita da mesma maneira que estimamos a proficiência após as calibrações e será descrita mais à frente. Em seguida, aplicamos as Equações 15 e 16 para obtermos as constantes de transformação da escala dos parâmetros divulgados para a escala oficial. A Tabela 7 apresenta os valores das constantes para cada área.

Tabela 7

Constantes de transformação dos parâmetros divulgados nos microdados para a escala oficial do Enem

Área	k	d
CH	501,489	112,310
CN	501,144	113,102
LC	499,978	108,086
MT	500,020	129,646

Assim, a transformação para a métrica oficial do Enem foi a seguinte:

$$b_{jy} = kb_{j\,div} + d \tag{23}$$

$$a_{jy} = \frac{a_{j\,div}}{k} \tag{24}$$

Onde b_{jy} e a_{jy} são os parâmetros a e b do item j na métrica oficial do Enem, cuja referência é a amostra normativa, e $a_{j\,div}$ e $b_{j\,div}$ são os valores desses parâmetros na métrica divulgada nos microdados, cuja referência é a amostra de calibração de 2009.

Após posicionarmos os itens na escala oficial do Enem, transformamos a escala para média 0 e desvio padrão 1 a partir das Equações 21 e 22. A Tabela 8 apresenta os parâmetros divulgados, os parâmetros transformados e o ponto na escala correspondente ao decil do item, de CH e CN. Para obter o ponto de cada decil, ordenamos os itens da prova segundo sua dificuldade e verificamos os pontos de corte de cada decil. A Tabela 9 apresenta os parâmetros de LC e MT.

Tabela 8

Parâmetros divulgados e transformados dos itens de Ciências Humanas e Ciências da Natureza

Item	CH				CN			
	a_{div} (a_{01})	b_{div} (b_{01})	c	decil	a_{div} (a_{01})	b_{div} (b_{01})	c	decil
1	3,533 (3,146)	1,341 (1,521)	0,137	1,787	1,854 (1,639)	0,631 (0,725)	0,178	0,773
2	2,478 (2,206)	1,829 (2,069)	0,309	2,106	1,712 (1,514)	-0,075 (-0,074)	0,174	0,599
3	2,158 (1,922)	1,830 (2,070)	0,134	2,106	2,744 (2,426)	1,554 (1,769)	0,186	1,884
4	3,114 (2,773)	1,338 (1,518)	0,158	1,519	4,355 (3,850)	1,092 (1,247)	0,147	1,254
5	1,339 (1,192)	-0,247 (-0,262)	0,202	0,207	2,915 (2,577)	1,124 (1,282)	0,112	1,413
6	1,048 (0,933)	0,551 (0,634)	0,219	0,840	3,012 (2,663)	1,502 (1,711)	0,187	1,711
7	2,107 (1,876)	0,739 (0,845)	0,148	1,133	1,991 (1,760)	0,674 (0,774)	0,187	0,969
8	1,939 (1,726)	0,768 (0,877)	0,169	1,133	2,298 (2,032)	1,503 (1,711)	0,131	1,884
9	3,864 (3,440)	0,089 (0,115)	0,131	0,207	3,214 (2,842)	1,908 (2,169)	0,172	2,347
10	1,633 (1,454)	3,518 (3,966)	0,123	3,966	2,398 (2,120)	1,699 (1,934)	0,190	2,026
11	1,731 (1,541)	1,001 (1,140)	0,270	1,210	2,404 (2,126)	0,720 (0,826)	0,213	0,969
12	2,049 (1,824)	2,080 (2,350)	0,119	3,966	3,312 (2,928)	1,884 (2,142)	0,139	2,347
13	1,340 (1,194)	1,036 (1,178)	0,197	1,210	2,275 (2,012)	1,858 (2,113)	0,263	2,347
14	1,304 (1,161)	1,051 (1,195)	0,232	1,210	3,038 (2,686)	0,624 (0,717)	0,208	0,773
15	1,947 (1,733)	1,073 (1,220)	0,329	1,384	0,295 (0,261)	3,800 (4,310)	0,027	4,310
16	1,412 (1,257)	1,608 (1,821)	0,211	1,915	3,172 (2,805)	1,357 (1,546)	0,218	1,711
17	2,201 (1,959)	1,674 (1,895)	0,241	1,915	0,505 (0,447)	0,958 (1,094)	0,042	1,254
18	4,436 (3,950)	1,680 (1,902)	0,270	1,915	1,295 (1,145)	1,179 (1,345)	0,159	1,413
19	3,041 (2,708)	1,738 (1,967)	0,297	2,106	1,454 (1,286)	1,677 (1,908)	0,140	2,026
20	2,997 (2,669)	1,181 (1,342)	0,188	1,384	2,117 (1,872)	1,594 (1,814)	0,092	1,884
21	1,281 (1,141)	-0,038 (-0,028)	0,187	0,207	2,977 (2,632)	1,695 (1,928)	0,295	2,026
22	1,639 (1,460)	3,229 (3,641)	0,142	3,966	1,647 (1,456)	0,850 (0,972)	0,175	1,254
23	2,051 (1,826)	0,960 (1,093)	0,143	1,133	3,261 (2,883)	0,475 (0,549)	0,249	0,599

24	2,320 (2,065)	1,589 (1,800)	0,119	1,915	2,283 (2,019)	2,629 (2,985)	0,244	4,310
25	1,925 (1,714)	1,502 (1,702)	0,121	1,787	1,794 (1,587)	0,792 (0,907)	0,207	0,969
26	1,722 (1,534)	0,225 (0,268)	0,221	0,840	1,829 (1,617)	2,088 (2,374)	0,184	4,310
27	2,148 (1,912)	2,285 (2,582)	0,171	3,966	2,118 (1,873)	1,572 (1,789)	0,144	1,884
28	5,866 (5,223)	1,029 (1,170)	0,115	1,210	1,000 (0,884)	1,194 (1,362)	0,213	1,413
29	2,633 (2,344)	1,211 (1,375)	0,281	1,384	1,701 (1,504)	1,151 (1,313)	0,147	1,413
30	2,840 (2,528)	1,084 (1,232)	0,177	1,384	2,820 (2,494)	2,010 (2,284)	0,124	2,347
31	1,656 (1,474)	1,874 (2,120)	0,196	3,966	1,883 (1,665)	1,487 (1,694)	0,270	1,711
32	2,110 (1,879)	1,582 (1,792)	0,231	1,915	2,907 (2,571)	0,128 (0,156)	0,210	0,599
33	3,980 (3,544)	0,789 (0,901)	0,190	1,133	1,929 (1,706)	2,723 (3,092)	0,177	4,310
34	1,682 (1,498)	0,134 (0,166)	0,176	0,207	2,545 (2,250)	0,324 (0,378)	0,184	0,599
35	1,637 (1,458)	1,317 (1,495)	0,189	1,519	1,575 (1,392)	1,653 (1,882)	0,220	1,884
36	2,115 (1,883)	1,843 (2,085)	0,129	2,106	2,215 (1,959)	0,671 (0,770)	0,173	0,773
37	2,740 (2,439)	0,994 (1,131)	0,264	1,133	0,742 (0,656)	0,632 (0,726)	0,011	0,773
38	2,308 (2,055)	1,492 (1,691)	0,220	1,787	2,940 (2,599)	1,730 (1,969)	0,232	2,026
39	2,205 (1,963)	1,564 (1,771)	0,140	1,787	2,618 (2,315)	0,816 (0,935)	0,058	0,969
40	2,178 (1,939)	1,219 (1,384)	0,265	1,384	1,006 (0,890)	2,269 (2,578)	0,121	4,310
41	3,553 (3,163)	1,234 (1,401)	0,184	1,519	3,993 (3,531)	0,946 (1,082)	0,167	1,254
42	3,718 (3,310)	1,259 (1,429)	0,169	1,519	1,469 (1,299)	-0,235 (-0,255)	0,199	0,599
43	1,027 (0,914)	0,109 (0,137)	0,196	0,207	2,955 (2,613)	1,284 (1,464)	0,164	1,711
44	2,785 (2,480)	0,558 (0,642)	0,198	0,840	2,470 (2,184)	0,951 (1,087)	0,131	1,254
45	1,027 (0,915)	0,716 (0,819)	0,174	0,840	-	-	-	-

Nota. CH = Ciências Humanas; CN = Ciências da Natureza; a_{div} = parâmetro de discriminação divulgado nos microdados; a_{01} = parâmetro de discriminação na métrica oficial do Enem padronizada; b_{div} = parâmetro de dificuldade divulgado nos microdados; b_{01} = parâmetro de dificuldade na métrica oficial do Enem padronizada; c = parâmetro de pseudochute; decil = ponto na escala correspondente ao decil do item, após ordenar os itens pela sua dificuldade; - = item anulado.

Tabela 9

Parâmetros divulgados e transformados dos itens de Linguagens e Códigos e Matemática

Item	LC				MT			
	a_{div} (a_{01})	b_{div} (b_{01})	c	decil	a_{div} (a_{01})	b_{div} (b_{01})	c	decil
1	2,315 (1,786)	0,452 (0,586)	0,171	0,700	1,501 (1,158)	2,331 (3,022)	0,122	3,244
2	3,077 (2,373)	0,195 (0,253)	0,166	0,347	1,587 (1,224)	-0,086 (-0,111)	0,201	1,096
3	1,813 (1,398)	2,023 (2,623)	0,121	7,068	3,351 (2,585)	3,342 (4,333)	0,166	4,333
4	2,482 (1,915)	1,238 (1,605)	0,165	1,616	2,756 (2,126)	0,624 (0,810)	0,092	1,096
5	2,301 (1,775)	0,769 (0,998)	0,173	1,031	1,181 (0,911)	2,233 (2,895)	0,207	2,905
6	1,110 (0,857)	0,907 (1,177)	0,126	1,371	1,738 (1,341)	2,680 (3,475)	0,140	3,666
7	2,641 (2,037)	1,657 (2,149)	0,261	2,287	2,092 (1,614)	0,994 (1,289)	0,133	1,424
8	2,609 (2,012)	1,224 (1,587)	0,195	1,616	1,516 (1,170)	1,923 (2,493)	0,230	2,670
9	2,088 (1,611)	0,681 (0,883)	0,158	1,031	3,254 (2,510)	1,577 (2,045)	0,136	2,164
10	1,584 (1,222)	0,696 (0,903)	0,084	1,031	2,731 (2,107)	2,499 (3,239)	0,135	3,244

11	3,188 (2,459)	1,259 (1,632)	0,174	2,010	3,296 (2,543)	2,862 (3,710)	0,255	4,333
12	2,154 (1,661)	1,146 (1,485)	0,106	1,616	2,265 (1,747)	2,777 (3,600)	0,235	3,666
13	1,729 (1,334)	1,851 (2,399)	0,149	2,501	2,961 (2,284)	0,896 (1,162)	0,185	1,424
14	1,521 (1,173)	1,918 (2,487)	0,185	2,501	2,887 (2,227)	2,823 (3,660)	0,140	3,666
15	1,107 (0,854)	1,640 (2,126)	0,255	2,287	2,137 (1,649)	1,782 (2,310)	0,156	2,422
16	1,930 (1,489)	1,024 (1,328)	0,142	1,371	2,990 (2,306)	1,890 (2,451)	0,159	2,670
17	1,742 (1,343)	1,512 (1,960)	0,244	2,010	2,947 (2,273)	2,066 (2,678)	0,121	2,905
18	1,610 (1,241)	0,270 (0,350)	0,205	0,700	1,084 (0,836)	0,843 (1,094)	0,015	1,096
19	1,572 (1,213)	2,746 (3,561)	0,222	7,068	2,183 (1,684)	2,311 (2,996)	0,149	3,244
20	2,909 (2,244)	-0,408 (-0,529)	0,129	0,012	2,597 (2,003)	1,340 (1,738)	0,154	2,164
21	2,420 (1,867)	1,104 (1,432)	0,149	1,616	2,047 (1,579)	1,082 (1,403)	0,179	1,424
22	1,686 (1,301)	0,306 (0,397)	0,146	0,700	-	-	-	-
23	1,356 (1,046)	1,445 (1,873)	0,235	2,010	1,782 (1,374)	3,299 (4,277)	0,189	4,333
24	2,863 (2,208)	1,449 (1,879)	0,214	2,010	0,966 (0,745)	2,506 (3,250)	0,205	3,666
25	3,215 (2,480)	0,546 (0,708)	0,164	1,031	1,758 (1,356)	2,408 (3,121)	0,166	3,244
26	0,359 (0,277)	5,451 (7,068)	0,163	7,068	4,063 (3,134)	2,180 (2,827)	0,216	2,905
27	1,069 (0,825)	-0,440 (-0,570)	0,014	0,012	2,145 (1,655)	1,736 (2,251)	0,226	2,422
28	2,787 (2,150)	1,753 (2,272)	0,150	2,287	1,570 (1,211)	2,177 (2,823)	0,114	2,905
29	3,323 (2,563)	1,492 (1,935)	0,186	2,010	2,312 (1,783)	0,441 (0,572)	0,189	1,096
30	1,617 (1,247)	1,839 (2,384)	0,261	2,501	2,597 (2,003)	2,830 (3,669)	0,097	4,333
31	1,598 (1,233)	-0,167 (-0,217)	0,196	0,012	2,590 (1,998)	2,034 (2,637)	0,184	2,670
32	1,737 (1,340)	-0,119 (-0,154)	0,051	0,012	2,006 (1,548)	1,154 (1,497)	0,141	1,737
33	1,161 (0,896)	1,007 (1,306)	0,219	1,371	3,199 (2,467)	1,608 (2,085)	0,172	2,164
34	1,627 (1,255)	0,535 (0,694)	0,102	0,700	1,895 (1,462)	1,652 (2,142)	0,280	2,164
35	2,174 (1,677)	1,807 (2,344)	0,147	2,501	1,221 (0,942)	2,109 (2,735)	0,090	2,905
36	1,407 (1,085)	0,812 (1,053)	0,175	1,371	1,340 (1,033)	1,935 (2,509)	0,187	2,670
37	1,288 (0,993)	0,075 (0,097)	0,013	0,347	2,732 (2,107)	1,770 (2,295)	0,141	2,422
38	1,844 (1,422)	0,862 (1,118)	0,249	1,371	2,424 (1,870)	3,020 (3,915)	0,154	4,333
39	2,149 (1,657)	2,255 (2,924)	0,256	7,068	1,539 (1,187)	1,603 (2,079)	0,197	2,164
40	1,774 (1,369)	0,024 (0,031)	0,187	0,347	2,759 (2,128)	1,152 (1,494)	0,152	1,737
41	1,705 (1,315)	0,362 (0,470)	0,178	0,700	1,201 (0,926)	0,850 (1,102)	0,164	1,424
42	1,439 (1,110)	1,090 (1,414)	0,125	1,616	1,257 (0,969)	-0,404 (-0,523)	0,204	1,096
43	1,483 (1,144)	1,914 (2,482)	0,165	2,501	0,934 (0,720)	1,109 (1,438)	0,074	1,737
44	2,208 (1,703)	0,259 (0,336)	0,172	0,347	2,993 (2,309)	1,847 (2,394)	0,064	2,422
45	1,221 (0,942)	1,683 (2,182)	0,212	2,287	3,025 (2,333)	1,334 (1,730)	0,115	1,737

Nota. LC = Linguagens e Códigos; MT = Matemática; a_{div} = parâmetro de discriminação divulgado nos microdados; a_{01} = parâmetro de discriminação na métrica oficial do Enem padronizada; b_{div} = parâmetro de dificuldade divulgado nos microdados; b_{01} = parâmetro de dificuldade na métrica oficial do Enem padronizada; c = parâmetro de pseudochute; decil = ponto na escala correspondente ao decil do item, após ordenar os itens pela sua dificuldade; - = item anulado.

Simulação das respostas

A simulação foi do tipo Monte Carlo. Nesse tipo de simulação, sorteia-se um número aleatório de uma distribuição uniforme entre zero e um e compara-se com a probabilidade de o

sujeito acertar um item (dados seus parâmetros). Se o número for maior ou igual à probabilidade, atribui-se erro. Se for menor, atribui-se acerto. A nota e os parâmetros dos itens utilizados para calcular a probabilidade de acerto estavam na métrica oficial do Enem padronizada.

Apesar de ser possível acessar as respostas dos participantes do Enem, optamos por simular as respostas para que os padrões de resposta seguissem os postulados do modelo de TRI adotado pelo Enem. Dessa maneira, tivemos uma linha de base em que a probabilidade de acerto dos itens era dada exclusivamente pelos parâmetros (dos itens e do sujeito) relacionados a somente uma dimensão. Ou seja, os padrões de respostas eram explicados pela variável latente do modelo e não foram influenciados por dimensões adicionais como a fadiga e a motivação.

Transformação dos parâmetros calibrados

Para que os resultados de todas as replicações de calibrações de uma mesma área fossem comparáveis, foi necessário posicionar os itens em uma mesma escala, no caso a métrica oficial do Enem padronizada. Para isso, transformamos os parâmetros calibrados (escala mirt) para a escala oficial do Enem e em seguida para a escala do Enem padronizada.

Primeiro, obtivemos as constantes de transformação da escala mirt para a escala oficial seguindo as Equações 15 e 16. Os parâmetros dos itens foram posicionados na métrica oficial do Enem seguindo as Equações 18 e 19. Após posicionar os itens na escala oficial do Enem, transformamos a escala para média 0 e desvio padrão 1 por meio das Equações 21 e 22. Ao posicionar os itens calibrados de uma área na mesma escala, foi possível realizar comparações diretas entre os parâmetros, como calcular a diferença absoluta entre o parâmetro estimado e o real.

Calibração e estimação da proficiência

A calibração foi realizada por meio do pacote mirt (v1.33.2; Chalmers, 2012). A distribuição prévia do parâmetro de discriminação foi log-normal de média 0 e desvio 0,5 (o

que garante valores positivos). Já para o pseudochute, foi adotada uma distribuição beta com parâmetros 7 e 28. Dessa forma, a distribuição prévia desse parâmetro ficou centrada em 0,2, o que é adequado para itens de cinco alternativas. O método de estimação foi o de máxima verossimilhança marginal (assumindo a distribuição de teta com média zero e desvio padrão um) e o critério de convergência foi 0,01. O comando utilizado foi o que segue:

```
model <- mirt.model(
             'F1 = 1-45
             PRIOR = (1-45, a1, lnorm, 0, 0.5),(1-45, g, expbeta, 7, 28)'
      )
calib <- mirt(sample, model, itemtype = '3PL', TOL = .01)
```

onde 45 corresponde à quantidade de itens na prova. Em CN e MT um item foi anulado, por isso foram 44 itens e em LC há cinco itens de cada língua estrangeira, portanto a quantidade é 50. O objeto *sample* contém a amostra sorteada para calibração.

A estimação da proficiência foi feita por meio da função *fscores* com o método EAP e 40 pontos de quadratura entre -4 e 4, valores utilizados pelo Inep para a estimação das notas do Enem 2020 (Inep, 2021). O comando utilizado foi o seguinte:

```
scores <- fscores (model, qdpts = 40, theta_lim = c(-4, 4))
```

Avaliação das calibrações

A avaliação da calibração foi feita por meio da MDA, do viés, da REQM e da correlação entre os parâmetros estimados e os parâmetros oficiais. A MDA foi calculada para cada item da seguinte maneira:

$$MDA_j = \frac{\sum_r \left| \widehat{par}_{jr} - par_j \right|}{R} \tag{25}$$

onde R é o total de replicações, $\widehat{par}_{jr}$ é o parâmetro estimado do item j na replicação r e par_j é o parâmetro oficial do item j. A MDA geral de cada condição correspondeu à média das MDA dos itens calibrados nessa condição. A MDA condicionada à dificuldade correspondeu à média da MDA dos itens agrupados de acordo com o decil de dificuldade. Esse agrupamento teve quatro ou cinco itens (aproximadamente 10% do total de itens da prova). Para calcular a MDA

condicional, ordenamos os itens da prova segundo sua dificuldade, verificamos os pontos de corte de cada decil e agrupamos os itens. Em seguida, calculamos a média da MDA dos itens de cada agrupamento. Os decis dos itens de CH e CN estão apresentados na Tabela 8 e os dos itens de LC e MT, na Tabela 9.

O viés de cada item foi calculado da seguinte maneira:

$$V_j = \frac{\sum_r \widehat{par}_{jr}}{R} - par_j \tag{26}$$

O viés geral da condição correspondeu à média dos vieses dos itens da área. A REQM de cada item foi calculada da seguinte maneira:

$$REQM_j = \sqrt{\frac{\sum_{j=1}^{n} (\widehat{par}_{jr} - par_j)^2}{R}} \tag{27}$$

A REQM geral da condição correspondeu à média dos valores de REQM dos itens da área. A correlação geral da condição foi calculada entre a média das estimativas do parâmetro de cada item nas 100 replicações e o parâmetro real.

Como utilizamos o modelo 3PL, para cada condição tivemos três medidas de cada índice, sendo uma para cada parâmetro (discriminação, dificuldade e pseudochute). Além do cálculo desses índices, construímos três modelos de ANOVA para cada área do conhecimento, um para cada parâmetro de item. A variável dependente foi a MDA do item na condição analisada (calculada com a Equação 25) e a variável independente de cada modelo foi o tipo de amostra (aleatória, retangular ou deslocada). Todos os modelos construídos utilizaram a correção de Welch, dado que os dados feriram os pressupostos da ANOVA. Nos modelos estatisticamente significativos, a significância estatística da comparação par a par dos tipos de amostra foi verificada com a correção de Bonferroni. Ainda, calculou-se o tamanho do efeito e verificou-se se seu intervalo de confiança incluía o zero por meio de 1.000 reamostragens (*bootstrap*).

Resultados

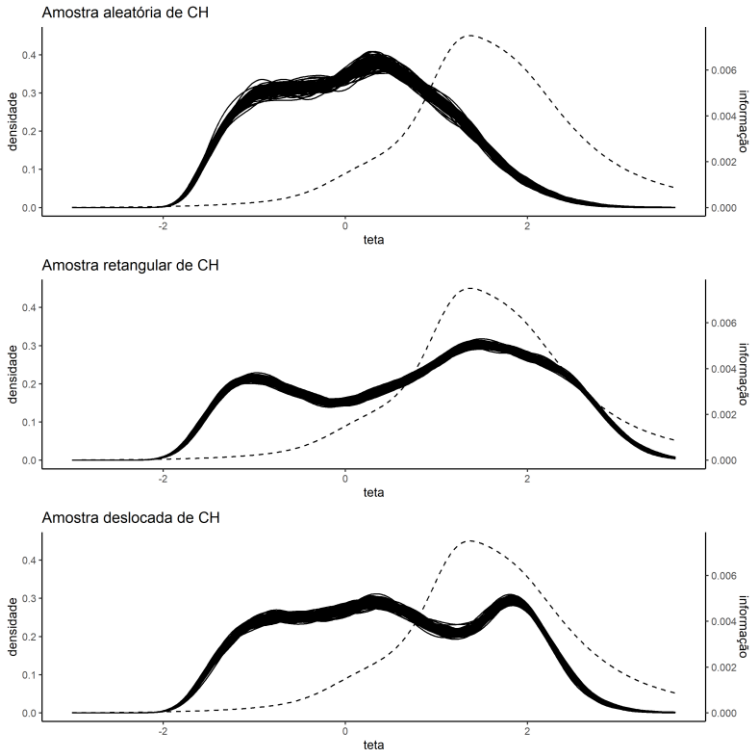
Amostras sorteadas

A Figura 4 mostra a distribuição das notas oficiais padronizadas das 100 amostras de cada desenho em CH e a curva de informação do teste. A Figura 5 mostra essas informações referentes à prova de CN, a Figura 6, à prova de LC e a Figura 7, à prova de MT. Nota-se que os três desenhos amostrais geraram curvas de distribuição de teta diferentes dentro da mesma área. Além disso, de uma maneira geral as amostras não abrangeram a região dos testes com itens mais difíceis.

Em CH, as amostras aleatórias tiveram um pico de teta próximo de 0,50, as retangulares tiveram dois picos (próximo de -1,00 e 1,50) e as deslocadas tiveram uma distribuição mais uniforme, com um vale próximo de 1,00. As amostras retangulares e deslocadas tiveram uma proporção de sujeitos maior na região com mais itens do que a amostra aleatória.

Figura 4

Curva de densidade das notas das 100 amostras sorteadas em Ciências Humanas para cada desenho amostral e curva de informação do teste

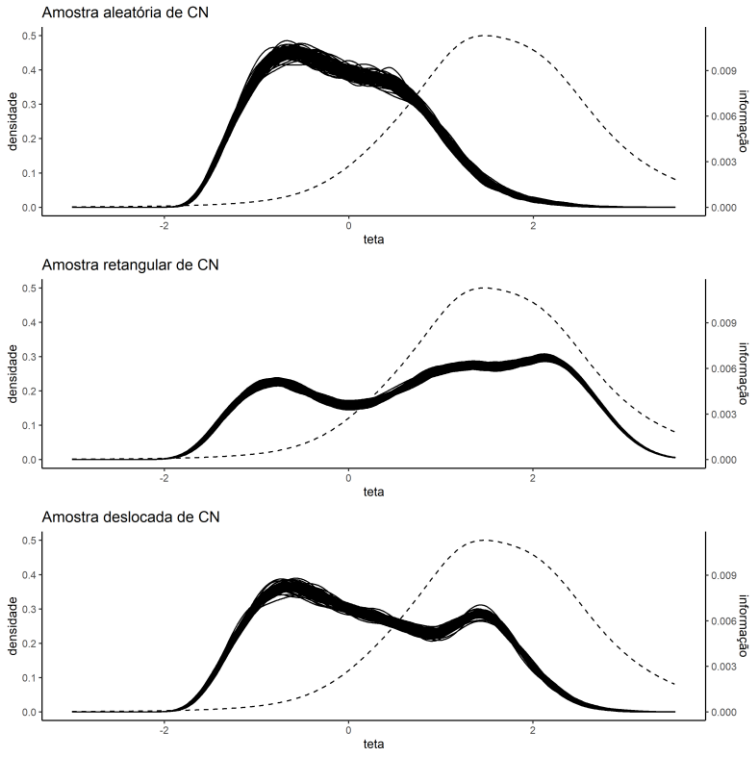


Nota. a linha pontilhada se refere à informação psicométrica da prova e as linhas contínuas, às curvas de densidade das notas das amostras.

Em CN, as amostras aleatórias tiveram um pico próximo de -1,00, as retangulares tiveram distribuição mais uniforme, com um vale próximo de 0,00 e as deslocadas tiveram dois picos (próximo de -0,50 e 1,75). As amostras retangulares tiveram uma proporção maior de sujeitos com notas mais altas e sua distribuição abrangeu mais a região com mais itens. As aleatórias foram as que menos abrangeram essa região.

Figura 5

Curva de densidade das notas das 100 amostras sorteadas em Ciências da Natureza para cada desenho amostral e curva de informação do teste



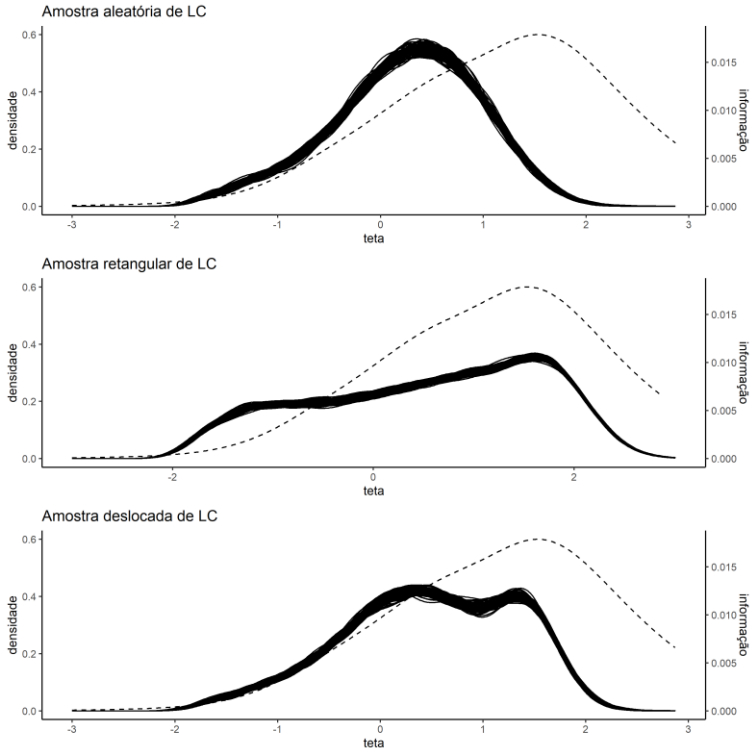
Nota. a linha pontilhada se refere à informação psicométrica da prova e as linhas contínuas, às curvas de densidade das notas das amostras.

Em LC, as amostras aleatórias tiveram uma forma de sino com pico próximo de 0,50, e as retangulares e as deslocadas tiveram distribuição mais uniforme. A diferença entre as duas últimas é que as retangulares tiveram mais sujeitos entre -1,50 e 2,00, com um pico próximo de

1,75, e as deslocadas tiveram mais sujeitos entre -0,50 e 1,75. As amostras retangulares conseguiram abranger uma amplitude maior da região do teste com mais informação, porém as deslocadas tiveram maior pico nessa região. As amostras aleatórias foram as que contemplaram menor amplitude da região da prova com mais informação.

Figura 6

Curva de densidade das notas das 100 amostras sorteadas em Linguagens e Códigos para cada desenho amostral e curva de informação do teste

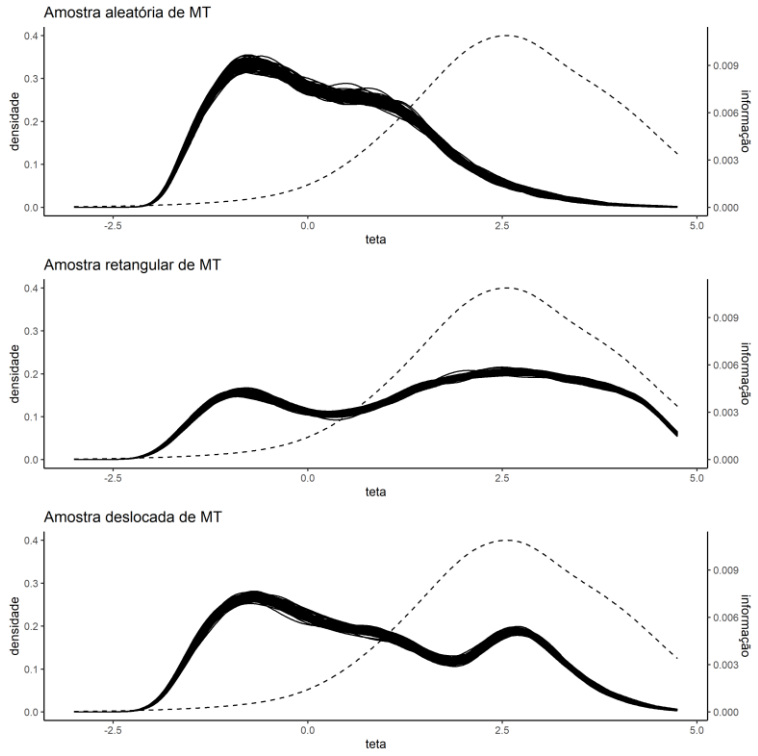


Nota. a linha pontilhada se refere à informação psicométrica da prova e as linhas contínuas, às curvas de densidade das notas das amostras.

Em MT, as amostras aleatórias tiveram um pico próximo de -1,00, as retangulares tiveram distribuição mais uniforme, com um vale próximo de 0,50, e as deslocadas tiveram dois picos (próximo de -1,00 e 3,00). As amostras retangulares abrangeram uma amplitude maior da região do teste com maior informação, e com maior proporção em relação às demais amostras. As amostras aleatórias foram as que menos abrangeram.

Figura 7

Curva de densidade das notas das 100 amostras sorteadas em Matemática para cada desenho amostral e curva de informação do teste



Nota. a linha pontilhada se refere à informação psicométrica da prova e as linhas contínuas, às curvas de densidade das notas das amostras.

Parâmetros estimados

A Figura 8 mostra os valores médios dos parâmetros estimados em CH obtidos nas 100 replicações dos três tipos de amostra, em função dos parâmetros reais. A Figura 9 mostra os parâmetros de CN, a Figura 10, os de LC e a Figura 11, os de MT. As figuras contêm a equação de regressão de cada condição, a reta esperada se os valores dos parâmetros estimados fossem exatamente iguais aos reais e o valor da correlação entre os valores médios estimados e os valores reais.

Em CH, a amostra deslocada retornou valores melhores de discriminação e de pseudochute, pois o intercepto da equação foi o mais próximo de zero e o coeficiente, o mais próximo de um. Já para a dificuldade, a amostra retangular retornou valores melhores. Em CN, a deslocada retornou valores melhores de dificuldade e de pseudochute. Já para a discriminação,

pelo gráfico não se pode concluir o melhor tipo de amostra, pois o coeficiente da retangular é mais próximo de um e o intercepto da deslocada, mais próximo de zero. Em LC, a aleatória retornou melhores valores de dificuldade e pseudochute. Para a discriminação, a aleatória apresentou melhor intercepto, a retangular melhor coeficiente e a deslocada, os segundos melhores valores dos parâmetros da regressão. Em MT, a retangular retornou melhores valores de discriminação e dificuldade, e a aleatória, de pseudochute.

Figura 8

Parâmetros estimados em função dos parâmetros reais em Ciências Humanas

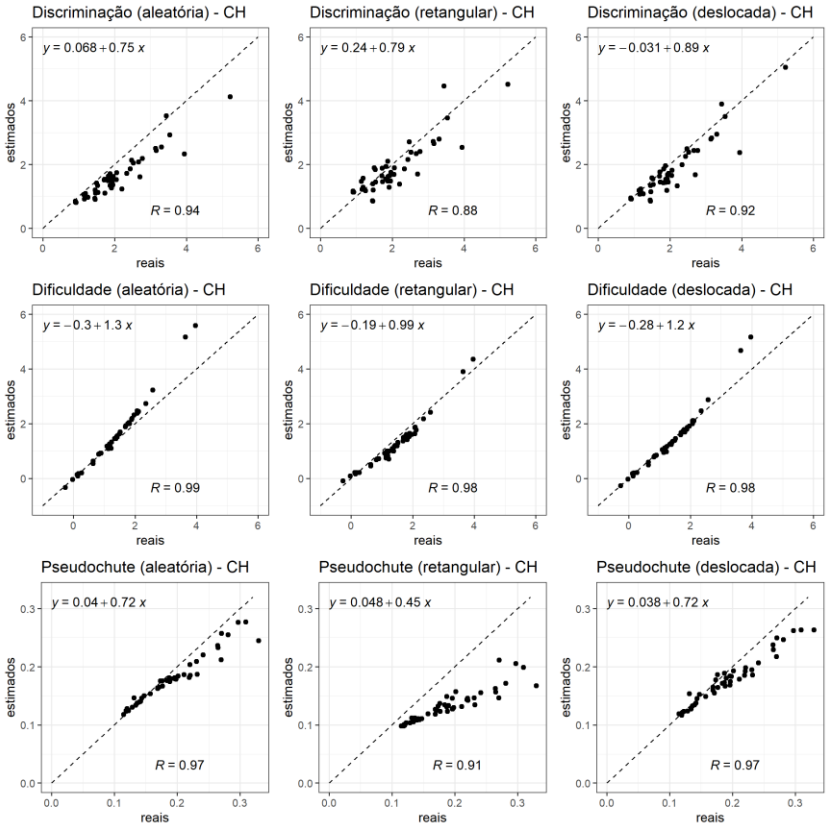


Figura 9

Parâmetros estimados em função dos parâmetros reais em Ciências da Natureza

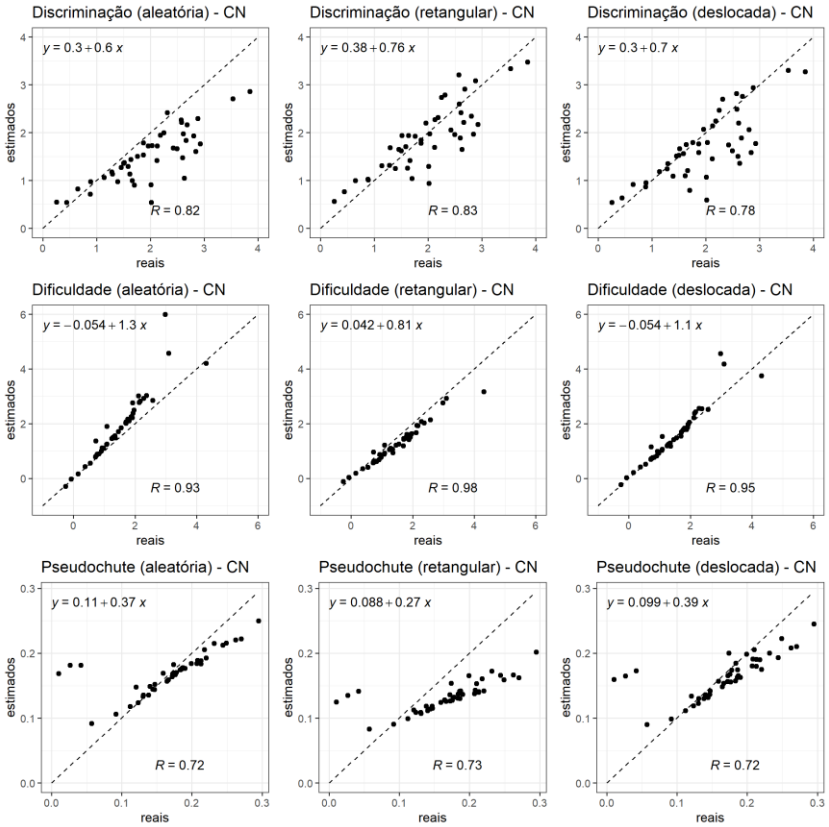


Figura 10

Parâmetros estimados em função dos parâmetros reais em Linguagens e Códigos

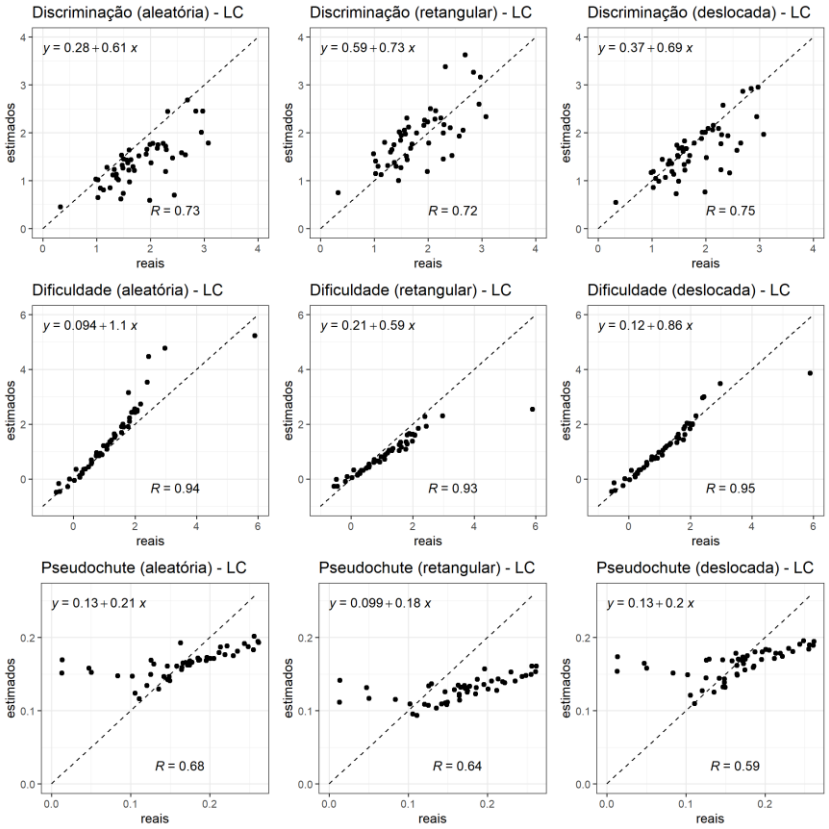
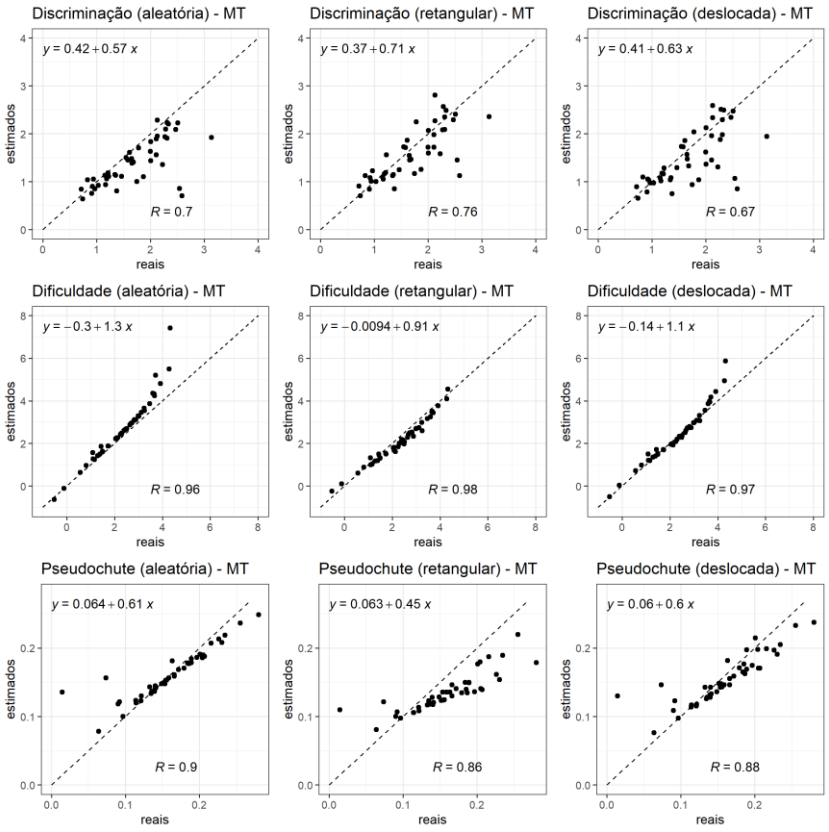


Figura 11

Parâmetros estimados em função dos parâmetros reais em Matemática



Em todas as áreas, em geral os parâmetros de discriminação foram subestimados, pois muitos ficaram abaixo da reta de igualdade dos valores. Já para os parâmetros de dificuldade, nota-se que os itens mais difíceis ficaram mais deslocados da reta. Os parâmetros de pseudochute em geral também foram subestimados, e em LC os valores estimados tiveram pouca relação com os valores reais.

No que diz respeito aos valores de correlação, em geral a variação entre as amostras foi pequena. Em CH, os valores de correlação foram ligeiramente superiores para as amostras aleatória e deslocada. A diferença para a amostra retangular chegou a 0,06 (discriminação, aleatória; e pseudochute, aleatória e deslocada). Em CN e LC, os valores de correlação foram mais parecidos entre si. Em CN, a maior diferença foi entre a retangular e a deslocada, para a discriminação (0,05). Em LC, a maior diferença foi entre a aleatória e a deslocada, para o pseudochute (0,09). A maior diferença foi observada em MT, em que a amostra retangular

apresentou valores de correlação ligeiramente superiores na discriminação, com diferença de 0,09 para a deslocada. Para os demais parâmetros, os valores foram mais parecidos, com uma diferença de 0,04 entre a aleatória e a retangular no pseudochute. Em suma, os valores de correlação foram semelhantes, porém nota-se uma pequena superioridade para a amostra aleatória no parâmetro do pseudochute, e um equilíbrio entre as amostras para a discriminação e dificuldade.

Viés, MDA e REQM

A Tabela 10 mostra os valores médios do viés, da MDA e da REQM dos 45 parâmetros para cada tipo de amostra em cada área. Para a discriminação, o viés médio variou de -0,482 (CN, aleatória) a 0,102 (LC retangular); a MDA média variou de 0,294 (MT, retangular) a 0,521 (CN, aleatória); a REQM média variou de 0,305 (MT, retangular) a 0,545 (CN, aleatória). Para a dificuldade, o viés médio variou de -0,268 (LC, retangular) a 0,378 (CN, aleatória); a MDA variou de 0,133 (CH, deslocada) a 0,395 (CN, aleatória); a REQM variou de 0,146 (CH, deslocada) a 0,429 (CN, aleatória). Para o pseudochute, o viés variou de -0,059 (CH, retangular) a 0,002 (LC e MT, aleatória); a MDA variou de 0,016 (MT, aleatória) a 0,059 (CH, retangular); a REQM variou de 0,017 (MT, aleatória) a 0,059 (CH, retangular).

Tabela 10

Média do viés, da média da diferença absoluta, e da raiz do erro quadrático médio dos 45 parâmetros estimados

Área e amostra	a			b			c		
	Viés	MDA	REQM	Viés	MDA	REQM	Viés	MDA	REQM
CH									
Aleatória	-0,447	0,454	0,474	0,191	0,220	0,242	-0,014	0,017	0,019
Deslocada	-0,256	0,306	0,326	-0,006	0,133	0,146	-0,016	0,020	0,022
Retangular	-0,189	0,368	0,379	-0,198	0,247	0,251	-0,059	0,059	0,059
CN									
Aleatória	-0,482	0,521	0,545	0,378	0,395	0,429	0,001	0,026	0,028
Deslocada	-0,290	0,412	0,430	0,070	0,164	0,179	-0,005	0,029	0,030
Retangular	-0,094	0,353	0,364	-0,235	0,263	0,266	-0,036	0,051	0,052
LC									
Aleatória	-0,431	0,457	0,472	0,263	0,326	0,354	-0,002	0,035	0,036

Deslocada	-0,201	0,317	0,334	-0,042	0,168	0,180	-0,002	0,036	0,037
Retangular	0,102	0,381	0,391	-0,268	0,328	0,330	-0,037	0,054	0,055
MT									
Aleatória	-0,311	0,350	0,372	0,365	0,371	0,411	0,002	0,016	0,017
Deslocada	-0,216	0,348	0,363	0,074	0,185	0,198	-0,005	0,020	0,021
Retangular	-0,122	0,294	0,305	-0,221	0,272	0,276	-0,025	0,033	0,034

Nota. a = parâmetro de discriminação; b = parâmetro de dificuldade; c = parâmetro de pseudochute; REQM = raiz do erro quadrático médio; CH = Ciências Humanas; CN = Ciências da Natureza; LC = Linguagens e Códigos; MT = Matemática

De uma maneira geral, a amostra retangular apresentou valores absolutos de viés menores para a discriminação, a deslocada, menores para a dificuldade e a aleatória, menores para o pseudochute. No que diz respeito à MDA e à REQM, a amostra deslocada apresentou valores melhores para a discriminação em CH e em LC e para a dificuldade em todas as áreas. Já a amostra retangular apresentou valores de MDA e REQM melhores para discriminação em CN e em MT. A amostra aleatória retornou valores de pseudochute melhores do que as demais amostras nas quatro provas.

MDA condicional

A Figura 12 mostra a MDA dos três parâmetros condicionada à dificuldade em CH. Cada gráfico possui uma linha contínua com pontos marcados e uma linha tracejada. Cada ponto da linha contínua representa a média da MDA dos itens de acordo com o decil de dificuldade. Ou seja, cada ponto representa um agrupamento de quatro ou cinco itens (10% do total de itens da prova). A linha tracejada corresponde à distribuição da densidade da primeira amostra sorteada de cada desenho amostral (primeira replicação). A Figura 13 mostra essas informações em CN, a Figura 14, em LC e a Figura 15, em MT. De forma geral, os valores de MDA dos três parâmetros foram maiores para os itens mais difíceis, em especial na região com menos sujeitos da amostra.

Em relação ao parâmetro de discriminação, em geral nas quatro áreas a amostra deslocada retornou menor MDA para itens com dificuldade até cerca de 1,00. Nessa região, a amostra retangular apresentou os valores mais altos de MDA. Para itens acima de 1,00, os

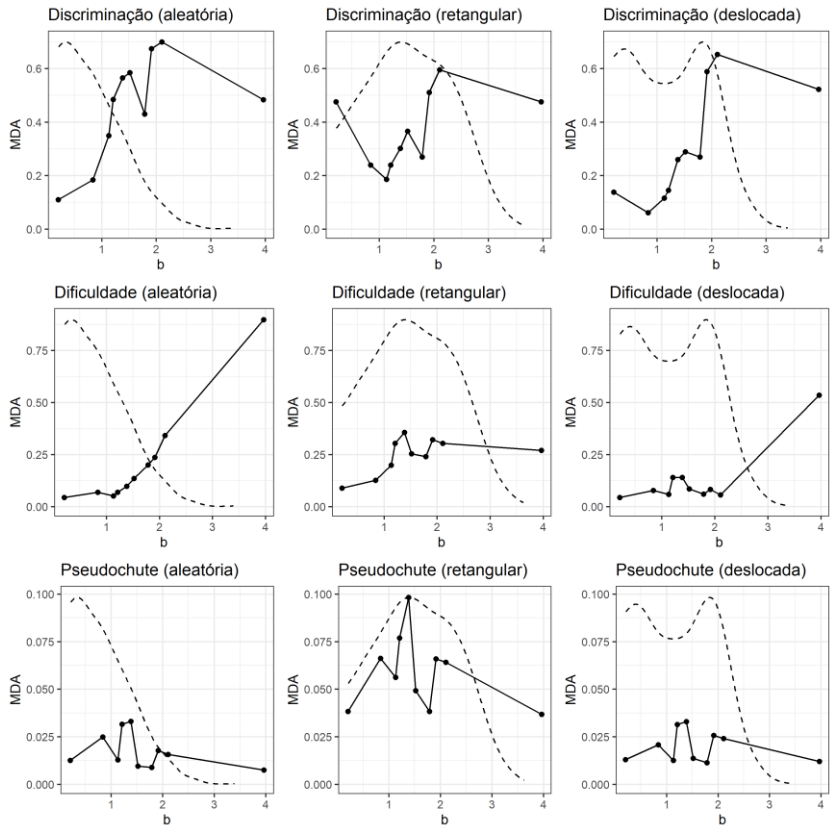
valores de MDA foram mais semelhantes entre as amostras, com a amostra retangular retornando valores ligeiramente menores para alguns itens, em especial os mais difíceis.

No que diz respeito à dificuldade, nas quatro áreas os valores de MDA dos itens extremamente fáceis foram semelhantes para os três tipos de amostra. Para os itens extremamente difíceis, a amostra retangular retornou valores menores do que as demais, com exceção de LC. E para os demais itens, a amostra deslocada retornou valores menores em comparação às outras amostras.

Em relação ao pseudochute, as amostras aleatória e deslocada tiveram valores consideravelmente menores de MDA do que a amostra retangular. Para os itens difíceis, a amostra aleatória apresentou ligeira vantagem em relação à deslocada.

Figura 12

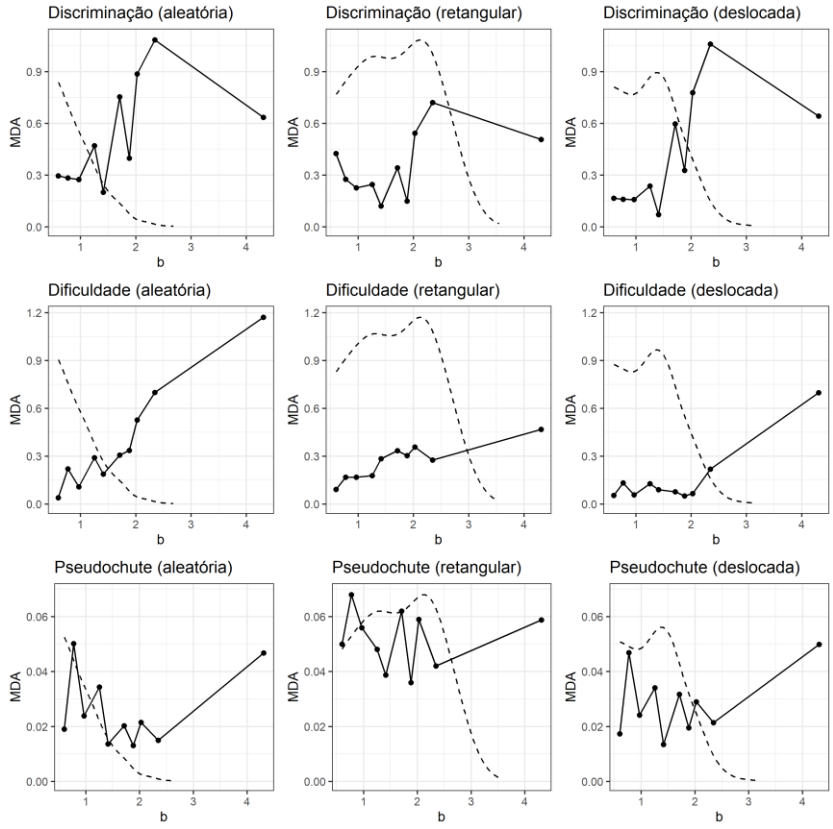
Média da diferença absoluta dos parâmetros condicionada à dificuldade em Ciências Humanas



Nota. A linha contínua com pontos marcados representa a média da diferença absoluta dos parâmetros e a linha tracejada, a densidade da distribuição da primeira amostra sorteada de cada desenho amostral.

Figura 13

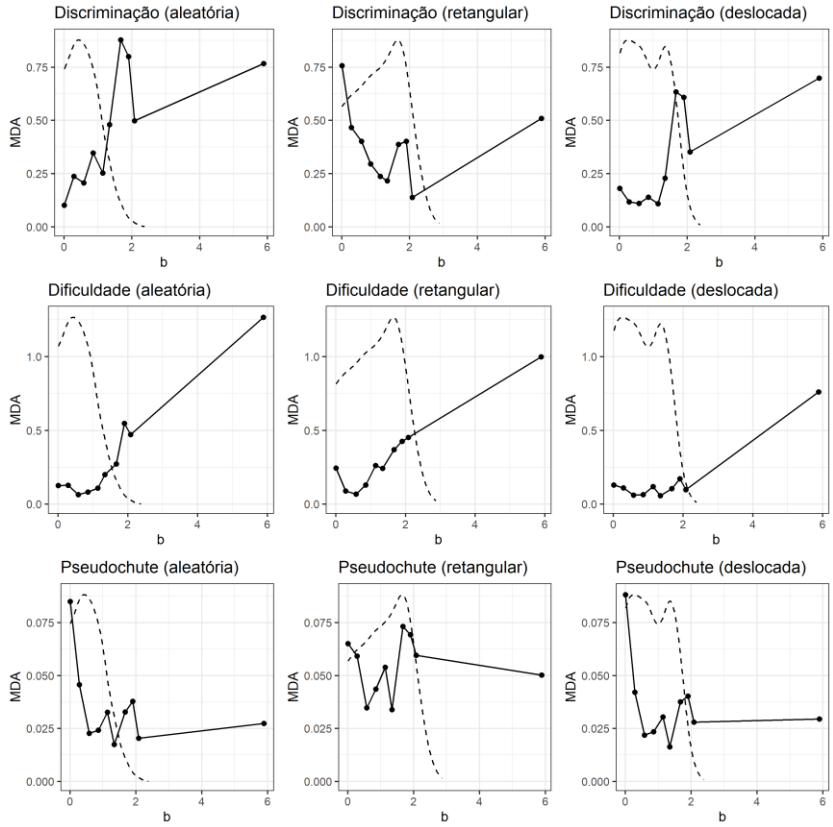
Média da diferença absoluta dos parâmetros condicionada à dificuldade em Ciências da Natureza



Nota. A linha contínua com pontos marcados representa a média da diferença absoluta dos parâmetros e a linha tracejada, a densidade da distribuição da primeira amostra sorteada de cada desenho amostral.

Figura 14

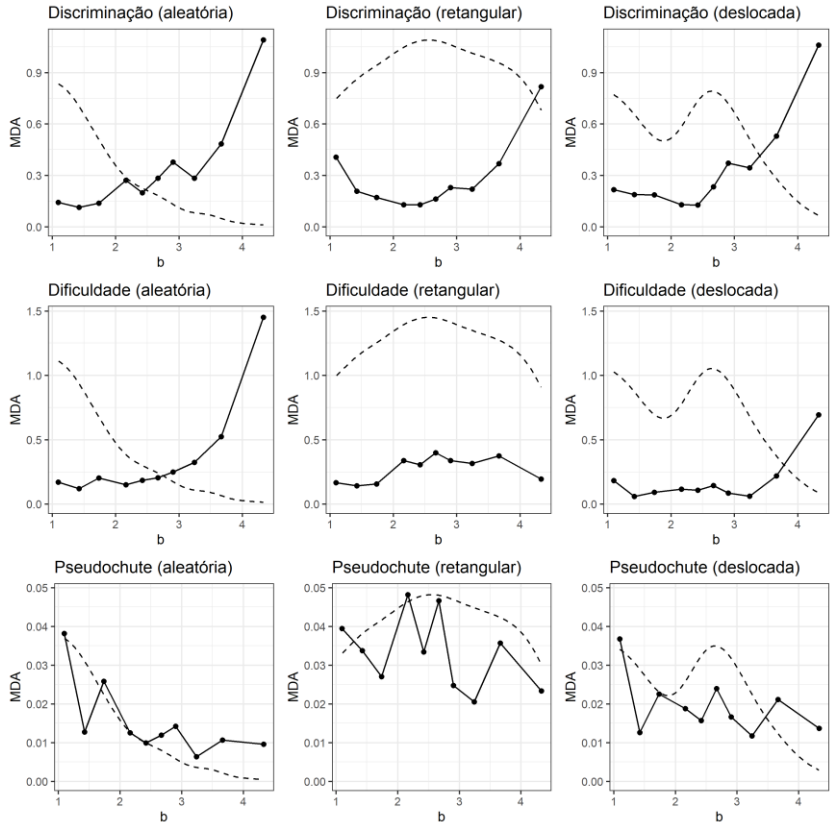
Média da diferença absoluta dos parâmetros condicionada à dificuldade em Linguagens e Códigos



Nota. A linha contínua com pontos marcados representa a média da diferença absoluta dos parâmetros e a linha tracejada, a densidade da distribuição da primeira amostra sorteada de cada desenho amostral.

Figura 15

Média da diferença absoluta dos parâmetros condicionada à dificuldade em Matemática



Nota. A linha contínua com pontos marcados representa a média da diferença absoluta dos parâmetros e a linha tracejada, a densidade da distribuição da primeira amostra sorteada de cada desenho amostral.

ANOVA

A Tabela 11 resume os resultados da ANOVA. Em nenhum dos modelos para o parâmetro de discriminação a diferença entre os desenhos amostrais foi estatisticamente significativa, o que indica que não há evidências de diferença de MDA entre os tipos de amostra. Já para a dificuldade e o pseudochute, em todas as áreas há evidência de diferença entre os tipos amostrais.

Nas comparações entre os grupos, para a dificuldade somente em duas comparações há significância estatística e em uma delas o intervalo de confiança do tamanho do efeito não incluiu o zero, com efeito médio (CN, média da deslocada menor que a aleatória). Para o pseudochute, em todas as áreas as médias das amostras aleatória e deslocada foram menores que as da retangular, com significância estatística e tamanho de efeito médio ou grande. Não foi observada diferença entre as amostras aleatória e deslocada para o pseudochute.

Tabela 11

Resultados da ANOVA com a média da diferença absoluta como variável dependente

Área	Parâmetro	Amostra	M	DP	F(gl)	p	Comparação	$p_{(bonf)}$	d
		Aleatória	0,45	0,32			Retangular	-	-
	a	Retangular	0,37	0,28	2,6(gl)	0,079	Deslocada	-	-
		Deslocada	0,31	0,30			Aleatória	-	-
		Aleatória	0,22	0,32			Retangular	1,000	-0,111
CH	b	Retangular	0,25	0,11	4,7(gl)	0,012	Deslocada	0,069	0,648
		Deslocada	0,13	0,22			Aleatória	0,243	-0,315
		Aleatória	0,02	0,02			Retangular	< 0,001	-1,616*
	c	Retangular	0,06	0,03	30,7(gl)	< 0,001	Deslocada	< 0,001	1,542*
		Deslocada	0,02	0,01			Aleatória	1,000	0,154
		Aleatória	0,52	0,40			Retangular	-	-
	a	Retangular	0,35	0,26	2,7(gl)	0,072	Deslocada	-	-
		Deslocada	0,41	0,40			Aleatória	-	-
		Aleatória	0,40	0,50			Retangular	0,228	0,355
CN	b	Retangular	0,26	0,18	3,9(gl)	0,024	Deslocada	0,550	0,414
		Deslocada	0,16	0,29			Aleatória	0,007	-0,569*
		Aleatória	0,03	0,04			Retangular	0,001	-0,785*
	c	Retangular	0,05	0,03	8,9(gl)	< 0,001	Deslocada	0,005	0,729*
		Deslocada	0,03	0,03			Aleatória	1,000	0,083
		Aleatória	0,46	0,39			Retangular	-	-
	a	Retangular	0,38	0,25	1,9(gl)	0,158	Deslocada	-	-
		Deslocada	0,32	0,33			Aleatória	-	-
		Aleatória	0,33	0,43			Retangular	1,000	-0,003
LC	b	Retangular	0,33	0,47	3,4(gl)	0,039	Deslocada	0,147	0,408
		Deslocada	0,17	0,30			Aleatória	0,154	-0,433
		Aleatória	0,03	0,03			Retangular	0,011	-0,605*
	c	Retangular	0,05	0,03	5,9(gl)	0,004	Deslocada	0,019	0,560*
		Deslocada	0,04	0,04			Aleatória	1,000	0,034
		Aleatória	0,35	0,40			Retangular	-	-
	a	Retangular	0,29	0,29	0,4(gl)	0,660	Deslocada	-	-
		Deslocada	0,35	0,38			Aleatória	-	-
		Aleatória	0,37	0,51			Retangular	0,503	0,268
MT	b	Retangular	0,27	0,13	3,2(gl)	0,047	Deslocada	0,682	0,435
		Deslocada	0,18	0,25			Aleatória	0,031	-0,465
		Aleatória	0,02	0,02			Retangular	< 0,001	-0,821*
	c	Retangular	0,03	0,02	7,9(gl)	< 0,001	Deslocada	0,008	0,649*
		Deslocada	0,02	0,02			Aleatória	1,000	0,200

Nota. CH = Ciências Humanas; CN = Ciências da Natureza; LC = Linguagens e Códigos; MT = Matemática; a = discriminação; b = dificuldade; c = pseudochute; M = média das 45 médias da diferença absoluta dos itens; DP = desvio padrão das 45 médias da diferença absoluta dos itens; F = estatística de Fisher; gl = graus de liberdade; p(bonf) = significância estatística com correção de Bonferroni; d = tamanho do efeito de Cohen; * = intervalo de confiança não inclui o zero.

Discussão

Este estudo teve como objetivo comparar diferentes distribuições amostrais na calibração de itens no modelo 3PL da TRI. Utilizamos um banco de respostas simuladas a partir das notas de sujeitos e dos parâmetros dos itens do Enem 2020. O desenho da simulação consistiu na manipulação da distribuição amostral (três tipos de amostra: aleatória, retangular e deslocada) para calibração dos itens das quatro provas do exame, com 100 replicações de cada condição. Os resultados não corroboram a hipótese H1 (a amostra com distribuição retangular de acertos retorna parâmetros mais semelhantes aos parâmetros reais). Para o parâmetro de discriminação, não houve diferença significativa entre os tipos de amostra. Para o parâmetro de dificuldade, a amostra deslocada retornou melhores valores em CH quando comparada com a retangular e em CN quando comparada com a aleatória. Nas demais áreas não houve diferença relevante. Para o parâmetro de pseudochute, as amostras deslocada e aleatória se mostraram melhores, sem diferença relevante entre elas. Apesar das diferenças encontradas, os resultados não apontam para a prevalência de um tipo de amostra para calibrar os itens do Enem 2020. A hipótese H2 do estudo (a amostra estratificada pelo número de acertos – deslocada – retorna parâmetros aceitáveis) foi corroborada.

Neste trabalho, o parâmetro de discriminação foi mais subestimado com a amostra aleatória do que com a retangular, pois ambos os vieses foram negativos, porém o valor absoluto da primeira foi maior. A dificuldade foi superestimada pela retangular e mais fortemente pela aleatória, ao passo que a deslocada teve viés próximo de zero. Para o pseudochute, o viés da retangular foi maior do que as demais amostras. Esses achados contrastam com os da simulação de Han (2012), em que a amostra retangular subestimou mais a discriminação do que a amostra normal, não se observou diferença entre as amostras para a dificuldade e a retangular apresentou um viés menor para o pseudochute. Cabe destacar que não utilizamos uma amostra com

distribuição normal neste estudo, porém as amostras aleatórias foram as que mais se aproximaram de um formato de sino.

Nossos resultados também refutam os achados de Wingersky e Lord (1984), que observaram uma precisão maior da calibração (medida pelo erro padrão do parâmetro) com amostra retangular em relação a uma distribuição em forma de sino, para a discriminação e o pseudochute. Neste estudo, a precisão (medida pela MDA) da discriminação foi semelhante entre os três tipos de amostra e a do pseudochute foi menor para a amostra retangular.

As divergências entre os achados deste estudo e os anteriores pode ser decorrente de dois fatores. O primeiro diz respeito à justaposição entre as amostras e os itens. No estudo de Han (2012), por exemplo, as amostras foram sorteadas de uma população com média zero e a média da dificuldade dos itens foi -0,068. Isso significa que em média os itens estavam posicionados próximos dos tetas dos sujeitos. No nosso estudo, as amostras em geral estavam deslocadas para a esquerda e não contemplavam a parte mais difícil da curva de informação dos testes, em especial as amostras aleatórias. O teta ideal para se estimar a dificuldade ou a discriminação do item no modelo 3PL não é igual à sua dificuldade. Para se estimar a dificuldade, o teta ideal é um pouco superior e para se estimar a discriminação é um pouco superior ou inferior ao valor da dificuldade do item (Stocking, 1990). Por isso, diferenças na distribuição da proficiência podem afetar a precisão da calibração de um item. A maior MDA da dificuldade dos itens mais à direita na escala reforça esse ponto.

O segundo fator que pode ter ocasionado as diferenças dos resultados é a diferença entre as amostras retangulares deste estudo e as dos anteriores. Uma vez que neste estudo a estratificação dos sujeitos se deu a partir da soma de acertos, a distribuição de teta na amostra retangular não foi de fato uniforme, pois a mesma quantidade de acertos pode resultar em tetas diferentes no modelo 3PL da TRI. Nos estudos anteriores, as amostras retangulares tiveram distribuição mais uniforme pois a estratificação foi pelo teta.

Isso mostra uma limitação deste estudo. Não controlamos experimentalmente a distribuição do banco de itens e da população, uma vez que utilizamos dados reais de uma aplicação do Enem. Isso limita as generalizações para situações em que a justaposição entre a população e os itens seja semelhante à justaposição observada nas quatro situações analisadas neste estudo. Por isso, sugerimos que novos estudos verifiquem o impacto dos graus de justaposição entre a população e a curva de informação do teste sobre a calibração dos itens.

Outra limitação do estudo é que utilizamos somente respostas simuladas, o que pode enviesar os resultados, pois possíveis interferências externas (como fadiga e motivação) não foram consideradas. Recomendamos estudos que verifiquem a estabilidade da calibração de itens a partir de respostas reais.

Outra ação que também recomendamos é utilizar os tetas dos sujeitos para sortear novamente a amostra de calibração. Neste estudo, a amostra retangular não tinha de fato distribuição retangular, porque a estratificação da população foi feita a partir da soma dos acertos. Para se obter uma amostra com distribuição mais próxima de retangular, pode-se estimar os tetas dos sujeitos a partir dos parâmetros calibrados provisoriamente e sortear nova amostra estratificada pelo teta provisório. Em seguida, os itens podem ser recalibrados com a nova amostra retangular e os resultados podem ser comparados com a calibração anterior.

Outra implementação pode ser utilizar os parâmetros da primeira calibração como valores prévios (*prior*) na nova calibração. Kim (2006) indicou que atualizar os valores prévios dos parâmetros dos itens melhora a calibração para amostras com distribuição diferente da normal padrão, por exemplo, normal com média um e desvio padrão 1,96.

Como conclusão, pode-se afirmar que este estudo possui evidências de que não há diferenças relevantes nos três desenhos amostrais avaliados para calibrar itens nas condições reais do Enem 2020. Ainda, o método adotado no exame performa de maneira semelhante aos

demais analisados. Esperamos que este estudo contribua para as decisões de amostragem para calibrar itens em testes educacionais de larga escala.

Referências

Akour, M., & AL-Omari, H. (2013). Empirical investigation of the stability of IRT item-parameters estimation. *International Online Journal of Educational Sciences*, *5*(2), 291–301.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. http://doi.org/10.18637/jss.v048.i06

Cheng, Y., & Yuan, K.-H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika*, *75*(2), 280–291. 10.1007/s11336-009-9144-x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model, *17*(1). University of Massachusetts Amherst. 10.7275/F0GZ-KC87

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2021). *Exame Nacional do Ensino Médio—Enem Procedimentos de análise*. Inep. Retirado de https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/enem_procedimentos_de_analise.pdf

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355–381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

Ministério da Educação. (2009). Retirado de http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=768-proposta-novovestibular1-pdf&category_slug=documentos-pdf&Itemid=30192

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Psicología Pirámide

Nunes, C. H. S. da S., & Primi, R. (2005). Impacto do tamanho da amostra na calibração de itens e estimativa de escores por teoria de resposta ao item. *Avaliação Psicológica*, *4*(2), 141–153.

Paek, I., Liang, X., & Lin, Z. (2021). Regarding item parameter invariance for the Rasch and the 2-parameter logistic models: An investigation under finite non-representative sample calibrations. *Measurement: Interdisciplinary Research and Perspectives*, *19*(1), 39–54. 10.1080/15366367.2020.1754703

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retirado de https://www.R-project.org/

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*(1), 321–335. http://doi.org/10.12738/estp.2017.1.0270

Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory & Practice*. 10.12738/estp.2015.6.0102

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika*, *55*(3), 461–475.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*(3), 347–364. 10.1177/014662168400800312

Produto 1 – Pacote simCAT

Para desenvolver uma CAT, são necessárias simulações que subsidiem as decisões para determinar suas especificações. Tais especificações incluem o algoritmo da CAT (início da aplicação, seleção dos itens, estimação do teta e critério de parada) e a composição do banco de itens que será utilizado para a população e para cada sujeito. Existem diversos programas e pacotes comerciais e gratuitos que permitem simular aplicações em CAT, como catIrt (Nydick, 2014), catR (Magis & Raîche, 2012), CATSim (Weiss & Guyer, 2012), Firestar (Choi, 2020) e mirtCAT (Chalmers, 2016).

De forma geral, os pacotes permitem gerar bancos de respostas a partir de simulações de Monte Carlo e especificar diferentes critérios para início da aplicação, métodos de seleção de itens, métodos de estimação do teta e critérios de parada. Apesar das similaridades, cada pacote possui suas especificidades. Por exemplo, o catR permite utilizar o método progressivo com diferentes valores de parâmetro de aceleração e o mirtCAT permite publicar a CAT final.

Apesar da vasta lista de funcionalidades desses pacotes, esta tese demandou algumas não contempladas por eles. Por isso, decidimos nos basear no pacote catR, que tinha a maior similaridade com aquilo que nossas especificações demandavam, para construir um pacote de simulação em CAT em ambiente R.

O objetivo do pacote simCAT (Jaloto & Primi, 2023a) é realizar simulações em CAT para itens dicotômicos. Ele permite gerar bancos de respostas com simulações de Monte Carlo, especificar diferentes métodos de seleção de itens (Máxima Informação de Fisher e aleatório), critérios de parada (tamanho fixo, erro padrão, mínima informação do banco, diferença absoluta do teta e redução do erro) e métodos de balanceamento de conteúdo (*constrained CAT*, *modified constrained CAT* e *modified multinomial model*). Além disso, o pacote permite utilizar o método progressivo restrito para controle de exposição, com as diferentes formas de calcular o seu peso, sejam para CATs de tamanho fixo, seja para CATs de tamanho variável.

Adicionalmente, permite controlar o parâmetro de aceleração desse peso. O pacote avalia a precisão CAT por meio da comparação entre o teta real e o teta estimado (correlação, raiz do erro quadrático médio e viés) e do erro padrão médio do teta estimado. A avaliação da eficiência é feita por meio da média dos valores mínimo, máximo, média e mediana do tamanho das aplicações. O pacote avalia a segurança por meio da taxa de exposição dos itens e da sobreposição. O avanço deste pacote está na reunião de todas essas funcionalidades.

O pacote é gratuito, de uso livre e de código aberto. Ele está disponível em http://github.com/alexandrejaloto/simCAT, juntamente com seu manual completo em PDF, cuja capa é apresentada na Figura 16. Para citar o pacote, sugerimos:

Jaloto, A., & Primi, R. (2023). *SimCAT. Computerized adaptive testing simulations*. https://github.com/alexandrejaloto/simCAT

Figura 16

Capa do manual do pacote simCAT

Package 'simCAT'

November 24, 2022

Title Implements Computerized Adaptive Testing Simulations

Version 0.0.0.9000

Author Alexandre Jaloto [aut, cre] (<https://orcid.org/0000-0002-5291-1768>)
Ricardo Primi [ths]

Description What the package does (one paragraph).

License use_mit_license() , use_gpl3_license() or friends to pick a license

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

Imports dplyr,
mirt,
mirtCAT

URL https://github.com/alexandrejaloto/simCAT

BugReports https://github.com/alexandrejaloto/simCAT/issues

R **topics documented:**

calc.info 2
calc.prob 2
cat.evaluation 3
content.balancing 4
eap 5
exposure.rate 5
gen.resp 6
rmse 6
select.item 7
simCAT 9
stop.cat 11

Index 13

1

Para instalar o pacote, sugere-se o uso do seguinte comando em R:

```
devtools::install_github('alexandrejaloto/simCAT')
```

Esperamos que o pacote passe por atualizações constantes, portanto encorajamos o seu uso, sua divulgação e as sugestões de aprimoramento.

Artigo 2 – Método progressivo restrito em CAT educacional de alto impacto: eficiência, precisão e segurança

Alexandre Jaloto

Ricardo Primi

Resumo

O método progressivo restrito (PR) aumenta a segurança de uma Testagem Adaptativa Computadorizada (CAT). Entretanto, pouco se sabe sobre o efeito do parâmetro de aceleração sobre a precisão e segurança. Por isso, avaliamos o PR com diferentes parâmetros de aceleração em CATs de tamanho fixo e variável. Combinamos métodos de seleção de itens (Máxima Informação de Fisher – MIF – e PR) com critérios de parada (Tamanho fixo, Erro padrão de 0,30 e Redução do erro de 0,015) e simulamos CATs para cada prova do Enem. Com o critério de redução do erro, a precisão com MIF foi menor. Nas demais CATs, a precisão foi semelhante. A segurança aumentou com o parâmetro de aceleração maior. Posteriormente, comparamos o formato linear do Enem com a CAT de 20 itens. A última teve precisão maior.
Palavras-chave: psicometria; avaliação educacional em larga escala; testagem adaptativa informatizada; simulação; controle de exposição de itens

Abstract

The Progressive-Restrict method (PR) increases the safety of Computerized Adaptive Testing (CAT). However, little is known about the effect of the acceleration parameter on accuracy and safety. Therefore, we evaluated the PR with different acceleration parameters in CATs with fixed and variable-length. We combined item selection methods (Maximum Fisher Information – MFI – and PR) with stopping criteria (Fixed length, Standard error of 0,30 and Error reduction of 0,015) and simulated CATs for each Enem test. With the error reduction criterion, the accuracy with MFI was lower. In the other CATs, the precision was similar. Security has increased with larger acceleration parameters. At last, we compared the linear format of the Enem with the 20-item CAT. The latter had greater accuracy.
Keywords: psychometrics; large-scale educational assessment; computerized adaptive testing; simulation; item exposure control.

Resumen

El Método Progresivo Restringido (PR) aumenta la seguridad de las Pruebas Adaptativas Computarizadas (CAT). Sin embargo, se sabe poco sobre el efecto del parámetro de aceleración en la precisión y la seguridad. Por lo tanto, evaluamos el PR con diferentes parámetros de aceleración en CAT de tamaño fijo y variable. Combinamos métodos de selección de ítems (Máxima Información de Fisher - MIF - y PR) con criterios de parada (Tamaño fijo, Error estándar de 0,30 y Reducción de error de 0,015) y simulamos CATs para cada prueba de Enem. Con el criterio de reducción de error, la precisión con MIF fue menor. En los demás CAT, la precisión fue similar. La seguridad ha aumentado con el parámetro de aceleración más grande. Posteriormente, comparamos el formato lineal del Enem con el CAT de 20 ítems. Este último tenía mayor precisión.

Palabras clave: psicometría; evaluación educativa a gran escala; pruebas adaptativas informatizadas; simulación; control de exposición de ítems

A Testagem Adaptativa Computadorizada (em inglês *Computerized Adaptive Testing*, CAT) pode aumentar a eficiência de um teste ao reduzir o seu tamanho (Bulut & Kan, 2012; Mizumoto et al., 2019; Spenassato et al., 2016). Porém, em testes de alto impacto (*high stakes*) a segurança do teste (em termos de exposição dos itens) também deve ser considerada, além da eficiência. Por meio de um componente aleatório, o método progressivo restrito (PR) controla a exposição dos itens com pouco impacto na eficiência e precisão (Leroux & Dodd, 2016; Leroux et al., 2013). A importância desse componente aleatório na seleção do item reduz ao longo da aplicação, e a velocidade dessa redução tem relação com a administração de itens pouco expostos (Barrada et al., 2008). Essa velocidade é determinada pelo parâmetro de aceleração na equação do método PR, o qual tem sido pouco explorado, em especial em CAT de tamanho variável. Por isso, neste trabalho avaliamos o uso do controle de exposição PR com diferentes parâmetros de aceleração em CATs de tamanho fixo e variável. Utilizamos dados do Exame Nacional do Ensino Médio (Enem), um teste educacional de alto impacto.

A CAT pode otimizar a aplicação de um teste porque os itens são administrados ao participante de acordo com sua resposta ao item anterior (Weiss & Kingsbury, 1984). Quando o método de Máxima Informação de Fisher (MIF) é utilizado, o item selecionado é aquele com maior informação para o teta (variável latente) provisório do sujeito. Por isso, a CAT aumenta a eficiência da aplicação (em termos do tamanho do teste) sem comprometer a precisão da medida, ou até a melhora. Por exemplo, a aplicação de Sulak e Kelecioğlu (2019) teve um tamanho médio de 7,07 itens com um banco de 250 itens. Spenassato et al. (2016) simularam a aplicação do Enem 2012 em formato CAT e reduziram o tamanho do teste de 45 para 33 itens, com correlação de 0,998 entre os tetas originais e simulados. Em testes educacionais de alto impacto como o Enem (que é utilizado para ingressar no ensino superior no Brasil), além da

eficiência e precisão, a segurança é um fator que precisa ser considerado. Uma forma de aumentar a segurança do teste é controlar a exposição dos itens.

O método de controle de exposição PR evita a superexposição dos itens e diminui a subexposição e a sobreposição. A sobreposição corresponde à proporção de itens iguais aplicados a dois participantes selecionados aleatoriamente. Na simulação de Leroux e Dodd (2016), as condições sem controle de exposição tiveram a raiz do erro quadrático médio (REQM) chegando a 0,31. Ao inserir o método PR, esse valor não passou de 0,34. Além da pouca piora na precisão, esse método proporcionou maior segurança para o teste, pois a menor taxa de sobreposição obtida sem controle de exposição foi cerca de 0,42, ao passo que com o PR a maior taxa foi cerca de 0,21. Ou seja, sem o uso do PR dois sujeitos aleatórios compartilham cerca de 42% do seu teste. Essa taxa cai pela metade quando esse método é adotado. Leroux et al. (2013) obtiveram resultados semelhantes em simulações de CAT de tamanho fixo e variável. O valor da REQM aumentou no máximo em 0,04 e a taxa de sobreposição chegou a reduzir de 0,54 para 0,25 quando o PR foi adotado. Outros estudos têm apontado para a potencialidade do uso do PR para aumentar a segurança do teste sem comprometer a precisão em relação ao método de MIF (e.g., Lee & Dodd, 2012; Leroux et al., 2019).

O PR combina dois métodos de controle de exposição: a máxima informação restrita e o progressivo (Revuelta & Ponsoda, 1998). No método da máxima informação restrita, estabelece-se uma taxa máxima de exposição permitida para os itens. No início da aplicação, somente os itens com taxa de exposição menor do que a máxima ficam disponíveis para administração. Os itens remanescentes são selecionados pelo método de MIF.

No método progressivo, o item administrado é o que apresenta a maior soma entre dois componentes: um aleatório e um informativo. No início da aplicação, a importância do componente aleatório é de 100% e a do informativo, 0%. A importância do componente

aleatório reduz ao longo da aplicação e a do informativo, aumenta. Em termos de notação, no método progressivo o item selecionado j^* será aquele que:

$$j^* = arg\max_{j \in S} \left|(1 - W)R_j + WI_j\left(\hat{\theta}_t\right)\right| \quad (28)$$

onde S representa o banco de itens disponíveis para aplicação, $\hat{\theta}_t$ corresponde ao teta provisório após a administração de t itens, $I_j\left(\hat{\theta}_t\right)$ é a informação do item j para o teta provisório, R_j é um número aleatório sorteado de uma distribuição uniforme com intervalo entre zero e o valor da maior informação dentre os itens disponíveis no banco, $\left[0; max_{j \in S} I_j\left(\hat{\theta}_t\right)\right]$, e W é o peso que determina a importância dos componentes aleatório e informativo da equação para o item j. A expressão $arg\max_{j \in S}$ indica que será selecionado o item j com o maior resultado da função, ou seja, que apresente a maior soma entre os componentes aleatório e informativo.

Originalmente o peso W foi concebido da seguinte maneira (Revuelta & Ponsoda, 1998):

$$W = \frac{t}{M} \quad (29)$$

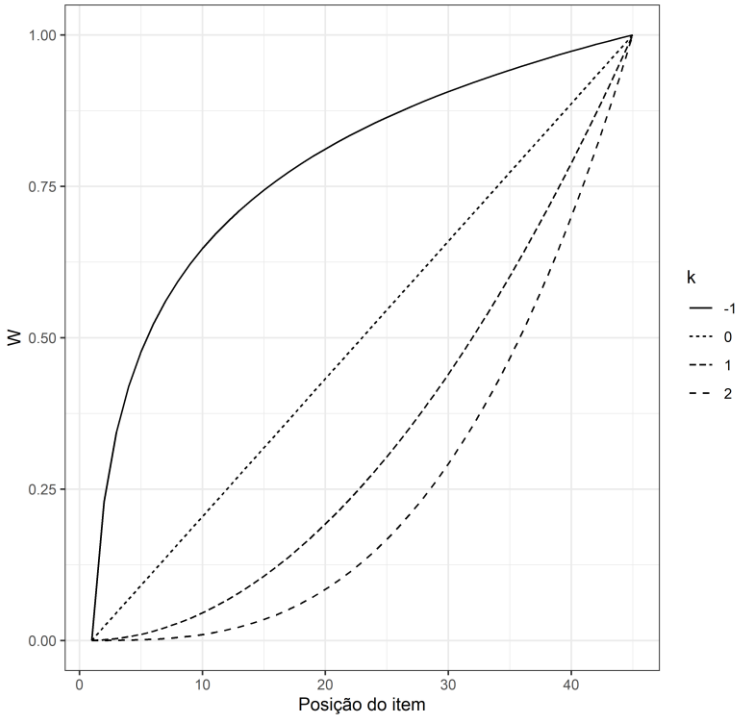
onde M é o tamanho do teste. Posteriormente, Barrada et al. (2008) propuseram outra forma de calcular esse peso:

$$W = \begin{cases} 0, & \text{se } q = 1 \\ \dfrac{\sum_{b=2}^{t}(b-1)^k}{\sum_{b=2}^{N}(b-1)^k}, & \text{se } q \neq 1 \end{cases} \quad (30)$$

onde N é o tamanho do teste, k é um parâmetro que permite controlar a velocidade com que W se distancia de 0 ao longo da aplicação e q é a posição do item na aplicação. Essa fórmula incorpora o parâmetro de aceleração k, que interfere na velocidade do aumento de W. Quanto maior o parâmetro k, mais lentamente W aumenta, e mais lentamente o componente aleatório perde importância. A Figura 17 ilustra quatro situações de valores de k em uma CAT de 45 itens.

Figura 17

Variação do peso no método progressivo restrito com diferentes parâmetros de aceleração, em CATs de tamanho fixo



Para testes de tamanho variável, W pode ser calculado da seguinte maneira (McClarty et al., 2006):

$$W = \frac{EP_{parada}}{EP_t} \tag{31}$$

onde EP_{parada} é o valor de parada do erro padrão e EP_t é o erro padrão do teta provisório. Magis e Barrada (2017) alteraram a forma de calcular o peso e incluíram o parâmetro de aceleração nela:

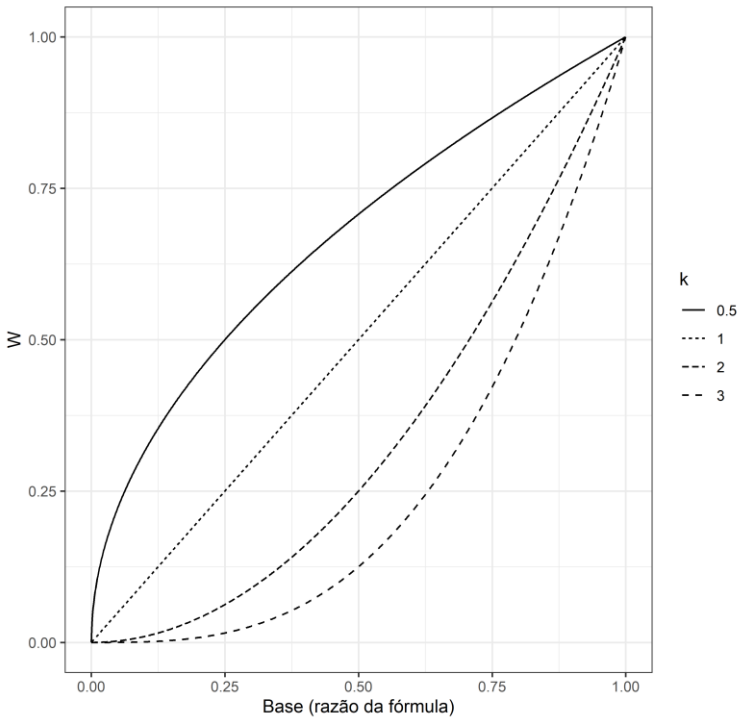
$$W = \begin{cases} 0, & \text{se } q = 1 \\ max\left[\frac{I(\hat{\theta}_t)}{I_{parada}}, \frac{q}{M-1}\right]^k, & \text{se } q \neq 1 \end{cases} \tag{32}$$

onde I_{parada} é a informação necessária para atingir o valor de parada do erro padrão e M é o tamanho máximo do teste. Neste trabalho utilizamos a Equação 32.

Existem duas situações na Equação 32 que afetam o peso do componente randômico na Equação 28 de forma indesejada: quando $k = 0$ e quando $k < 0$. Quando $k = 0$, a partir do segundo item tem-se que $W = 1$ constantemente (Equação 32). Nesse caso, o peso do componente randômico será sempre zero a partir da seleção do segundo item (Equação 28) e os itens serão selecionados exclusivamente segundo sua informação psicométrica. Já para $k < 0$, o valor de W diminui conforme a aplicação progride (Equação 32), e consequentemente o peso do componente randômico aumenta (Equação 28). O efeito é que o peso da informação do item será menor conforme a aplicação progride. Por isso, neste trabalho utilizamos somente valores positivos de k. A Figura 18 ilustra quatro situações de valores de k em uma CAT de tamanho variável. O eixo das abscissas corresponde à razão da fórmula, que é a base na Equação 32, ou seja, $max\left[\frac{I(\hat{\theta}_t)}{I_{parada}}, \frac{q}{M-1}\right]$.

Figura 18

Variação do peso no método progressivo restrito com diferentes parâmetros de aceleração, em CATs de tamanho variável



Nota. o eixo x corresponde à base na Equação 32.

A partir da Figura 17 e da Figura 18, nota-se que quanto maior o parâmetro de aceleração k, mais lentamente W aumenta, e o componente aleatório perde importância mais lentamente. Paralelamente, a importância da informação aumenta mais lentamente. Como consequência, a importância da discriminação do item (que tem relação direta com sua informação) também aumentará mais lentamente. Por isso, os itens menos discriminativos são mais expostos nessas condições em comparação a situações em que a importância da informação é maior (Barrada et al., 2008). Aumentar a exposição de itens pouco discriminativos pode reduzir a demanda por itens novos no banco, já que a exposição dos demais itens diminuirá. Apesar de o parâmetro de aceleração poder ter um impacto positivo nas características de uma CAT, seu efeito sobre a eficiência, precisão e segurança do teste tem sido pouco estudado (e.g., Barrada et al., 2008). Adicionalmente, não se sabe o efeito desse parâmetro em CAT de tamanho variável.

Exemplos de critérios de parada de uma CAT de tamanho variável são o erro padrão e a redução do erro após a administração de um item. A combinação desses critérios aumenta a eficiência da CAT em relação ao uso do erro padrão sozinho, pois a aplicação encerra quando a administração de novos itens não contribui para a redução do erro de medida. Kallen et al. (2018) reduziram uma CAT que tinha como critério de parada o erro padrão (0,3) e máximo de 12 itens. Os autores incluíram a redução do erro (0,01) como critério de parada e verificaram que as proporções de aplicações encerradas por esse critério chegaram a 13,3%, com pouco impacto na precisão. O tamanho dos testes reduziu em até 3,3 itens em média. Utilizar a diferença de 0,015 como valor da redução predita do erro como critério de parada também aumenta a eficiência da CAT com pouco impacto na precisão, quando não se tem um banco de itens com distribuição uniforme (Morris et al., 2020). A combinação do critério do erro padrão com o da redução do erro aumenta a eficiência da CAT, no entanto ela é pouco explorada.

Dadas as lacunas apontadas nos parágrafos acima, este trabalho selecionou o método PR para investigar em que medida ele altera a eficiência (em termos de tamanho do teste), a

precisão e a segurança (em termos de exposição dos itens) de uma CAT com diferentes critérios de parada. Nosso objetivo foi avaliar o controle de exposição PR com diferentes parâmetros de aceleração em CATs de tamanho fixo e variável. Simulamos a aplicação do Enem (um teste educacional de alto impacto aplicado em papel) em formato de CAT. O Enem, que é aplicado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), é composto por quatro provas de 45 itens de múltipla escolha, a saber: Ciências Humanas e suas Tecnologias (CH); Ciências da Natureza e suas Tecnologias (CN); Linguagens, Códigos e suas Tecnologias (LC); e Matemática e suas Tecnologias (MT).

Nossas perguntas de pesquisa foram as seguintes:

1. Como a eficiência, a precisão e a segurança do teste variam ao se adotar o controle de exposição PR?
2. Como a eficiência, a precisão e a segurança variam em função do parâmetro de aceleração do método PR?
3. Qual é o impacto sobre a precisão do Enem ao reduzir seu tamanho por meio de uma CAT com controle de exposição?

Elaboramos as seguintes hipóteses de estudo: (H1) a eficiência e a precisão do método PR serão semelhantes à do MIF; (H2) a segurança do PR será maior do que a do MIF; (H3) a eficiência e a precisão do método PR serão semelhantes com todos os parâmetros de aceleração; (H4) quanto maior o parâmetro de aceleração, maior a segurança; e (H5) a precisão de uma CAT com PR será maior do que a do teste linear. Este estudo avança porque avalia o método PR e diferentes valores do seu parâmetro de aceleração, que é pouco explorado na literatura, em especial em CAT de tamanho variável. Além disso, é inédito porque avalia a combinação do método PR com o critério de parada de redução do erro observado, que apesar de eficiente também é pouco explorado. Por último, o estudo de simulação utiliza bancos de itens robustos (mais de 700 itens) aplicados previamente em situação real.

Método

Desenho do estudo

Para cada tipo de CAT (tamanho fixo e tamanho variável), manipulamos duas variáveis: o controle de exposição de itens e o critério de parada. Como controle de exposição, utilizamos o método PR com dois parâmetros de aceleração: para as CATs de tamanho fixo, PR com $k = 1$ (PR1) e PR com $k = 2$ (PR2); para as CATs de tamanho variável, PR com $k = 2$ (PR2) e PR com $k = 3$ (PR3). Por isso, tivemos três condições de controle de exposição para cada tipo de CAT, sendo uma delas a ausência de controle. A taxa máxima de exposição dos itens para o PR foi estabelecida em 0,30.

Como método de seleção de itens, adotamos o MIF (sem controle de exposição e em conjunto com o método PR) e o aleatório. Este último foi utilizado somente para fins de referência, com o objetivo de quantificar a melhora das demais condições. Não utilizamos restrição de exposição com o método aleatório. Assim, tivemos de fato cinco métodos de seleção de itens (aleatório, MIF, PR1, PR2 e PR3), sendo quatro para cada tipo de CAT. Utilizamos quatro critérios de parada: para as CATs de tamanho fixo, 45 itens (TF45) e 20 itens (TF20); para as CATs de tamanho variável, erro padrão de 0,30 (EP30) e combinação de erro padrão 0,30 com redução do erro de 0,015 (EP30RE015).

Como para cada tipo de CAT tivemos quatro métodos de seleção de itens, quatro critérios de parada, replicamos cada banco 20 vezes e aplicamos o desenho às quatro provas do Enem, o total de simulações foi $4 \times 4 \times 20 \times 4 = 1280$. A Tabela 12 mostra as condições das simulações de CAT de tamanho fixo e a Tabela 13 mostra as condições das CATs de tamanho variável. Todos os comandos estão disponíveis em http://github.com/alexandrejaloto/tese_artigo2.

Tabela 12

Condições das simulações de CAT de tamanho fixo

Critério de parada	Método de seleção			
	Aleatório	Máxima informação de Fisher	PR com taxa de 0,30 e PA 1	PR com taxa de 0,30 e PA 2
Tamanho fixo de 45 itens	ALETF45	MIFTF45	PR1TF45	PR2TF45
Tamanho fixo de 20 itens	ALETF20	MIFTF20	PR1TF20	PR2TF20
Erro padrão de 0,30	ALEEP30	MIFEP30	PR1EP30	PR2EP30
Erro padrão de 0,30 ou redução do erro em 0,015	ALEEP30RE015	MIFEP30RE015	PR1EP30RE015	PR2EP30RE015

Nota. PR = progressivo restrito; PA = parâmetro de aceleração do método progressivo restrito.

Tabela 13

Condições das simulações de CAT de tamanho variável

Critério de parada	Método de seleção			
	Aleatório	Máxima informação de Fisher	PR com taxa de 0,30 e PA 2	PR com taxa de 0,30 e PA 3
Tamanho fixo de 45 itens	ALETF45	MIFTF45	PR2TF45	PR3TF45
Tamanho fixo de 20 itens	ALETF20	MIFTF20	PR2TF20	PR3TF20
Erro padrão de 0,30	ALEEP30	MIFEP30	PR2EP30	PR3EP30
Erro padrão de 0,30 ou redução do erro padrão em 0,015	ALEEP30RE015	MIFEP30RE015	PR2EP30RE015	PR3EP30RE015

Nota. PR = progressivo restrito; PA = parâmetro de aceleração do método progressivo restrito.

Banco de respostas

Para cada área do Enem, foi sorteada uma amostra aleatória simples dos participantes da edição de 2020. Os dados foram obtidos a partir dos microdados do Enem (disponíveis em https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem) em 01/11/2022. O tamanho amostral foi tal que garantisse uma média com erro padrão de 5 pontos na escala do Enem (o que equivale a um coeficiente de variação de 0,01) com intervalo de confiança de 95%. O cálculo do tamanho amostral n foi feito da seguinte maneira:

$$n = \left(\frac{\sigma \times Z}{d}\right)^2 \tag{33}$$

Onde σ é o desvio padrão da população, d é o erro aceitável (5) e Z é o valor de z na curva da distribuição normal com área de $[1 - 0{,}05/2]$. O valor de Z que garante um intervalo de confiança de 95% é aproximadamente 1,96.

Ao adotarmos esse procedimento, pudemos generalizar nossos resultados para a população dessa edição do Enem, o que potencialmente aproxima nossa simulação de situações esperadas para as futuras edições com características semelhantes. Adicionalmente, complementamos o tamanho amostral até que ele ficasse três vezes maior do que o número de itens no banco. Isso permitiu comparar com mais segurança resultados relacionados à exposição de itens, pois tornou constante a razão entre o tamanho do banco e o dos participantes. Por exemplo, se tivéssemos proporções diferentes, o fato de haver algum item não apresentado na simulação de uma área poderia ser devido ao grande tamanho do banco de itens em relação ao tamanho amostral. A título de ilustração, em LC o tamanho amostral necessário para garantir um erro padrão de 5 pontos é menor do que o tamanho do banco de itens.

As estatísticas descritivas dos participantes do Enem 2020 e das amostras sorteadas de cada área são apresentadas na Tabela 14. Apesar de as notas de cada prova serem divulgadas em uma métrica com média 500 e desvio padrão 100 (a referência são os participantes do Enem 2009 de escola pública concluintes do ensino médio regular), os parâmetros dos itens são divulgados em uma métrica com média 0 e desvio padrão 1 (a referência é a amostra de calibração dos itens do Enem 2009). Para facilitar as análises, neste trabalho optamos por utilizar a métrica em que os parâmetros são divulgados. A descrição das transformações das métricas está disponível no Artigo 1 desta tese.

Tabela 14

Estatísticas descritivas dos participantes do Enem 2020 e das amostras das simulações

Área	n	Média (desvio padrão)	Intervalo
Ciências Humanas			
Participantes de 2020	2.749.073	0,09 (0,83)	-1,67–3,22
Amostra da simulação	2.268	0,06 (0,82)	-1,53–2,43
Ciências da Natureza			
Participantes de 2020	2.596.735	-0,09 (0,70)	-1,57–3,13
Amostra da simulação	2.247	-0,09 (0,71)	-1,48–2,51
Linguagens e Códigos			
Participantes de 2020	2.751.791	0,22 (0,68)	-1,95–2,79
Amostra da simulação	2.649	0,22 (0,67)	-1,79–2,07
Matemática			
Participantes de 2020	2.596.527	0,16 (0,90)	-1,33–3,66

Amostra da simulação	2.376	0,14 (0,90)	-1,33–3,35

As respostas aos itens foram geradas a partir de uma simulação de Monte Carlo por meio da função gen.resp do pacote simCAT (disponível em http://github.com/alexandrejaloto/simCAT). Com isso, obtivemos quatro matrizes de respostas, uma para cada área do conhecimento. As linhas de cada matriz corresponderam aos sujeitos da amostra, e as colunas corresponderam aos itens de cada área. Ou seja, produzimos um banco de respostas como se cada sujeito tivesse respondido a todos os itens de uma área. Cada banco de respostas foi gerado 20 vezes para fins de replicação da simulação.

Especificações da CAT

O banco de itens foi composto pelos itens aplicados no Enem de 2009 a 2020 disponibilizados nos microdados. Os arquivos foram obtidos em 01/11/2022, com exceção dos microdados do Enem 2011, que foram baixados em 18/11/2022.

Excluímos os itens que não possuíam informação sobre o conteúdo avaliado (a habilidade da matriz de referência do Enem) e aqueles que a equipe do Inep excluíra das análises (esses não possuíam parâmetros da TRI). Alguns itens das edições de 2016 e 2020 foram utilizados em duas aplicações do mesmo ano e foram contabilizados somente uma vez no banco deste trabalho. Os bancos variaram de 749 (CN) a 883 (LC) itens. A Tabela 15 apresenta a descrição dos quatro bancos de itens. A Figura 19 mostra a curva de informação do teste e a curva de densidade do teta da amostra.

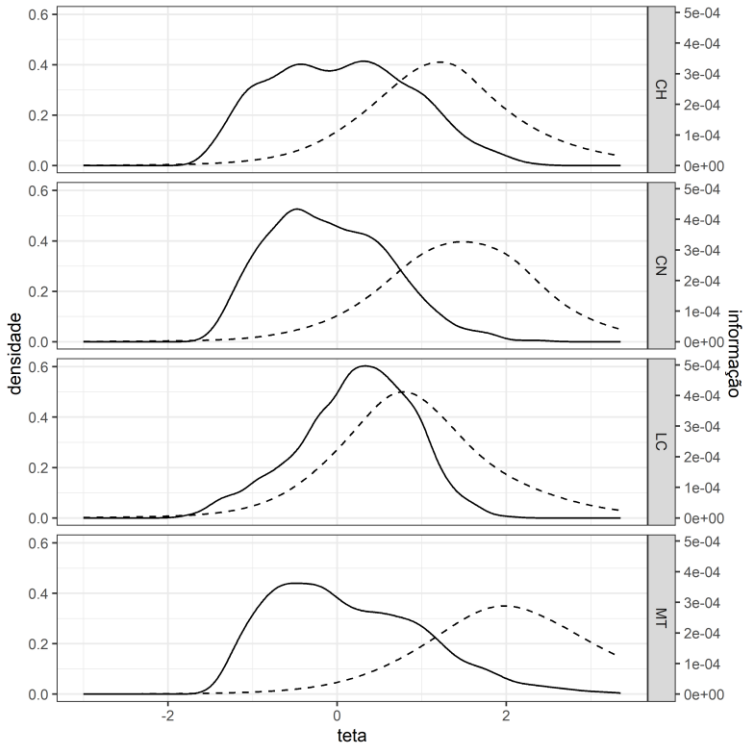
Tabela 15

Descrição dos quatro bancos de itens

Área	n	média (desvio padrão)		
		a	b	c
CH	756	2,18 (0,95)	1,09 (0,80)	0,17 (0,07)
CN	749	2,34 (1,01)	1,39 (2,35)	0,17 (0,07)
LC	883	2,21 (0,9)	0,79 (0,82)	0,16 (0,07)
MT	792	2,09 (0,78)	1,97 (1,33)	0,16 (0,06)

Figura 19

Curva de informação dos bancos e itens e distribuição da densidade dos tetas de cada amostra



Nota. A linha contínua representa a curva de densidade do teta da amostra e a linha tracejada representa a informação do banco de itens.

A seleção do primeiro item foi aleatória para as condições que utilizaram o método aleatório ou o método PR. Para o método MIF, o teta inicial foi a média das notas de cada área no Enem 2020.

Para o balanceamento de conteúdo do teste, utilizamos o método CAT restrita modificado (*modified constrained* CAT, MCCAT; Leung et al., 2000). O pacote simCAT restringe o teste para que uma mesma habilidade só seja aplicada novamente depois de todas as outras já terem sido aplicadas. Nas aplicações com menos de 30 itens, algumas habilidades não foram contempladas. No entanto, não consideramos esse ponto uma limitação para a simulação de uma situação real de aplicação, pois é comum uma prova do Enem não conter itens de todas as habilidades (e.g., LC 2009, MT 2009, CH 2019 e CN 2019).

Estimamos o teta pelo método EAP, com 40 pontos de quadratura de -4,0 a 4,0. As aplicações de tamanho variável tinham no mínimo 15 itens e no máximo 60. A simulação se deu com o pacote simCAT, que se baseia no pacote catR (Magis & Raîche, 2012) mas possui algumas diferenças, como inclusão de balanceamento de conteúdo por MCCAT e critério de parada da redução do erro.

Avaliação da CAT

A eficiência das CATs de tamanho variável foi avaliada em termos dos valores mínimo, máximo, média e mediana do tamanho das aplicações. Quanto menores os valores, maior a redução que a condição da CAT proporciona, portanto, mais eficiente ela se torna. A precisão da CAT foi avaliada a partir da correlação, do viés e da REQM entre o teta real e o teta simulado. Além disso, avaliamos o erro padrão de medida do teta simulado. O teta real correspondeu à nota oficial do participante no Enem 2020 na métrica transformada com média 0 e desvio padrão 1 e o teta simulado correspondeu à nota obtida na simulação. Neste trabalho, reportamos a média dos indicadores nas 20 replicações das condições.

O viés é uma medida de distância entre o teta real e o simulado. Ele corresponde à diferença média entre o teta real e o simulado e é calculado da seguinte maneira

$$V = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i\right)}{n} \tag{34}$$

onde n é o total de sujeitos, $\hat{\theta}_i$ é o teta estimado do sujeito i e θ_i é o teta real do sujeito i. A REQM é calculada da seguinte maneira

$$REQM = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \tag{35}$$

Quanto menor a REQM, mais precisa é a CAT, pois menor a distância entre o teta simulado e o teta real.

A segurança da CAT foi avaliada a partir da exposição e da sobreposição dos itens. A exposição do item varia de zero a um e foi medida pela razão entre a quantidade de vezes que

ele foi administrado e a quantidade de sujeitos. Um item com taxa de exposição de zero não foi aplicado. Quanto mais itens não administrados, mais o banco está subaproveitado. Uma taxa de exposição maior do que 0,30 indica que pode haver superexposição de algum item. Quanto mais itens superexpostos, menos seguro o banco, pois os itens foram apresentados a grandes proporções dos participantes. Além de calcularmos a taxa de exposição do item, estabelecemos intervalos de taxa de exposição e verificamos a quantidade de itens em cada intervalo.

A sobreposição corresponde à proporção de itens iguais aplicados a dois participantes selecionados aleatoriamente. Quanto maior a sobreposição, menos seguro o banco. Ela é calculada da seguinte maneira (Chen et al., 2003)

$$\bar{T} = \frac{S^2 + \bar{r}}{\bar{r}} \tag{36}$$

onde S^2 é a variância das taxas de exposição dos itens e $\bar{r}$ é a taxa de exposição média dos itens.

Além de reportarmos a média geral dos indicadores nas 20 replicações, analisamos graficamente o erro padrão e a REQM condicionados ao teta em cada área. Para esse gráfico, também utilizamos a média das 20 replicações. Por último, selecionamos a condição com melhores indicadores de eficiência, precisão e segurança e comparamos sua precisão com a precisão do teste linear do Enem 2020.

Simulação do teste linear

De posse das matrizes de respostas simuladas, selecionamos as respostas aos itens da primeira aplicação do Enem 2020 e calculamos a nota de cada participante em cada prova, da mesma maneira que o cálculo do teta nas CATs. Em seguida, comparamos os indicadores de precisão (erro padrão, viés, correlação e REQM) do teste linear com os indicadores da CAT selecionada na etapa anterior. Adicionalmente, comparamos a REQM condicionada ao teta dos dois testes.

Resultados

Eficiência das CATs

A Tabela 16 mostra os indicadores de eficiência (média dos valores mínimo, máximo, média e mediana do tamanho das aplicações) das oito condições de CAT de tamanho variável de cada área. No que diz respeito ao critério de parada EP30, nas quatro áreas o tamanho médio das aplicações com o método MIF foi menor do que com o PR2 e PR3. Em LC, a diferença foi de menos de dois itens, e em CH, CN e MT a diferença passou de cinco itens, chegando a 14. Os três métodos superaram de maneira mais relevante o aleatório, com diferença de até 30 itens (CN, MIF). O valor médio das medianas para o MIF foi 15 em CH, CN e LC, o que significa que nessas áreas 15 itens foram suficientes para medir o teta de 50% da amostra com esse método de seleção. Em MT, a mediana para essa condição foi 20. Para o método PR2, esse valor variou de 15 (LC) a 40,1 (MT). Para o PR3, os valores variaram de 15 (LC) a 46,1 (MT). As aplicações com o método aleatório tiveram valores de mediana superiores, chegando a 60 em CN e MT. Não houve grandes diferenças entre os métodos para os valores médios do tamanho mínimo (de 15 a 15,4) e máximo (em todas as condições, 60) das aplicações nos quatro métodos de seleção e nas quatro áreas.

As aplicações com o critério de parada EP30RE015 tiveram tamanho menor em comparação com EP30. Os valores médios do tamanho máximo das aplicações não passaram de 23,5 e o valor médio da mediana foi 15 em todas as áreas. Ou seja, em geral 50% da amostra respondeu 15 itens em todas as condições com EP30RE015. Adicionalmente, a diferença do tamanho médio entre os métodos de seleção com EP30RE015 foi irrelevante e não chegou a um item.

Tabela 16

Média dos valores mínimo, máximo, da média e da mediana de itens apresentados nas simulações

Parada	Seleção	CH				CN				LC				MT			
		Min	Máx	NIA	MIA	Min	Máx	NIA	MIA	Min	Máx	NIA	MIA	Min	Máx	NIA	MIA
EP30	ALE	15,0	60,0	45,1	56,6	15,0	60,0	50,8	60,0	15,0	60,0	33,5	27,2	15,4	60,0	54,6	60,0
	MIF	15,0	60,0	21,7	15,0	15,0	60,0	20,7	15,0	15,0	60,0	16,5	15,0	15,0	60,0	31,2	20,0

	PR2	15,0	60,0	26,9	19,0	15,0	60,0	29,6	24,0	15,0	60,0	19,6	15,0	15,0	60,0	40,5	40,1
	PR3	15,0	60,0	31,2	26,3	15,0	60,0	34,7	33,3	15,0	60,0	18,3	15,0	15,0	60,0	44,2	46,1
	ALE	15,0	21,4	15,4	15,0	15,0	20,7	15,3	15,0	15,0	20,9	15,5	15,0	15,0	20,8	15,3	15,0
EP30RE015	MIF	15,0	18,9	15,1	15,0	15,0	19,1	15,1	15,0	15,0	18,6	15,0	15,0	15,0	19,5	15,2	15,0
	PR2	15,0	22,1	15,4	15,0	15,0	22,4	15,5	15,0	15,0	19,4	15,1	15,0	15,0	23,5	15,7	15,0
	PR3	15,0	21,9	15,5	15,0	15,0	21,6	15,4	15,0	15,0	19,0	15,1	15,0	15,0	22,5	15,5	15,0

Nota. CH = Ciências Humanas; CN = Ciências da Natureza; LC = Linguagens e Códigos; MT = Matemática; NIA = número médio de itens aplicados; min = mínimo; máx = máximo; MIA = mediana de itens aplicados; ALE = aleatório; MIF = Máxima Informação de Fisher; PR2 = progressivo restrito com $k = 2$; PR3 = progressivo restrito com $k = 3$; EP30 = erro padrão de 0,30; EP30RE015 = erro padrão de 0,30 ou redução do erro padrão em 0,015.

Precisão das CATs

A Tabela 17 mostra os indicadores de precisão (média dos valores de erro padrão, correlação, viés e REQM) das oito condições de CAT de tamanho fixo de cada área. Em todas as áreas o método MIF apresentou os melhores valores desses indicadores, quando comparado com os demais métodos de seleção dentro do mesmo critério de parada. Já o método aleatório apresentou os piores valores de precisão. Apesar da superioridade do MIF, de maneira geral a diferença entre ele e os métodos PR1 e PR2 foi irrelevante. A maior média de erro padrão excluindo-se o método aleatório foi 0,376, valor que equivale a uma confiabilidade de 0,859. A menor correlação das condições sem método aleatório foi 0,926, a maior REQM foi 0,347 e o maior viés, 0,058. O viés médio foi positivo em todas as condições, ou seja, as simulações superestimaram as notas dos sujeitos. As maiores diferenças nos indicadores de precisão foram observadas em MT e não chegaram a 0,07 em todos os indicadores. Isso mostra que as condições de tamanho fixo com PR tiveram valores satisfatórios de precisão.

Tabela 17

Média do erro padrão de medida, correlação, viés e raiz do erro quadrático médio das replicações das CATs de tamanho fixo

Parada	Seleção	CH				CN				LC				MT			
		EP	COR	Viés	REQM	EP	COR	Viés	REQM	EP	COR	Viés	REQM	EP	COR	Viés	REQM
	ALE	0,370	0,918	0,053	0,340	0,423	0,870	0,066	0,372	0,271	0,929	0,045	0,268	0,510	0,880	0,078	0,439
TF45	MIF	0,176	0,976	0,021	0,186	0,183	0,967	0,025	0,189	0,129	0,981	0,010	0,134	0,241	0,965	0,035	0,243
	PR1	0,200	0,969	0,024	0,208	0,224	0,953	0,031	0,227	0,189	0,960	0,021	0,197	0,288	0,954	0,042	0,281

TF20	PR2	0,211	0,967	0,026	0,217	0,232	0,950	0,032	0,233	0,187	0,961	0,021	0,194	0,297	0,952	0,045	0,287
	ALE	0,496	0,857	0,060	0,436	0,552	0,782	0,065	0,463	0,398	0,860	0,076	0,378	0,639	0,805	0,088	0,544
	MIF	0,240	0,960	0,029	0,239	0,242	0,948	0,032	0,238	0,185	0,964	0,019	0,187	0,311	0,947	0,046	0,299
	PR1	0,269	0,951	0,034	0,264	0,284	0,932	0,039	0,273	0,209	0,955	0,023	0,209	0,360	0,933	0,053	0,337
	PR2	0,285	0,947	0,035	0,275	0,300	0,926	0,042	0,285	0,218	0,952	0,024	0,217	0,376	0,928	0,058	0,347

Nota. CH = Ciências Humanas; CN = Ciências da Natureza; LC = Linguagens e Códigos; MT = Matemática; EP = erro padrão de medida; COR = correlação; REQM = raiz do erro quadrático médio; ALE = aleatório; MIF = Máxima Informação de Fisher; PR1 = progressivo restrito com $k = 1$; PR2 = progressivo restrito com $k = 2$; TF20 = tamanho fixo de 20 itens; TF45 = tamanho fixo de 45 itens.

A Tabela 18 mostra os indicadores de precisão para as CATs de tamanho variável. Nessas condições, o método MIF também apresentou os melhores valores de precisão e o aleatório, os piores. Com o critério de parada EP30, a diferença entre o MIF e os métodos PR2 e PR3 foi irrelevante e não chegou a 0,06 em todas as áreas, para todos os indicadores de precisão. Com esse critério de parada, a maior média de erro padrão excluindo-se o método aleatório foi 0,342, valor que equivale a uma confiabilidade de 0,883. A menor correlação das condições sem método aleatório foi 0,921, a maior REQM foi 0,323 e o maior viés, 0,048. O viés médio foi positivo em todas as condições.

As condições com o critério EP30RE015 tiveram uma precisão mais baixa em relação ao EP30. Adicionalmente, as diferenças entre o MIF e os métodos PR2 e PR3 foram maiores, chegando a 0,29 na média do erro padrão em MT. Em CH, CN e MT o PR3 se comportou de forma similar ao aleatório. Com o critério de parada EP30RE015, a maior média de erro padrão excetuando-se o método aleatório foi 0,628 (confiabilidade de 0,606). A menor correlação das condições sem método aleatório foi 0,792, a maior REQM foi 0,541 e o maior viés, 0,088. O viés médio foi positivo em todas as condições. Nas CATs de tamanho variável, a precisão do PR com critério EP30 foi satisfatória e similar à observada nas CATs de tamanho fixo. No entanto, com critério EP30RE015 a precisão foi insatisfatória.

Tabela 18

Média do erro padrão de medida, correlação, viés e raiz do erro quadrático médio das replicações das CATs de tamanho variável

Parada	Seleção	CH				CN				LC				MT			
		EP	COR	Viés	REQM	EP	COR	Viés	REQM	EP	COR	Viés	REQM	EP	COR	Viés	REQM
EP30	ALE	0,375	0,918	0,048	0,338	0,404	0,881	0,054	0,360	0,313	0,911	0,034	0,293	0,481	0,894	0,076	0,418
	MIF	0,239	0,959	0,017	0,238	0,244	0,946	0,018	0,237	0,202	0,959	0,016	0,198	0,287	0,952	0,034	0,282
	PR2	0,274	0,947	0,019	0,270	0,286	0,925	0,015	0,278	0,240	0,943	0,020	0,233	0,333	0,940	0,043	0,317
	PR3	0,286	0,943	0,020	0,279	0,294	0,921	0,015	0,287	0,231	0,947	0,018	0,224	0,342	0,937	0,048	0,323
EP30RE015	ALE	0,540	0,828	0,064	0,474	0,594	0,745	0,064	0,494	0,446	0,829	0,088	0,420	0,682	0,772	0,095	0,583
	MIF	0,268	0,952	0,032	0,261	0,264	0,940	0,034	0,257	0,211	0,955	0,022	0,209	0,334	0,941	0,048	0,316
	PR2	0,384	0,910	0,055	0,354	0,434	0,862	0,068	0,381	0,270	0,928	0,038	0,268	0,519	0,874	0,077	0,450
	PR3	0,469	0,868	0,061	0,420	0,538	0,792	0,072	0,453	0,254	0,936	0,031	0,251	0,628	0,808	0,088	0,541

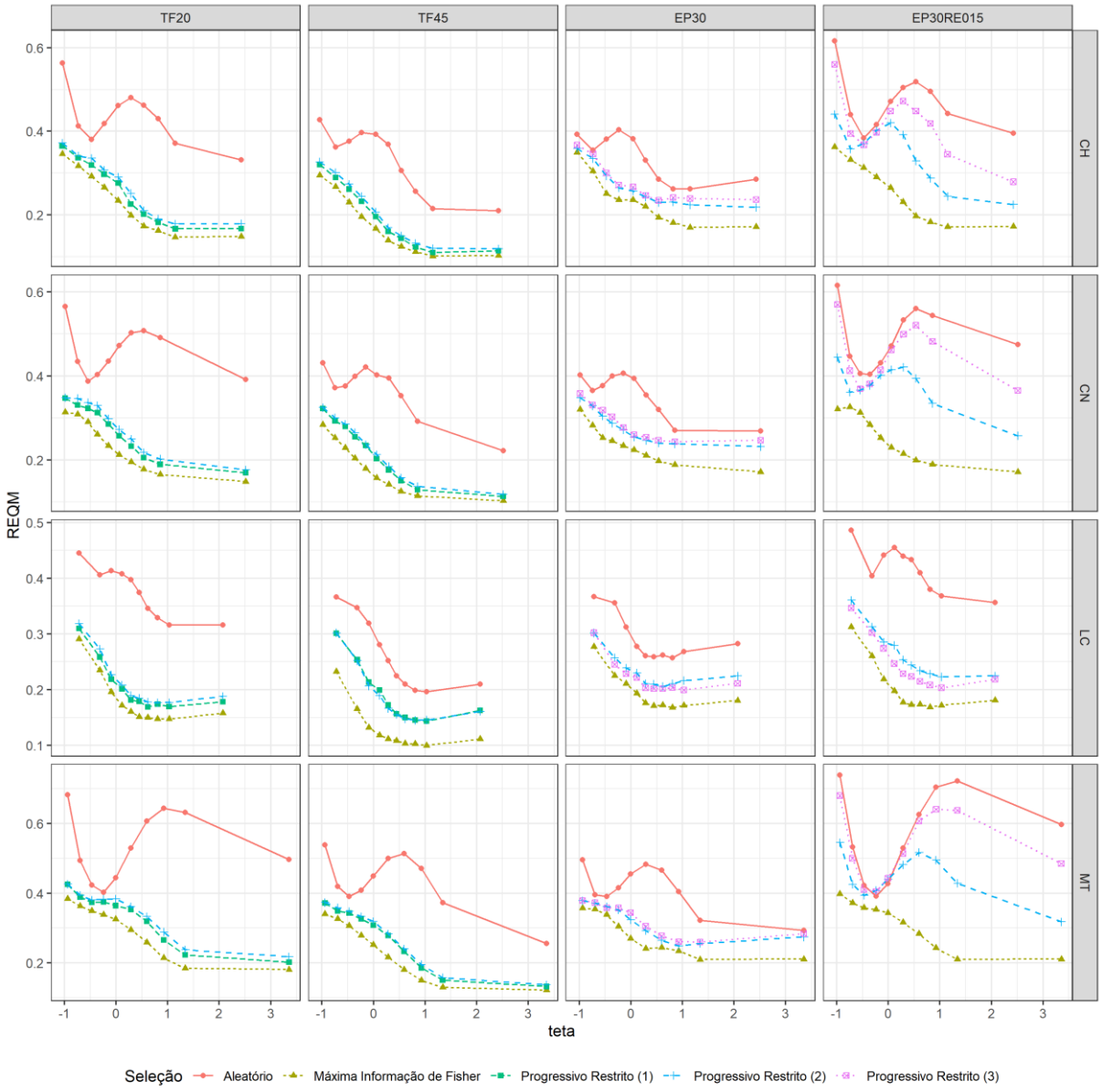
Nota. CH = Ciências Humanas; CN = Ciências da Natureza; LC = Linguagens e Códigos; MT = Matemática; EP = erro padrão de medida; COR = correlação; REQM = raiz do erro quadrático médio; ALE = aleatório; MIF = Máxima Informação de Fisher; PR1 = progressivo restrito com $k = 1$; PR2 = progressivo restrito com $k = 2$; EP30 = erro padrão de 0,30; EP30RE015 = erro padrão de 0,30 ou redução do erro padrão em 0,015.

REQM e erro padrão condicionado ao teta

A Figura 20 mostra a REQM de cada condição ao longo da escala do Enem. Nas CATs de tamanho fixo o uso de controle de exposição teve pouco impacto na REQM ao longo de toda a escala. Adicionalmente, a diferença entre PR1 e PR2 foi praticamente inexistente. Ainda, MIF, PR1 e PR2 apresentaram um ganho relevante na precisão em relação ao método aleatório. O mesmo padrão se observa na CAT de tamanho variável com critério de parada EP30: MIF, PR2 e PR3 apresentaram um ganho relevante na precisão, o controle de exposição não impactou de forma relevante a precisão e a diferença entre PR2 e PR3 foi praticamente inexistente. Já na CAT com critério EP30RE015, somente em LC esse padrão se replicou. No entanto, em CH, CN e MT o ganho na precisão proporcionado pelo MIF não foi observado em PR2 e PR3. Ao incluir o controle de exposição, a REQM nas regiões mais baixas da escala (abaixo de zero) ficou semelhante à do método aleatório. Nas regiões mais altas, apesar de PR2 e PR3 terem melhorado a precisão, ela ainda ficou bem abaixo da precisão do MIF. De forma geral, a Figura 20 mostra que a precisão com uso do PR como controle de exposição foi adequada nas CATs de tamanho fixo e com o critério de parada EP30.

Figura 20

REQM condicionada ao teta

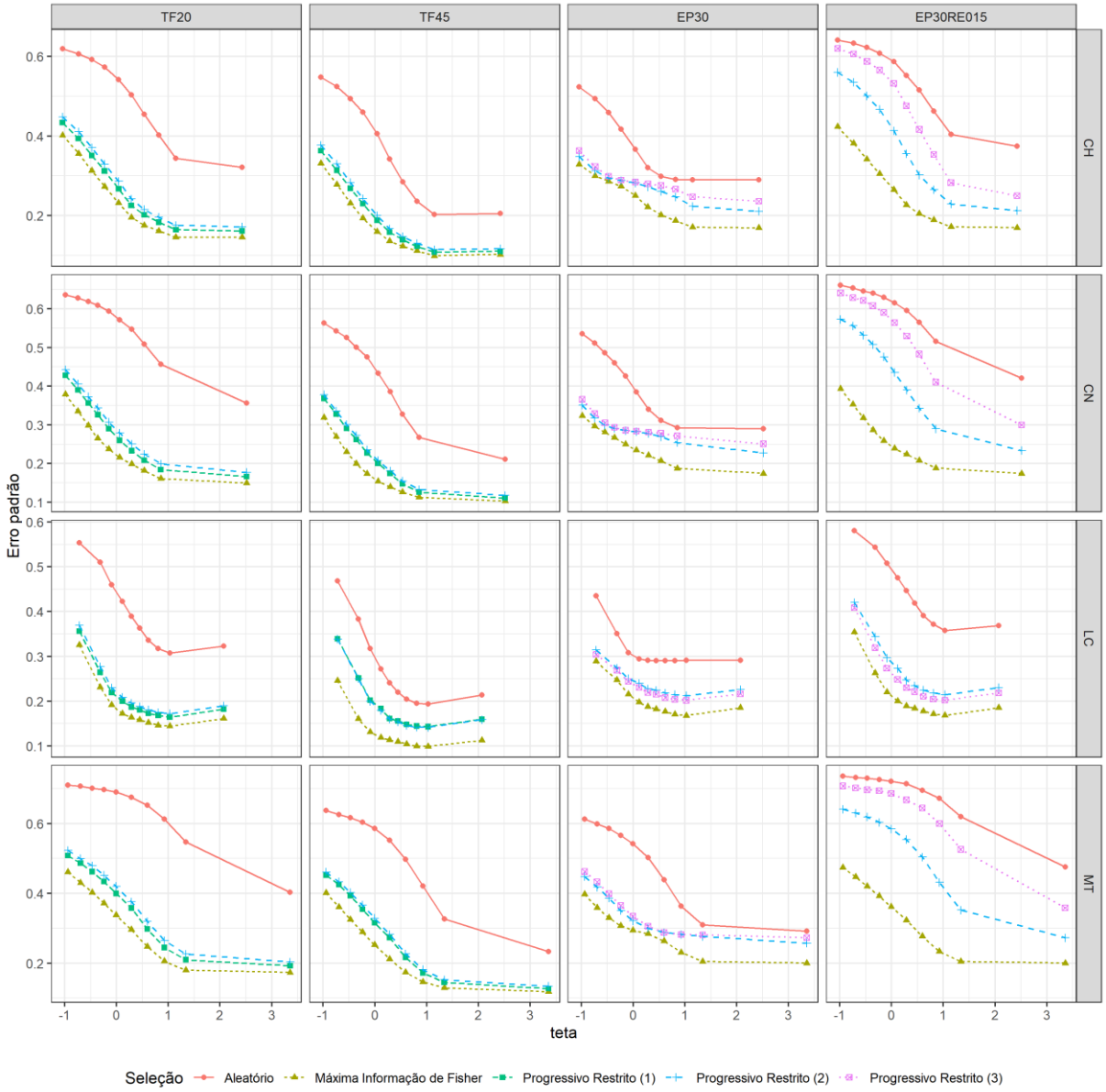


A Figura 21 mostra o erro padrão de cada condição ao longo da escala do Enem. Nas CATs de tamanho fixo, de forma geral o padrão de ganho no erro padrão foi similar ao padrão de ganho na REQM: MIF, PR1 e PR2 apresentaram um ganho relevante na precisão, o controle de exposição não impactou de forma relevante a precisão e a diferença entre PR1 e PR2 foi praticamente inexistente. Nas CATs de tamanho variável, em LC esse padrão também se repetiu. Já em CH, CN e MT com critério EP30, os métodos MIF, PR2 e PR3 apresentaram

ganho relevante de precisão nas partes mais baixas da escala e sem grandes diferenças entre esses métodos. No entanto, nas partes mais altas da escala a precisão teve um ganho relevante com o método MIF, mas não com os métodos PR2 e PR3. Nessas três áreas, com o critério EP30RE015 a precisão com o método MIF teve um ganho relevante em toda a escala. Porém, o ganho com PR2 e PR3 não foi relevante nas partes baixas da escala e a precisão ficou insatisfatória. Adicionalmente, a precisão com PR3 foi semelhante à precisão com seleção aleatória. Nas partes altas da escala, o método PR2 teve um ganho relevante, porém não tão alto quanto o do MIF. O mesmo ocorreu com o método PR3. De forma geral, a Figura 21 mostra que a precisão com uso do PR como controle de exposição foi adequada nas CATs de tamanho fixo e com o critério de parada EP30.

Figura 21

Erro padrão condicionado ao teta



Segurança das CATs

Os resultados sobre a segurança dos itens não são facilmente comparáveis entre todas as condições, diferentemente dos resultados sobre a precisão. Isso porque testes maiores naturalmente utilizam mais itens em uma administração, o que aumenta a quantidade de vezes que um item é administrado. Portanto, é esperado que a proporção de itens não administrados seja menor. Além disso, uma taxa de sobreposição de 0,30 em um teste de 20 itens representa a administração de seis itens comuns entre dois sujeitos, enquanto em um teste de 40 itens isso

representa 12 itens. Por isso, comparamos as exposições dos itens com cautela entre as diferentes condições de CAT.

A Tabela 19 mostra os indicadores de segurança das CATs de tamanho fixo de cada área (média das taxas mínima e máxima de exposição e da taxa de sobreposição). Em todas as áreas, nas condições com método aleatório ou PR2 todos os itens foram apresentados pelo menos uma vez, independentemente do tamanho da prova. Com o método PR1, somente em LC (TF20) houve item não administrado. Com o método MIF, houve item não administrado em todas as áreas, com os dois tamanhos de prova.

Como esperado, as taxas máximas de exposição com o método MIF foram de 1,00. Já com o controle de exposição, essas taxas foram de no máximo 0,30 (com arredondamento). Alguns itens ultrapassaram essa taxa, porém isso é inerente ao método de controle de exposição e ocorre devido à forma como o filtro de itens é realizado no início da aplicação. Com o método de máxima informação restrita, os itens que possuem mais de 30% de exposição são excluídos e voltam a ser disponibilizados após algumas aplicações, quando sua exposição fica menor do que o valor limite. Os itens que superaram 30% de exposição após as simulações são aqueles que necessariamente não seriam disponibilizados nas próximas aplicações, caso houvesse outras.

A taxa de sobreposição dos itens reduziu com a inclusão do controle de exposição. Com o método MIF, chegou a 0,432 (MT) com o critério TF45 e a 0,404 (MT) com o critério TF20. Com o PR2, as maiores taxas de sobreposição foram 0,151 (CN) com TF45 e 0,125 (MT) com TF20. As taxas de sobreposição do PR1 ficaram ligeiramente maiores do que as do PR2.

Tabela 19

Média das taxas mínima e máxima de exposição e da taxa de sobreposição nas CATs de tamanho fixo

Seleção	Parada	CH			CN			LC			MT		
		r_{min}	r_{max}	SP	r_{min}	r_{max}	SP	r_{min}	r_{max}	SP	r_{min}	r_{max}	SP

TF45	ALE	0,037	0,092	0,061	0,038	0,089	0,062	0,014	0,098	0,056	0,035	0,098	0,059
	MIF	0,000	1,000	0,335	0,000	1,000	0,419	0,000	1,000	0,279	0,000	1,000	0,432
	PR1	0,004	0,300	0,174	0,005	0,300	0,182	0,002	0,300	0,155	0,005	0,300	0,181
	PR2	0,010	0,300	0,143	0,012	0,300	0,151	0,004	0,300	0,128	0,011	0,300	0,149
TF20	ALE	0,015	0,040	0,027	0,015	0,041	0,027	0,006	0,039	0,024	0,014	0,038	0,026
	MIF	0,000	1,000	0,323	0,000	1,000	0,383	0,000	1,000	0,264	0,000	1,000	0,404
	PR1	0,001	0,300	0,143	0,001	0,300	0,156	0,000	0,300	0,120	0,001	0,300	0,157
	PR2	0,002	0,300	0,110	0,003	0,300	0,122	0,001	0,299	0,091	0,003	0,300	0,125

Nota. CH = Ciências Humanas; CN = Ciências da Natureza; LC = Linguagens e Códigos; MT = Matemática; r_{min} = menor taxa de exposição média de um item; r_{max} = maior taxa de exposição média de um item; SP = sobreposição média; ALE = aleatório; MIF = Máxima Informação de Fisher; PR1 = progressivo restrito com $k = 1$; PR2 = progressivo restrito com $k = 2$; TF20 = tamanho fixo de 20 itens; TF45 = tamanho fixo de 45 itens.

A Tabela 20 mostra os indicadores de segurança das CATs de tamanho variável de cada área. Em CH, CN e MT, nas condições com método aleatório ou com PR todos os itens foram apresentados pelo menos uma vez, independentemente do critério de parada. Em LC, nos dois critérios de parada houve item não administrado mesmo com controle de exposição. Assim como nas CATs de tamanho fixo, o MIF foi o que mais subutilizou o banco, pois teve a maior quantidade de itens não apresentados. Com o método MIF, houve item não administrado em todas as áreas, com os dois critérios de parada. Como esperado, as taxas máximas de exposição com o método MIF foram de 1,00. Já com o controle de exposição, essas taxas foram de no máximo 0,30 (com arredondamento). Com o critério de parada EP30RE015, em CH, CN e MT as taxas com controle de exposição chegaram a no máximo 0,217 (CH).

A taxa de sobreposição dos itens reduziu com a inclusão do controle de exposição. Com o método MIF, chegou a 0,406 (MT) com ambos os critérios de parada. Com o PR2, as maiores taxas de sobreposição foram 0,167 (MT) com critério EP30 e 0,151 (LC) com EP30RE015. As taxas de sobreposição do PR3 chegaram a no máximo 0,139 (MT) com critério EP30. Com o critério de parada EP30RE015, em CH, CN e MT os valores de sobreposição ficaram semelhantes aos do método aleatório.

Tabela 20

Média das taxas mínima e máxima de exposição e da taxa de sobreposição nas CATs de tamanho variável

Seleção	Parada	CH			CN			LC			MT		
		r_{min}	r_{max}	SP	r_{min}	r_{max}	SP	r_{min}	r_{max}	SP	r_{min}	r_{max}	SP
EP30	ALE	0,036	0,096	0,061	0,040	0,105	0,070	0,009	0,075	0,042	0,041	0,131	0,072
	MIF	0,000	1,000	0,299	0,000	1,000	0,340	0,000	1,000	0,265	0,000	1,000	0,406
	PR2	0,005	0,300	0,119	0,007	0,300	0,127	0,000	0,300	0,148	0,007	0,300	0,167
	PR3	0,011	0,299	0,095	0,014	0,300	0,103	0,000	0,300	0,125	0,014	0,300	0,139
EP30RE015	ALE	0,011	0,031	0,021	0,011	0,032	0,021	0,004	0,030	0,019	0,011	0,029	0,020
	MIF	0,000	1,000	0,335	0,000	1,000	0,389	0,000	1,000	0,284	0,000	1,000	0,406
	PR2	0,004	0,217	0,047	0,006	0,177	0,042	0,000	0,300	0,151	0,006	0,127	0,033
	PR3	0,008	0,109	0,025	0,010	0,073	0,023	0,000	0,300	0,125	0,009	0,045	0,021

Nota. CH = Ciências Humanas; CN = Ciências da Natureza; LC = Linguagens e Códigos; MT = Matemática; r_{min} = menor taxa de exposição média de um item; r_{max} = maior taxa de exposição média de um item; SP = sobreposição; ALE = aleatório; MIF = Máxima Informação de Fisher; PR2 = progressivo restrito com $k = 2$; PR3 = progressivo restrito com $k = 3$; EP30 = erro padrão de 0,30; EP30RE015 = erro padrão de 0,30 ou redução do erro padrão em 0,015.

A Tabela 21 mostra a média da porcentagem de itens por intervalo de taxa de exposição nas aplicações de tamanho fixo. Para obter esse valor, calculamos a média das taxas de exposição de cada item nas 20 replicações e verificamos a quantidade de itens em cada intervalo estabelecido. Para os dois tamanhos de prova o método MIF foi o que apresentou maiores proporções de itens não administrados (chegando a 82,6% em MT com TF20) e de itens superexpostos (chegando a 7,5% em CH e CN com TF45). Como esperado, a inclusão do controle de exposição reduziu a proporção de itens não apresentados e superexpostos. A média de itens não administrados com os métodos PR1 e PR2 em todas as áreas e com os dois tamanhos de prova foi zero. Alguns poucos itens tiveram taxas de exposição maiores do que 0,30, porém isso é inerente ao método de controle de exposição, como já explicado. No que diz respeito ao parâmetro de aceleração, quanto maior seu valor, menor a proporção de itens com taxas de exposição mais elevadas, para os dois tamanhos de prova.

Tabela 21

Média geral das porcentagens de itens para cada intervalo de taxa de exposição nas CATs de tamanho fixo

Critério de parada		CH				CN				LC				MT			
	Exposição	ALE	MIF	PR1	PR2	ALE	MIF	PR1	PR2	ALE	MIF	PR1	PR2	ALE	MIF	PR1	PR2
TF20	0	0,0	79,8	0,0	0,0	0,0	80,2	0,0	0,0	0,0	74,5	0,0	0,0	0,0	82,6	0,0	0,0
	(0;0,02]	0,0	6,7	78,7	76,1	0,0	7,7	80,6	78,8	24,3	9,7	80,1	77,3	0,0	5,8	81,6	80,4
	(0,02;0,05]	100,0	2,4	9,3	12,8	100,0	2,7	8,0	11,3	75,7	5,0	8,4	11,9	100,0	2,7	7,4	9,2
	(0,05;0,1]	0,0	2,1	4,4	5,0	0,0	1,6	3,9	3,3	0,0	2,0	4,5	5,2	0,0	2,0	3,5	4,4
	(0,1;0,15]	0,0	2,4	2,5	2,1	0,0	1,6	1,9	2,0	0,0	3,5	3,5	2,7	0,0	1,5	2,1	1,5
	(0,15;0,2]	0,0	1,3	1,7	1,7	0,0	0,8	1,3	1,5	0,0	1,2	1,4	1,6	0,0	1,3	1,5	1,4
	(0,2;0,25]	0,0	1,3	1,5	0,9	0,0	0,9	1,6	1,3	0,0	1,2	0,9	1,0	0,0	0,6	0,8	1,3
	(0,25;0,3]	0,0	1,5	1,6	1,3	0,0	1,2	2,7	1,7	0,0	1,0	1,2	0,2	0,0	0,3	2,5	1,6
	(0,3;0,4]	0,0	1,2	0,4	0,0	0,0	1,1	0,0	0,0	0,0	0,9	0,0	0,0	0,0	0,9	0,5	0,1
	(0,4;1]	0,0	1,3	0,0	0,0	0,0	2,1	0,0	0,0	0,0	0,8	0,0	0,0	0,0	2,4	0,0	0,0
		CH				CN				LC				MT			
	Exposição	ALE	MIF	PR1	PR2	ALE	MIF	PR1	PR2	ALE	MIF	PR1	PR2	ALE	MIF	PR1	PR2
TF45	0	0,0	59,1	0,0	0,0	0,0	60,2	0,0	0,0	0,0	57,1	0,0	0,0	0,0	62,4	0,0	0,0
	(0;0,02]	0,0	12,8	48,5	21,3	0,0	15,8	53,4	11,9	2,3	13,6	52,9	35,9	0,0	13,6	55,2	18,6
	(0,02;0,05]	11,6	4,9	22,8	50,4	13,2	4,4	20,4	62,1	37,9	5,2	20,2	38,7	25,8	5,1	19,8	57,8
	(0,05;0,1]	88,4	4,6	11,0	13,0	86,8	3,2	7,9	11,5	59,8	6,7	11,9	13,3	74,2	5,2	8,8	10,4
	(0,1;0,15]	0,0	3,4	4,1	4,9	0,0	2,5	4,8	3,7	0,0	4,1	4,6	3,9	0,0	1,8	3,3	3,2
	(0,15;0,2]	0,0	3,2	3,6	2,8	0,0	2,9	2,7	2,8	0,0	2,8	3,3	2,9	0,0	1,6	2,3	2,1
	(0,2;0,25]	0,0	2,0	2,6	2,1	0,0	1,9	2,4	1,2	0,0	3,7	2,2	1,9	0,0	1,8	2,3	1,5
	(0,25;0,3]	0,0	2,4	5,8	4,4	0,0	1,6	8,4	6,8	0,0	1,9	4,6	3,2	0,0	1,3	5,7	4,4
	(0,3;0,4]	0,0	3,4	1,6	1,2	0,0	1,9	0,0	0,0	0,0	2,6	0,3	0,2	0,0	2,0	2,7	2,0
	(0,4;1]	0,0	4,1	0,0	0,0	0,0	5,6	0,0	0,0	0,0	2,3	0,0	0,0	0,0	5,3	0,0	0,0

Nota. CH = Ciências Humanas; CN = Ciências da Natureza; LC = Linguagens e Códigos; MT = Matemática; ALE = aleatório; MIF = Máxima Informação de Fisher; PR1 = progressivo restrito com $k = 1$; PR2 = progressivo restrito com $k = 2$; TF20 = tamanho fixo de 20 itens; TF45 = tamanho fixo de 45 itens.

A Tabela 22 mostra a média da porcentagem de itens por intervalo de taxa de exposição nas aplicações de tamanho variável. Para os dois critérios de parada o método MIF foi o que apresentou maiores proporções de itens não administrados (chegando a 87,4% em MT com EP30RE015) e de itens superexpostos (chegando a 5,2% em MT com EP30). Assim como nas CATs de tamanho fixo, a inclusão do controle de exposição reduziu a proporção de itens não apresentados e superexpostos. A porcentagem média de itens não administrados nas quatro áreas com o método PR2 ou PR3 para ambos os critérios de parada foi zero. Adicionalmente, poucos itens tiveram taxas de exposição maiores do que 0,30 com PR2 ou PR3, porém pelo

motivo já descrito anteriormente. No que diz respeito ao parâmetro de aceleração, quanto maior seu valor, menor a proporção de itens com taxas de exposição mais elevadas, para os dois critérios de parada.

Tabela 22

Média geral das porcentagens de itens para cada intervalo de taxa de exposição nas CATs de tamanho variável

Critério de parada		CH				CN				LC				MT			
	Exposição	ALE	MIF	PR2	PR3	ALE	MIF	PR2	PR3	ALE	MIF	PR2	PR3	ALE	MIF	PR2	PR3
EP30	0	0,0	75,1	0,0	0,0	0,0	76,4	0,0	0,0	0,0	71,2	0,0	0,0	0,0	78,5	0,0	0,0
	(0;0,02]	0,0	8,2	65,7	18,8	0,0	9,1	64,5	0,8	8,3	15,2	80,1	79,8	0,0	6,7	52,4	1,9
	(0,02;0,05]	16,0	3,3	18,3	64,7	0,3	2,0	20,2	81,8	80,2	4,8	8,2	9,2	0,4	2,4	28,2	78,7
	(0,05;0,1]	84,0	3,7	7,1	8,6	99,7	3,7	5,2	8,1	11,6	3,1	5,2	5,4	96,8	1,5	5,9	6,8
	(0,1;0,15]	0,0	2,5	2,6	2,9	0,0	2,4	2,7	3,3	0,0	1,9	2,0	2,4	2,8	1,1	2,7	3,3
	(0,15;0,2]	0,0	1,9	2,5	2,2	0,0	1,9	3,2	2,4	0,0	0,8	1,5	1,0	0,0	0,9	2,4	2,0
	(0,2;0,25]	0,0	1,9	1,7	1,5	0,0	1,1	2,0	1,5	0,0	0,9	1,0	1,0	0,0	1,6	1,4	1,9
	(0,25;0,3]	0,0	0,8	2,0	1,3	0,0	0,8	2,3	2,0	0,0	0,8	1,9	1,1	0,0	2,0	5,7	4,5
	(0,3;0,4]	0,0	1,2	0,0	0,0	0,0	0,9	0,0	0,0	0,0	0,6	0,1	0,0	0,0	1,8	1,4	0,9
	(0,4;1]	0,0	1,5	0,0	0,0	0,0	1,7	0,0	0,0	0,0	0,8	0,0	0,0	0,0	3,4	0,0	0,0
		CH				CN				LC				MT			
	Exposição	ALE	MIF	PR2	PR3	ALE	MIF	PR2	PR3	ALE	MIF	PR2	PR3	ALE	MIF	PR2	PR3
EP30RE015	0	0,0	85,3	0,0	0,0	0,0	85,0	0,0	0,0	0,0	81,4	0,0	0,0	0,0	87,4	0,0	0,0
	(0;0,02]	39,2	4,0	74,3	70,0	41,8	6,0	76,0	68,5	67,0	7,8	86,3	85,1	69,4	4,0	76,0	70,6
	(0,02;0,05]	60,8	2,1	18,9	27,6	58,2	1,6	18,0	30,3	33,0	2,4	4,8	6,1	30,6	2,0	17,6	29,4
	(0,05;0,1]	0,0	2,1	5,0	2,1	0,0	1,2	3,9	1,2	0,0	2,6	3,6	4,5	0,0	1,0	5,6	0,0
	(0,1;0,15]	0,0	1,5	1,1	0,3	0,0	1,7	1,6	0,0	0,0	2,0	1,8	1,2	0,0	1,3	0,9	0,0
	(0,15;0,2]	0,0	1,2	0,5	0,0	0,0	0,5	0,5	0,0	0,0	0,7	0,8	1,1	0,0	1,1	0,0	0,0
	(0,2;0,25]	0,0	0,9	0,1	0,0	0,0	0,7	0,0	0,0	0,0	0,9	1,2	1,0	0,0	0,3	0,0	0,0
	(0,25;0,3]	0,0	1,1	0,0	0,0	0,0	0,7	0,0	0,0	0,0	0,8	1,5	0,9	0,0	0,5	0,0	0,0
	(0,3;0,4]	0,0	0,7	0,0	0,0	0,0	1,2	0,0	0,0	0,0	0,6	0,0	0,0	0,0	1,0	0,0	0,0
	(0,4;1]	0,0	1,2	0,0	0,0	0,0	1,3	0,0	0,0	0,0	0,8	0,0	0,0	0,0	1,4	0,0	0,0

Nota. CH = Ciências Humanas; CN = Ciências da Natureza; LC = Linguagens e Códigos; MT = Matemática; EP = erro padrão de medida; COR = correlação; REQM = raiz do erro quadrático médio; ALE = aleatório; MIF = Máxima Informação de Fisher; PR2 = progressivo restrito com $k = 2$; PR3 = progressivo restrito com $k = 3$; EP30 = erro padrão de 0,30; EP30RE015 = erro padrão de 0,30 ou redução do erro padrão em 0,015.

De forma geral, nas condições com o mesmo critério de parada, o método aleatório foi o que proporcionou aplicações mais seguras e o MIF, as menos seguras. A inclusão do método

de exposição PR aumentou a segurança de forma relevante, e quanto maior o parâmetro de aceleração, maior a segurança do teste.

Comparação entre PR2TF20 e linear

A combinação mais satisfatória de CAT foi o tamanho fixo de 20 itens com o método PR2 (PR2TF20). Em termos de eficiência, o tamanho médio com EP30RE015 foi menor que 16 em todas as áreas. Além disso, algumas condições de CAT de tamanho variável tiveram tamanho máximo menor que 20 itens. No entanto, essa superação não chegou a cinco itens, e em termos de precisão, as CATs com TF20 superaram as de tamanho variável, em especial as com EP30RE015. Portanto, a eficiência ligeiramente maior do critério EP30RE015 não justifica arcar com a perda na precisão em relação ao TF20. Em paralelo, o uso do controle de exposição aumentou a segurança do teste de forma relevante, sem comprometer a precisão da aplicação. Portanto, a condição adotada para se comparar com a aplicação linear do Enem foi a PR2TF20.

A Tabela 23 apresenta os indicadores de precisão da simulação das provas lineares. Adicionalmente, essa tabela mostra novamente os valores das CATs com a condição PR2TF20 para facilitar a comparação. A precisão obtida no teste linear foi satisfatória, pois a maior média de erro padrão foi 0,495 (MT), o que equivale a uma confiabilidade de 0,755, a menor média de correlação foi 0,860 (CN) e o maior valor de REQM foi 0,433 (MT). O maior valor absoluto de viés foi 0,078 (CN). No entanto, em todas as áreas os indicadores de precisão foram melhores na CAT com PR2TF20.

Tabela 23

Média do erro padrão de medida, da correlação, do viés e da raiz do erro quadrático médio das replicações das simulações com a aplicação linear e da condição PR2TF20 da CAT

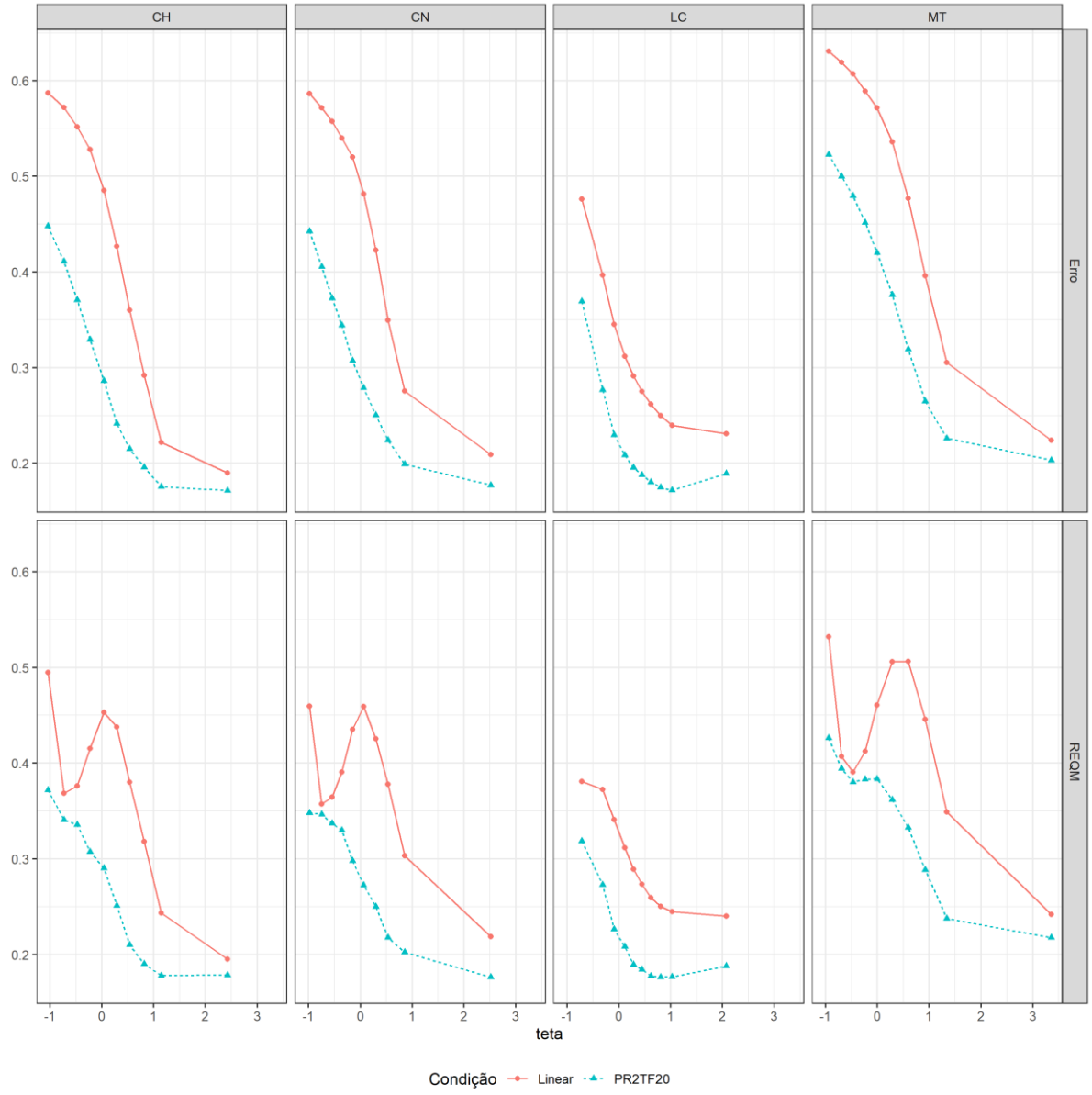
	Linear				PR2TF20			
	Erro Padrão	Correlação	Viés	REQM	Erro Padrão	Correlação	Viés	REQM
CH	0,422	0,895	-0,061	0,379	0,285	0,947	0,035	0,275
CN	0,451	0,860	-0,078	0,386	0,300	0,926	0,042	0,285
LC	0,308	0,908	-0,052	0,301	0,218	0,952	0,024	0,217
MT	0,495	0,885	-0,077	0,433	0,376	0,928	0,058	0,347

Nota. PR2TF20 = progressivo restrito com $k = 2$ e tamanho fixo de 20 itens; REQM = raiz do erro quadrático médio.

A Figura 22 mostra a REQM e o erro padrão da aplicação linear ao longo da escala do Enem, bem como da CAT com PR2TF20. Ao longo de toda a escala a REQM da CAT apresenta um ganho relevante em relação à aplicação linear. O mesmo acontece para o erro padrão.

Figura 22

Erro padrão e REQM condicionados ao teta do teste linear e da CAT com PR2TF20



Nota. PR2TF20 = progressivo restrito com $k = 2$ e tamanho fixo de 20 itens; REQM = raiz do erro quadrático médio.

Discussão

O objetivo deste trabalho foi avaliar o controle de exposição PR com diferentes parâmetros de aceleração em CATs de tamanho fixo e variável. Simulamos a aplicação do Enem (um teste educacional de alto impacto aplicado em papel) em formato de CAT. Os resultados corroboram parcialmente a hipótese H1 (a eficiência e a precisão do método PR serão semelhantes à do MIF). A eficiência do método PR foi inferior à do MIF nas CATs de tamanho variável, pois o tamanho das aplicações com o primeiro método foi maior. Já a precisão com o PR foi ligeiramente menor do que a com MIF nas CATs de tamanho fixo e na CAT com critério de parada EP30. Com o critério EP30RE015, a precisão teve um impacto negativo relevante. A hipótese H2 (a segurança do PR será maior do que a do MIF) foi corroborada. A hipótese H3 (a eficiência e a precisão do método PR serão semelhantes com todos os parâmetros de aceleração) foi parcialmente corroborada. Nas CATs de tamanho fixo e na CAT com critério de parada EP30, a mudança no parâmetro de aceleração não impactou a precisão de maneira relevante. Neste último critério, a eficiência reduziu com o aumento do parâmetro de aceleração. Com o critério EP30RE015, a eficiência não foi fortemente afetada, mas a precisão teve um impacto negativo. As hipóteses H4 (quanto maior o parâmetro de aceleração, maior a segurança) e H5 (a precisão de uma CAT com PR será maior do que a do teste linear) foram corroboradas. Especificamente sobre a H5, comparamos a precisão do teste linear com a de uma CAT de tamanho fixo de 20 itens e método PR com um parâmetro de aceleração de 2.

Nossos resultados sobre a eficiência da CAT com critério de parada EP30 contrastam com achados prévios. Neste estudo, a média de itens aplicados aumentou consideravelmente nas condições com o PR, com diferenças que chegam a 14 itens em relação ao MIF. No estudo de Leroux et al. (2013), a diferença chegou a no máximo sete itens. Já em Leroux e Dodd (2016), essa diferença não passou de três itens e em Leroux et al. (2019) a diferença na média não chegou a um item. O fato de a diferença do tamanho da aplicação ser maior neste estudo

pode ser devido às características do banco de itens e da amostra. Os estudos de Leroux e Dodd (2016) e Leroux et al. (2019) utilizaram itens politômicos, o que pode ampliar o intervalo de teta em que o teste apresenta informação total satisfatória, quando comparado a um teste de itens dicotômicos. Com isso, a administração aleatória de itens pode ter menor impacto, pois é provável que o item selecionado contribua de forma relevante para a redução do erro de medida. Essa hipótese é reforçada pelo fato de Leroux et al. (2013), que utilizaram itens dicotômicos, terem encontrado uma diferença média maior do que os demais estudos. Outra causa possível da diferença entre nossos achados e os prévios é o uso do parâmetro de aceleração. Neste estudo, quanto maior esse parâmetro, maior foi o tamanho médio da aplicação. Não utilizamos PR1 nas CATs de tamanho variável, mas nas condições com PR2, a diferença máxima para o MIF foi de nove itens, mais próxima de Leroux et al. (2013). Cabe destacar que a comparação entre os achados demanda cautela, pois a fórmula utilizada por esses estudos prévios para a seleção do item no PR é a da Equação 31, que não contém parâmetro de aceleração e se aproximaria da Equação 32 com parâmetro de aceleração 1.

Encontramos uma redução irrelevante na precisão da CAT ao incluir o controle de exposição PR em CATs de tamanho fixo e com critério de parada do erro padrão, o que corrobora achados anteriores (Barrada et al., 2008; Leroux & Dodd, 2016; Leroux et al., 2013, 2019). No entanto, o uso do método PR com o critério de parada da redução do erro teve um impacto relevante no erro de medida. Esse resultado contrasta com o obtido por Leroux et al. (2019), que utilizaram o critério de parada da redução predita do erro padrão (PSER). A diferença adotada como critério no estudo citado foi 0,020, valor menos rigoroso que o adotado neste trabalho. No entanto, no PSER, ainda que o erro padrão atinja o valor desejado, a aplicação continua até que a redução do erro atinja um valor determinado. Diferentemente, nas condições deste estudo com o critério da redução do erro, a aplicação encerrava caso o erro padrão atingisse 0,30. Por isso, um motivo possível para as diferenças dos resultados entre os

estudos é a redução da média do erro padrão causada por essa continuidade da aplicação. Portanto, recomendamos estudos que comparem a precisão dos casos em que a aplicação encerrou com erro padrão maior do que o desejado ao utilizar os dois critérios, bem como a precisão dos casos com erro padrão menor do que o desejado.

Outro possível motivo da piora na precisão ao se combinar o PR com o critério de redução do erro é um término precoce da aplicação. O peso do componente aleatório do método PR reduz mais lentamente com valores mais altos de parâmetro de aceleração. Dada a seleção aleatória de itens, podem ser administrados consecutivamente dois itens pouco informativos para o teta provisório, mesmo após o tamanho mínimo da aplicação (no caso deste trabalho, 15). Caso isso ocorra, pode haver uma baixa redução no erro. Se essa redução for menor do que a do critério de parada (no nosso caso, 0,015), a aplicação encerra precocemente. No caso do PSER, o algoritmo verifica a potencial redução do erro de medida para cada item disponível no banco, e posteriormente seleciona o item a ser administrado. Essa diferença no algoritmo pode mitigar o efeito da aleatoriedade do PR. Por isso, recomendamos estudos que verifiquem o impacto da aleatoriedade do método PR sobre o término precoce de CATs com o critério de parada da redução do erro. Ainda, sugerimos que os estudos busquem determinar tamanhos mínimos de prova ou algoritmos que evitem esse término precoce.

Nossos achados sobre a segurança do teste corroboram achados anteriores. Verificamos uma redução da sobreposição dos itens ao utilizar o método PR, e a redução aumentou quando o parâmetro de aceleração também aumentou. Esses resultados também foram encontrados por Barrada et al. (2008) em CATs de tamanho fixo. Não encontramos estudos que avaliassem o efeito do parâmetro de aceleração sobre a segurança do teste em CATs de tamanho variável, por isso recomendamos a replicação das condições de simulação utilizadas neste estudo com outros bancos de itens e sujeitos.

As diferenças nos indicadores de eficiência, precisão e segurança ao longo das quatro provas reforçam que seus valores são impactados pela distribuição dos itens e dos tetas ao longo da escala. De forma geral, os indicadores foram melhores para a prova de LC, cuja curva de informação do teste é a que mais coincide com a curva de densidade do teta da amostra. Esse achado está em acordo com o de Lee e Dodd (2012), que simularam CATs com bancos de diferentes distribuições (um com pico na região fácil da escala, outro na região média e outro na região difícil) e dois grupos de sujeitos, sendo um com média zero e outro com média 0,74. Os melhores resultados de precisão e segurança foram obtidos com o banco com itens médios, independente do grupo de sujeitos. Os piores resultados foram obtidos com o banco fácil e o grupo de sujeitos de média 0,74. Isso reforça a importância de se investigar o efeito da distribuição do banco de itens, para além do algoritmo da CAT.

A redução do tamanho da prova do Enem neste estudo para 20 itens foi superior à redução para 33 itens proposta por Spenassato et al. (2016) para a prova de MT. O aumento da eficiência neste estudo era esperado, uma vez que o banco de MT era composto por 792 itens e o do estudo citado, por 45 itens. Os autores encontraram um erro padrão médio de 0,351, ao passo que neste trabalho essa medida foi 0,376. Essa diferença pode ser devida à utilização de controle de exposição, pois neste trabalho a média do erro padrão em MT na condição com MIF foi 0,311. Ou seja, provavelmente o tamanho expressivamente maior do nosso banco de itens compensou o impacto da redução do tamanho da prova e do controle de exposição com PR2TF20, pois a diferença no erro em relação ao estudo prévio foi pequena. Apesar de nossos valores de REQM e correlação (0,347 e 0,928) terem ficado menos satisfatórios do que os do trabalho anterior (0,088 e 0,998), isso pode ser explicado pelo fato de os autores terem considerado como o teta real aquele calculado com os 45 itens do banco. Como foram aplicados 33 itens, faz sentido que o teta estimado na CAT seja muito próximo do teta real. Neste trabalho,

em vez de calcular o teta considerado real a partir das respostas, simulamos as respostas a partir do teta considerado real.

A redução observada neste estudo foi parecida com a observada por Jatobá et al. (2020), que também utilizaram os 45 itens da prova de MT do Enem 2012. Nesse estudo, os autores utilizaram um método de seleção de item personalizado, o que permitiu a redução da aplicação a 21 itens, enquanto a aplicação com MIF reduziu a 35 itens. Neste estudo, chegamos a um tamanho de 20 itens com um método de seleção de itens universal, portanto estudos futuros devem verificar a potencialidade de outros métodos de seleção para aumentar a eficiência da CAT do Enem.

Nossa redução também superou a de Kalender e Berberoglu (2017). A CAT de tamanho fixo de 25 itens desse estudo teve médias de erro padrão variando de 0,25 a 0,32. Neste trabalho, as CATs de 20 itens com MIF tiveram erro padrão médio de até 0,311. O banco dos autores tinha 45 itens, portanto essa superação era esperada. Na CAT que elegemos como a mais satisfatória, que controla a exposição dos itens, o erro padrão médio variou de 0,218 a 0,376. Trata-se de uma perda irrelevante na precisão, dado o ganho na segurança de um teste de alto impacto.

Este estudo tem como limitação o fato de ter utilizado somente bancos com respostas simuladas. Por isso, as respostas não tiveram possíveis interferências externas (como fadiga, motivação e conhecimento prévio do participante sobre um item). Recomendamos estudos que comparem a precisão e eficiência de CATs com provas lineares em contextos de alto impacto. Apontamos para a possibilidade de se desenvolver uma CAT eficiente para o Enem que melhore sua precisão e segurança. Considerando as quatro provas, a CAT eleita neste estudo totaliza 80 itens, quantidade menor do que o total de itens em um dia de aplicação do exame. Esperamos contribuir para otimizar testes educacionais de alto impacto, em especial o Enem, de modo que ele se torne mais justo ao eliminar interferências indesejáveis no processo de medição.

Referências

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 493–513. 10.1348/000711007X230937

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, (49), 61–80.

Chen, S.-Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129–145.

Jatobá, V. M. G., Farias, J. S., Freire, V., Ruela, A. S., & Delgado, K. V. (2020). ALICAT: A customized approach to item selection process in computerized adaptive testing. *Journal of the Brazilian Computer Society*, *26*(1), 4. 10.1186/s13173-020-00098-z

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey?. *Educational Sciences: Theory & Practice*, *17*(2), 573–596. http://doi.org/10.12738/estp.2017.2.0280

Kallen, M. A., Cook, K. F., Amtmann, D., Knowlton, E., & Gershon, R. C. (2018). Grooming a CAT: Customizing CAT administration rules to increase response efficiency in specific research and clinical settings. *Quality of Life Research*, *27*(9), 2403–2413. 10.1007/s11136-018-1870-z

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement*, *72*(1), 159–175. 10.1177/0013164411411296

Leroux, A. J., & Dodd, B. G. (2016). A comparison of exposure control procedures in CATs using the GPC model. *The Journal of Experimental Education*, *84*(4), 666–685. 10.1080/00220973.2015.1099511

Leroux, A. J., Lopez, M., Hembry, I., & Dodd, B. G. (2013). A comparison of exposure control procedures in CATs using the 3PL model. *Educational and Psychological Measurement*, *73*(5), 857–874. 10.1177/0013164413486802

Leroux, A. J., Waid-Ebbs, J. K., Wen, P.-S., Helmer, D. A., Graham, D. P., O'Connor, M. K., & Ray, K. (2019). An investigation of exposure control methods with variable-length CAT using the partial credit model. *Applied Psychological Measurement*, *43*(8), 624–638. 10.1177/0146621618824856

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2000). Content balancing in stratified computerized adaptive testing designs. *Annual Meeting of the American Educational Research Association*, New Orleans. Retirado de https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1058.3442&rep=rep1&type=pdf

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: recent updates of the package catR. *Journal of Statistical Software*, *76*(Code Snippet 1). 10.18637/jss.v076.c01

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8). 10.18637/jss.v048.i08

McClarty, K. L., Sperling, R. A., & Dodd, B. G. (2006). A variant of the progressive-restricted item exposure control procedure in computerized adaptive testing systems based on the 3PL and partial credit models. *Annual Meeting of the American Educational Research Association*, San Francisco

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, *36*(1), 101–123. 10.1177/0265532217725776

Morris, S. B., Bass, M., Howard, E., & Neapolitan, R. E. (2020). Stopping rules for computer adaptive testing when item banks have nonuniform information. *International Journal of Testing*, *20*(2), 146–168. 10.1080/15305058.2019.1635604

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4), 311–327.

Spenassato, D., Trierweiller, A. C., Andrade, D. F. de, & Bornia, A. C. (2016). Testes Adaptativos Computadorizados Aplicados em Avaliações Educacionais. *Revista Brasileira de Informática na Educação*, *24*(02), 1. http://doi.org/10.5753/rbie.2016.24.02.1

Sulak, S., & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *10*(3), 315–326. 10.21031/epod.530528

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

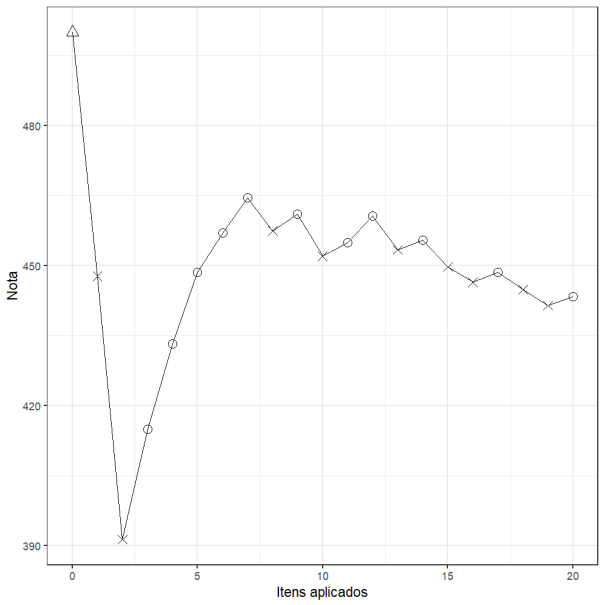
Produto 2 – Publicação da CAT Enem

Uma das etapas do desenvolvimento de uma CAT é a sua publicação, para que ela de fato possa ser aplicada. Nesta tese, desenvolvemos uma versão do Enem em formato de CAT. Nas etapas anteriores deste trabalho, verificamos a qualidade do banco de itens do Enem por meio da avaliação da sua calibração e determinamos as especificações finais para a CAT Enem (início da aplicação, método de seleção de itens e critério de parada). Nesta seção da tese, descrevemos a publicação e a disponibilização da CAT Enem.

A CAT Enem foi desenvolvida em ambiente R de programação com o pacote golem (Fay et al., 2022), que otimiza o desenvolvimento de aplicações no pacote shiny (Chang et al., 2021). A aplicação conta com uma página inicial com instruções. Nela, é possível escolher dentre as quatro provas do Enem, incluindo a língua estrangeira. Ao final, o programa retorna a nota do sujeito e um gráfico que mostra a variação da sua nota ao longo da aplicação, como o da Figura 23.

Figura 23

Ilustração de gráfico apresentado ao final da aplicação da CAT Enem



As especificações do algoritmo da CAT foram modificadas a partir das obtidas no Artigo 2 desta tese e são apresentadas na Tabela 24. A diferença está no método de seleção de

itens e no critério de parada. O método de seleção é o progressivo e não o progressivo restrito, pois para restringir a taxa de exposição dos itens a 30%, seria necessário armazenar as respostas dos participantes em um servidor, o que complexificaria a aplicação desnecessariamente. Uma vez que a aplicação possui finalidades pedagógicas, a superexposição dos itens não se torna uma questão. Ainda assim, adotamos o método progressivo para reduzir a probabilidade de um sujeito responder o mesmo item em uma eventual segunda participação. Como critério de parada, adotamos o tamanho fixo de 20 itens e um limite de tempo de 70 minutos. Esse tempo equivale a 3,50 minutos em média para cada item, o que é aproximadamente a proporção de tempo na aplicação oficial. No segundo dia prova, cada participante tem até cinco horas para responder os 180 itens, o que equivale a 3,33 minutos por item em média.

Tabela 24

Especificações do algoritmo da CAT Enem

Especificação	Descrição
Início da aplicação	Item selecionado aleatoriamente
Método de seleção dos itens	Progressivo com parâmetro de aceleração 2
Método de balanceamento do conteúdo	Modified constrained CAT
Método de estimação do teta	Expected a Posteriori
Critério de parada	Tamanho fixo de 20 itens ou 70 minutos

A aplicação da CAT Enem é gratuita, de uso livre e de código aberto. Ela está disponível para baixar em http://github.com/alexandrejaloto/CATEnem. Para utilizar em computador próprio, basta utilizar os seguintes comandos:

```
devtools::install_github('alexandrejaloto/CATEnem')
CATEnem::run_app()
```

Para utilizar a CAT Enem diretamente do navegador, basta acessar http://jalotoalexandre.shinyapps.io/CATEnem. Encorajamos o seu uso, sua divulgação e as sugestões de aprimoramento.

Considerações finais

Esta tese teve como objetivo desenvolver uma CAT do Enem mais eficiente, precisa e segura do que o formato atual do exame. Para atingir a este objetivo, foi necessário: comparar diferentes distribuições amostrais na calibração de itens no modelo 3PL; desenvolver um pacote de simulação de CAT; avaliar o controle de exposição PR com diferentes parâmetros de aceleração em uma CAT em termos de eficiência, precisão e segurança; e publicar a CAT Enem com as especificações selecionadas a partir de simulações. De forma geral, este trabalho contribuiu para o processo de reflexão do desenvolvimento de uma CAT para testes educacionais em larga escala e de alto impacto. Especificamente, contribuímos para avançar na reflexão sobre a implementação de um formato de CAT para o Enem.

Elaboramos sete hipóteses, algumas delas rejeitadas e outras corroboradas (parcial ou integralmente). A hipótese H1 (a amostra com distribuição retangular de acertos retorna parâmetros mais semelhantes aos parâmetros reais) foi rejeitada, pois não houve diferença significativa entre os tipos de amostra para o parâmetro de discriminação. Para o parâmetro de dificuldade, a amostra deslocada retornou melhores valores em CH e CN e para o parâmetro de pseudochute, as amostras deslocada e aleatória se mostraram melhores. Apesar dessas diferenças, os resultados não apontaram para a prevalência de um tipo de amostra para calibrar os itens do Enem 2020. A hipótese H2 (a amostra estratificada pelo número de acertos – deslocada – retorna parâmetros aceitáveis) do estudo foi corroborada.

A hipótese H3 (a eficiência e a precisão do método PR serão semelhantes à do MIF) foi parcialmente corroborada. A eficiência do método PR foi inferior à do MIF nas CATs de tamanho variável. Já a precisão com o PR foi ligeiramente menor do que a com MIF nas CATs de tamanho fixo. Nas CATs de tamanho variável, essa precisão também foi ligeiramente menor com critério de parada EP30. Já com o critério EP30RE015, a precisão teve um impacto negativo relevante. A hipótese H4 (a segurança do PR será maior do que a do MIF) foi

corroborada. A hipótese H5 (a eficiência e a precisão do método PR serão semelhantes com todos os parâmetros de aceleração) foi parcialmente corroborada. O aumento do parâmetro de aceleração não impactou a precisão de maneira relevante nas CATs de tamanho fixo e na CAT com critério de parada EP30, porém a eficiência reduziu neste último critério. Com o critério EP30RE015, a precisão reduziu com o aumento do parâmetro de aceleração e a eficiência não teve impacto relevante. As hipóteses H6 (quanto maior o parâmetro de aceleração, maior a segurança) e H7 (a precisão de uma CAT com PR será maior do que a do teste linear) foram corroboradas.

No que diz respeito à comparação entre a CAT e o teste linear, utilizamos a CAT de tamanho fixo de 20 itens e método PR com um parâmetro de aceleração de 2. A CAT com essas especificações reduziu o tamanho dos testes do Enem a menos da metade e se mostrou adequada em termos de precisão e segurança para ser aplicada à amostra deste estudo. Portanto, após avaliar a calibração do banco de itens de nossa CAT, realizar simulações com o pacote estatístico desenvolvido e determinar as especificações para a publicação da CAT Enem, nossa pergunta de pesquisa “é possível reduzir o tamanho do Enem e melhorar sua precisão e segurança com uma CAT?” foi respondida positivamente.

Este trabalho avança ao avaliar a qualidade da calibração dos itens do Enem, uma investigação necessária tanto academicamente (dada a ausência de estudos com o desenho amostral do Enem) quanto socialmente, pois o Inep ainda não possui uma publicação que avalie o desenho utilizado na calibração dos itens do Enem. Demonstramos que o desenho amostral adotado pelo Inep retornou valores aceitáveis de parâmetros na calibração dos itens do Enem 2020. Esperamos contribuir para aprimorar os desenhos amostrais para calibrar itens de testes educacionais. Estudos futuros sobre calibração devem incluir o sorteio de amostras estratificadas pelo teta calculado a partir de uma calibração provisória. Ainda, recomendamos

que se investigue o impacto dos graus de justaposição entre a população e a curva de informação do teste sobre a calibração dos itens.

Este trabalho também avança ao desenvolver um pacote de simulação de CAT em R com funcionalidades inexistentes em outros pacotes. O pacote simCAT se diferencia dos demais pacotes estatísticos que simulam aplicações em CAT por trazer algumas funcionalidades não incluídas por eles. Esperamos que este produto seja utilizado por pessoas interessadas em desenvolver suas próprias CATs que busquem realizar simulações do tipo Monte Carlo, *post-hoc* ou híbridas. Os avanços futuros possíveis desse pacote incluem a possibilidade de se utilizar itens politômicos no teste, a publicação de uma CAT após se determinar o seu algoritmo e uma interface amigável para as pessoas que não dominam a linguagem R.

Outro avanço deste trabalho foi preencher lacunas relacionadas ao entendimento do funcionamento do método PR e seu parâmetro de aceleração em CATs de tamanho variável. Ao utilizar o critério de parada do erro padrão, esse método funcionou adequadamente e não impactou a precisão e a eficiência de maneira relevante, quando comparado ao MIF. No entanto, com o critério da redução do erro, esse método não funcionou adequadamente. Portanto, recomendamos novos estudos que investiguem o motivo da piora da precisão ao se combinar o PR com esse critério de parada e que busquem desenvolver especificações de CAT que superem essa limitação.

Nosso trabalho tem o potencial de contribuir para práticas pedagógicas que envolvem o uso de provas pretéritas do Enem ao disponibilizar uma aplicação em formato de CAT do exame. Nossa aplicação utiliza itens do Enem e fornece uma devolutiva com a nota calculada na mesma escala do exame, por isso pode ser utilizada como um exercício preparatório para a aplicação oficial. Apesar desse potencial pedagógico, cabe destacar que não recomendamos o uso da nota na CAT Enem como se fosse uma réplica fiel do que o sujeito obteria caso

participasse da aplicação oficial do Enem. Para podermos ter essa equivalência, o ambiente deveria ser controlado de tal maneira que replicasse uma situação de aplicação oficial. Além disso, precisaríamos ter evidências de que a nota na CAT Enem possui a mesma interpretação que a nota no Enem.

Nesse sentido, recomendamos estudos de validade que produzam evidências para as possíveis interpretações e usos da nota na CAT Enem. Este trabalho não realizou estudos de validade e uma agenda para implementar uma CAT inclui fundamentalmente estudos dessa natureza. Exemplos de estudos incluem: verificar a associação entre a nota na CAT Enem e a nota no Enem no formato tradicional; verificar a capacidade da CAT Enem de predizer o desempenho no ensino superior; verificar a estrutura interna da CAT Enem incluindo o efeito de outros fatores, como a fadiga; determinar a amostra de conteúdos que devem estar presentes no teste.

Além dos estudos de validade, uma agenda para implementação da CAT Enem deve incluir ações e estudos que subsidiam as decisões sobre os elementos da CAT, incluindo a composição do banco de itens. Em nossa CAT Enem, o erro padrão de medida não é o mesmo para todas as faixas de teta, o que configura uma limitação, pois quando o teste objetiva selecionar, a estimativa dos tetas deve possuir nível de precisão fixo para que não haja injustiças (Labarrère et al., 2011). Por isso, este trabalho aponta para a necessidade de se elaborar itens com dificuldade nos intervalos da escala com menor informação psicométrica, notadamente as regiões mais fáceis.

No que se refere às especificações do algoritmo da CAT Enem, ainda que este trabalho tenha mostrado a potencialidade do método MIF com ou sem controle de exposição, estudos sobre a seleção de itens devem incluir outros métodos, sejam universais ou personalizados como o ALICAT (Jatobá et al., 2020). Para além do método de seleção em si, outros formatos de

CAT também devem ser explorados, como *Multistage* CAT (Zenisky et al., 2010) e *shadow test* (van der Linden, 2002).

Para a estimação do teta, recomendamos estudos que verifiquem o impacto da distribuição prévia na estimação. Sabe-se que adotar valores próximos da distribuição da amostra reduz o viés da estimação (Cúri & Silva, 2019), no entanto é preciso verificar o impacto da mudança desses valores sobre a qualidade da linkagem dos itens e a equalização, o que pode ter impacto sobre a comparabilidade ao longo dos anos de aplicação.

Ainda no que concerne às especificações da CAT, recomendamos estudos que investiguem critérios de parada mais eficientes que não comprometam a precisão do teste, como tamanho fixo com menos itens. Outra possibilidade é investigar outros critérios de parada relacionados ao erro padrão, como o da mínima informação do banco. No entanto, cabe destacar que adotar uma CAT de tamanho variável em um teste de alto impacto como o Enem pode gerar questionamentos por parte da sociedade quanto à isonomia, pois os sujeitos responderão a testes de tamanhos diferentes. Portanto, caso essa decisão seja tomada, o Inep deve prover documentos que evidenciem a isonomia da prova.

Outro ponto não abordado neste trabalho diz respeito à primeira etapa de um plano de desenvolvimento da CAT, qual seja, a execução de estudos de viabilidade, aplicabilidade e planejamento. Uma agenda de implementação da CAT Enem também deve incluir estudos dessa natureza, por exemplo verificar a variação do custo relacionada à redução do tamanho da prova, à mudança na logística do exame e à elaboração dos novos itens do banco.

Consideramos que a grande limitação deste trabalho, e que perpassou seus dois artigos, é o fato de utilizar somente respostas simuladas. Por isso, reforçamos a necessidade de estudos com respostas reais que contemplem os pontos elencados nos parágrafos acima.

Destacamos uma contribuição indireta deste trabalho, que é a facilitação do acesso aos microdados do Enem, na medida em que descrevemos como operar com informações

complexas contidas nos seus arquivos. Por exemplo, apresentamos as transformações necessárias nas escalas para posicionar os itens e as pessoas na mesma métrica. Esperamos com isso incentivar outras pesquisas robustas com dados oficiais do Inep. Por último, esperamos que este trabalho possa contribuir para o aprimoramento do Enem e a ampliação do acesso ao conhecimento de CAT no Brasil.

Referências

Aytuğ Koşan, A. M., Koç, N., Elhan, A. H., & Öztuna, D. (2019). Developing an item bank for Progress Tests and application of computerized adaptive testing by simulation in medical education. *International Journal of Assessment Tools in Education*, *6*(4), 656–669. 10.21449/ijate.635675

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement?. *Journal of Computerized Adaptive Testing*, *1*(1), 1–18. 10.7333/1212-0101001

Barichello, L., Guimarães, R. S., & Figueiredo Filho, D. B. (2022). A formatação da prova afeta o desempenho dos estudantes? Evidências do Enem (2016). *Educação e Pesquisa*, *48*, e241713. 10.1590/s1678-4634202248241713por

Barrada, J. R., Abad, F. J., & Veldkamp, B. P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21*(2), 313–320.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 493–513. 10.1348/000711007X230937

Bekman, R. M. (2001). Aplicação dos blocos incompletos balanceados na teoria de resposta ao item. *Estudos em Avaliação Educacional*, (24), 119–138.

Portaria n. 438, de 28 de maio de 1998. http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?jornal=1&pagina=5&data=01/06/1998

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, (49), 61–80.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. http://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–39.

Chang, H.-H., Qian, J., & Ying, Z. (2001). A-Stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, *25*(4), 333–341. 10.1177/01466210122032181

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229. 10.1177/014662169602000303

Chang, H.-H., & Ying, Z. (1999). A-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222. 10.1177/01466219922031338

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny Web application framework for R*. https://CRAN.R-project.org/package=shiny

Chen, S.-Y., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, *41*(2), 149–174.

Choi, S. W. (2020). *Firestar Computerized Adaptive Testing (CAT) simulation program*. https://github.com/choi-phd/Firestar

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*(1), 37–53. http://doi.org/10.1177/0013164410387338

Costa, D. R., Karino, C. A., Moura, F. A. S., & Andrade, D. F. de. (2009). A comparison of three methods of item selection for computerized adaptive testing. In D. J. Weiss (Org.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. http://www.iacat.org/sites/default/files/biblio/cat09costa.pdf

Cúri, M., & Silva, V. (2019). Academic english proficiency assessment using a computerized adaptive test. *Tendencias em Matemática Aplicada e Computacional*, *20*(2), 381–401. https://doi.org/10.5540/tema.2019.020.02.0381

von Davier, A. A. (2011). A statistical perspective on equating test scores. In A. A. von Davier (Org.). *Statistical Models for Test Equating, Scaling, and Linking* (pp. 1–20). Springer New York. 10.1007/978-0-387-98138-3

Domingue, B., Kanopka, K., Stenhaug, B., Sulik, M., Beverly, T., Brinkhuis, M. J. S., Circi, R., Faul, J., Liao, D., McCandliss, B., Obradovic, J., Piech, C., Porter, T., Soland, J., Weeks, J., Wise, S., & Yeatman, J. D. (2020). Speed accuracy tradeoff? Not so fast: Marginal changes in speed have inconsistent relationships with accuracy in real-world settings. *PsyArXiv*. http://doi.org/10.31234/osf.io/kduv5

Fay, C., Guyader, V., Rochette, S., & Girard, C. (2022). *Golem A framework for robust shiny applications*. https://CRAN.R-project.org/package=golem

Ferreira-Rodrigues, C. F. (2015). *Estudos com o Enem a partir de uma abordagem psicométrica da inteligência* (Tese de Doutorado, Universidade São Francisco, Itatiba, SP). https://www.usf.edu.br/galeria/getImage/427/2977366806369866.pdf

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The*

Journal of Technology, Learning, and Assessment, *5*(8). https://ejournals.bc.edu/index.php/jtla/article/view/1647

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model, *17*(1). University of Massachusetts Amherst. 10.7275/F0GZ-KC87

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Orgs.). *Computerized adaptive testing From inquiry to operation*. American Psychological Association. http://content.apa.org/books/10244-000

Huang, H.-Y., Chen, P.-H., & Wang, W.-C. (2012). Computerized adaptive testing using a class of high-order item response theory models. *Applied Psychological Measurement*, *36*(8), 689–706. 10.1177/0146621612459552

Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection, *17*(12). University of Massachusetts Amherst. 10.7275/NR1C-YV82

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2009a). *Exame Nacional do Ensino Médio (ENEM): Textos teóricos e metodológicos*. MEC/INEP

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2009b). *Matriz de referência ENEM*. MEC/INEP. http://download.inep.gov.br/download/enem/matriz_referencia.pdf

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2012). *Entenda a sua nota no Enem: Guia do participante*. MEC/INEP. http://download.inep.gov.br/educacao_basica/enem/guia_participante/2013/guia_do_participante_notas.pdf

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2021). *Exame Nacional do Ensino Médio—Enem Procedimentos de análise*. Inep.

https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/enem_procedimentos_de_analise.pdf

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2023a). *Encceja*. https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/encceja

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2023b). *Microdados*. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2023c). *Exame Nacional do Ensino Médio (Enem)*. https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

Jaloto, A. (2018). É possível reduzir o número de questões do Enem por meio de uma testagem adaptativa computadorizada? *III Seminário Internacional de Estatística com R*, Niterói. https://periodicos.uff.br/anaisdoser/article/view/29248

Jaloto, A., & Primi, R. (2021a). Using CAT to reduce test length of the Brazilian National High School Exam. *ITC 2021 Colloquium on Tests and Testing*,

Jaloto, A., & Primi, R. (2021b). E se o Enem fosse aplicado na forma de uma testagem adaptativa computadorizada? *XI Reunião da Abave e VII Conbratri*,

Jaloto, A., & Primi, R. (2022). Can we improve Brazilian High School Exam with CAT? *8th Conference of the International Association for Computerized Adaptive Testing*, Frankfurt am Main

Jaloto, A., & Primi, R. (2023a). *SimCAT. Computerized adaptive testing simulations*. https://github.com/alexandrejaloto/simCAT

Jaloto, A., & Primi, R. (2023b). *CATEnem*. https://github.com/alexandrejaloto/CATEnem

Jaloto, A., & Primi, R. (2023c). Next-generation Enem assessment with fewer items and high reliability using CAT. *SciELO Preprints*. https://doi.org/10.1590/SciELOPreprints.5339

Jatobá, V. M. G., Farias, J. S., Freire, V., Ruela, A. S., & Delgado, K. V. (2020). ALICAT: A customized approach to item selection process in computerized adaptive testing. *Journal of the Brazilian Computer Society*, *26*(1), 4. 10.1186/s13173-020-00098-z

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey?. *Educational Sciences: Theory & Practice*, *17*(2), 573–596. http://doi.org/10.12738/estp.2017.2.0280

Kallen, M. A., Cook, K. F., Amtmann, D., Knowlton, E., & Gershon, R. C. (2018). Grooming a CAT: Customizing CAT administration rules to increase response efficiency in specific research and clinical settings. *Quality of Life Research*, *27*(9), 2403–2413. 10.1007/s11136-018-1870-z

Karino, C. A., Costa, D. R., & Laros, J. A. (2009). Adequacy of an item pool measuring proficiency in English language to implement a CAT procedure. In D. J. Weiss (Org.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. http://www.iacat.org/sites/default/files/biblio/cat09karino.pdf

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359–375. 10.1207/s15324818ame0204_6

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. (3º ed.). Springer

Labarrère, J. G., Silva, C. Q. da, & Costa, D. R. (2011). Testes adaptativos computadorizados. *Revista Brasileira de Biometria*, *29*(2), 229–261.

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement*, *72*(1), 159–175. 10.1177/0013164411411296

Leroux, A. J., & Dodd, B. G. (2016). A comparison of exposure control procedures in CATs using the GPC model. *The Journal of Experimental Education*, *84*(4), 666–685. 10.1080/00220973.2015.1099511

Leroux, A. J., Lopez, M., Hembry, I., & Dodd, B. G. (2013). A comparison of exposure control procedures in CATs using the 3PL model. *Educational and Psychological Measurement*, *73*(5), 857–874. 10.1177/0013164413486802

Leroux, A. J., Waid-Ebbs, J. K., Wen, P.-S., Helmer, D. A., Graham, D. P., O'Connor, M. K., & Ray, K. (2019). An investigation of exposure control methods with variable-length CAT using the partial credit model. *Applied Psychological Measurement*, *43*(8), 624–638. 10.1177/0146621618824856

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2000). Content balancing in stratified computerized adaptive testing designs. *Annual Meeting of the American Educational Research Association*, New Orleans. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1058.3442&rep=rep1&type=pdf

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement*, *26*(4), 376–392. 10.1177/014662102237795

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, *2*(5). https://ejournals.bc.edu/index.php/jtla/article/view/1665

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201–216. 10.1007/BF02294775

van der Linden, W. J. (2002). Constrained adaptive testing with shadow tests. In C. A. W. Glas, & W. J. van der Linden (Orgs.). *Computerized adaptive testing Theory and practice* (pp. 27–52). Kluwer Academic Publishers

Luijten, M., Schalet, B., Roorda, L., Grootenhuis, M., Haverman, L., & Terwee, C. (2021). Optimizing the efficiency of computerized adaptive tests using real data: A machine learning approach. *Journal of PatientReported Outcomes*, *5*(SUPPL 1), O38. https://doi.org/10.1186/s41687-021-00349-3

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: recent updates of the package catR. *Journal of Statistical Software*, *76*(Code Snippet 1). 10.18637/jss.v076.c01

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8). 10.18637/jss.v048.i08

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R*. Springer International Publishing. 10.1007/978-3-319-69218-0

McBride, J. R. (1976). Bandwidth, fidelity, and adaptive tests. *CAT/C 21975 The second conference on computer-assisted test construction*. Atlanta Public Shools

McBride, J. R., & Martin, J. T. (2014). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Org.). *New horizon testing Latent trait test theory and computerized adaptive testing* (pp. 223–236). Elsevier Science

McClarty, K. L., Sperling, R. A., & Dodd, B. G. (2006). A variant of the progressive-restricted item exposure control procedure in computerized adaptive testing systems based on the

3PL and partial credit models. *Annual Meeting of the American Educational Research Association*, San Francisco

Ministério da Educação. (2009). http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=768-proposta-novovestibular1-pdf&category_slug=documentos-pdf&Itemid=30192

Ministério da Educação. (2023). *SiSU - Dados abertos*. https://dadosabertos.mec.gov.br/sisu

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, *36*(1), 101–123. 10.1177/0265532217725776

Moreira, G. de O. (2017). *Validade preditiva do processo seletivo para admissão em medicina e o papel das ações afirmativas em relação ao desempenho durante a graduação e na seleção para a residência médica* (Universidade Estadual de Campinas, Campinas). https://repositorio.unicamp.br/acervo/detalhe/990235

Morris, S. B., Bass, M., Howard, E., & Neapolitan, R. E. (2020). Stopping rules for computer adaptive testing when item banks have nonuniform information. *International Journal of Testing*, *20*(2), 146–168. 10.1080/15305058.2019.1635604

Nicewander, W. A., & Thomasson, G. L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement*, *23*(3), 239–247. https://doi.org/10.1177/01466219922031356

Nunes, C. H. S. da S., & Primi, R. (2005). Impacto do tamanho da amostra na calibração de itens e estimativa de escores por teoria de resposta ao item. *Avaliação Psicológica*, *4*(2), 141–153.

Nydick, S. W. (2014). *CatIrt Aan R package for simulating IRT-based computerized adaptive tests*. https://CRAN.R-project.org/package=catIrt

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Org.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. http://iacat.org/sites/default/files/biblio/cat09nydick.pdf

Owen, R. J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*(350), 351–356. 10.1080/01621459.1975.10479871

Paek, I., Liang, X., & Lin, Z. (2021). Regarding item parameter invariance for the Rasch and the 2-parameter logistic models: An investigation under finite non-representative sample calibrations. *Measurement: Interdisciplinary Research and Perspectives*, *19*(1), 39–54. 10.1080/15366367.2020.1754703

Peres, A. J. de S. (2019). Testagem adaptativa por computador (CAT): Aspectos conceituais e um panorama da produção brasileira. *Revista Examen*, *3*(3), 66–86.

Primi, R. (2006a). Evidências de validade das provas do ENADE 2004. In D. Ristoff, A. Limana, & M. R. F. de Brito (Orgs.). *Enade Perspectiva de avaliação dinâmica e análise de mudanças* (pp. 59–73). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. https://download.inep.gov.br/publicacoes/diversas/temas_da_educacao_superior/enade_perspectiva_de_avaliacao_dinamica_e_analise_de_mudanca.pdf

Primi, R. (2006b). *A validade do ENADE para avaliação da qualidade dos cursos de instituições de ensino superior* (Projeto de pesquisa aprovado no edital INEP/CAPES nº 001/2006)

Primi, R., Santos, A. A. A. dos, Vendramini, C. M., Taxa, F., Muller, F. A., Maria de Fátima Lukjanenko, & Isabel Silva Sampaio. (2001). Competências e habilidades cognitivas:

Diferentes definições dos mesmos construtos. *Psicologia: Teoria e Pesquisa*, *17*(2), 151–159. http://dx.doi.org/10.1590/S0102-37722001000200007

Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 176–186. http://doi.org/10.1037/aca0000230

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4), 311–327.

Revuelta, J., Ponsoda, V., & Olea, J. (1998). Métodos para el control de las tasas de exposición en tests adaptativos informatizados. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, *4*(2), preprint 4.

Rudner, L. M., & Guo, F. (2011). Computer adaptive testing for small scale programs and instructional systems. *Journal of Applied Testing Technology*, *12*(1). http://jattjournal.net/index.php/atp/article/view/48363

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*(1), 321–335. http://doi.org/10.12738/estp.2017.1.0270

Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory & Practice*. 10.12738/estp.2015.6.0102

Seo, D. G., & Choi, J. (2018). Post-hoc simulation study of computerized adaptive testing for the Korean Medical Licensing Examination. *Journal of Educational Evaluation for Health Professions*, *15*, 14. 10.3352/jeehp.2018.15.14

Souza, A. M. de, Vendramini, C. M. M., & Silva, M. C. R. da. (2013). Validade preditiva de um processo seletivo em relação ao desempenho de universitários de psicologia. *Encontro: Revista de Psicologia*, *16*(24), 55–68.

Spenassato, D., Trierweiller, A. C., Andrade, D. F. de, & Bornia, A. C. (2016). Testes Adaptativos Computadorizados Aplicados em Avaliações Educacionais. *Revista Brasileira de Informática na Educação*, *24*(02), 1. http://doi.org/10.5753/rbie.2016.24.02.1

Stafford, R. E., Runyon, C. R., Casabianca, J. M., & Dodd, B. G. (2019). Comparing computer adaptive testing stopping rules under the generalized partial-credit model. *Behavior Research Methods*, *51*(3), 1305–1320. 10.3758/s13428-018-1068-x

Sulak, S., & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *10*(3), 315–326. 10.21031/epod.530528

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1). https://doi.org/10.7275/wqzt-9427

Travitzki, R., Alavarse, O. M., De Rizzo Meneghetti, D., & De Toledo Catalani, É. M. (2021). Teste adaptativo informatizado da Provinha Brasil – Leitura: Resultados e perspectivas. *Estudos em Avaliação Educacional*, *31*(78), 1–29. 10.18222/eae.v0ix.7216

Tsaousis, I., Sideridis, G. D., & AlGhamdi, H. M. (2021). Evaluating a computerized adaptive testing version of a Cognitive Ability Test using a simulation study. *Journal of Psychoeducational Assessment*, *39*(8), 954–968. 10.1177/07342829211027753

Urry, V. W. (1970). *A Monte Carlo investigation of logistic mental test models* (Tese de doutorado, Universidade de Purdue, West Lafayette, Indiana, EUA). https://www.proquest.com/docview/302519686

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*(2), 203. 10.2307/1165378

Veldkamp, B. P., & Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: Avaliação e Políticas Públicas em Educação*, *21*(78), 57–72. https://doi.org/10.1590/S0104-40362013005000001

Veldkamp, B. P., Verschoor, A. J., & Eggen, T. J. H. M. (2010). A multiple objective test assembly approach for exposure control problems in computerized adaptive testing. *Psicológica*, *31*(2), 335–355.

Wang, C., Weiss, D. J., & Shang, Z. (2019). Variable-Length Stopping Rules for Multidimensional Computerized Adaptive Testing. *Psychometrika*, *84*(3), 749–771. 10.1007/s11336-018-9644-7

Wang, Z., Wang, C., & Weiss, D. J. (2022). Termination criteria for grid multiclassification adaptive testing with multidimensional polytomous Items. *Applied Psychological Measurement*, *46*(7), 014662162211083. 10.1177/01466216221108383

Weiss, D. J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological Measurement*, *6*(4), 473–492. 10.1177/014662168200600408

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27. https://doi.org/10.2458/v2i1.12351

Weiss, D. J., & Guyer, R. (2012). *Manual for CATSim Comprehensive simulation of computerized adaptive testing*. Assessment Systems Corporation. http://www.iacat.org/sites/default/files/biblio/CATSIM%20Manual.pdf

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*(3), 347–364. 10.1177/014662168400800312

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press. https://www.rasch.org/BTD_RSA/pdf%20[reduced%20size]/Best%20Test%20Design.pdf

Yasuda, J., Hull, M. M., & Mae, N. (2022). Improving test security and efficiency of computerized adaptive testing for the Force Concept Inventory. *Physical Review Physics Education Research*, *18*(1), 010112. 10.1103/PhysRevPhysEducRes.18.010112

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden, & C. A. W. Glas (Orgs.). *Elements of adaptive testing* (pp. 355–372). Springer New York. http://link.springer.com/10.1007/978-0-387-85461-8