https://doi.org/10.18222/eae.v35.10142_port
DOI do *preprint*: https://doi.org/10.1590/SciELOPreprints.5339

ENEM DE PRÓXIMA GERAÇÃO COM MENOS ITENS E ALTA CONFIABILIDADE USANDO CAT

ALEXANDRE JALOTO[I]
RICARDO PRIMI[II]
Tradução de: Laura Mendes Loureiro[III]

[I] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), Brasília-DF, Brasil; *alexandre.jaloto@inep.gov.br*
[II] Universidade São Francisco (USF), Campinas-SP, Brasil; *rprimi@mac.com*
[III] Viamundi Idiomas e Traduções, Belo Horizonte-MG, Brasil; *laura@viamundi.com.br*

RESUMO

O Exame Nacional do Ensino Médio (Enem) inclui uma redação e quatro provas com 45 itens cada. Sua confiabilidade e o impacto da fadiga nas pontuações são considerações importantes e, por isso, o Teste Adaptativo Computadorizado (CAT) pode ser uma maneira de resolver essas questões. Dessa forma, o presente estudo teve como objetivo verificar, por meio de um CAT, a possibilidade de redução do número de itens no Enem. Utilizando provas das edições de 2009 a 2019 do Enem, simulamos um CAT, que terminou quando o erro foi menor que 0,30, ou quando 45 itens foram aplicados. Em média, a aplicação variou de 12,0 (Linguagens e Códigos – LC) a 29,2 (Matemática – MT) itens. Os resultados apontam para o potencial de redução do tamanho do Enem para 20 itens para uma proporção que varia de 39,8% (MT) a 94,8% (LC) da população.

PALAVRAS-CHAVE AVALIAÇÃO DO ALUNO • PSICOMETRIA • TEORIA DA RESPOSTA AO ITEM • ENSINO SUPERIOR.

COMO CITAR:

Jaloto, A., & Primi, R. (2024). Enem de próxima geração com menos itens e alta confiabilidade usando CAT. *Estudos em Avaliação Educacional*, *35*, Artigo e10142. https://doi.org/10.18222/eae.v35.10142_port

NEXT-GENERATION ENEM ASSESSMENT WITH FEWER ITEMS AND HIGH RELIABILITY USING CAT

ABSTRACT

The Exame Nacional do Ensino Médio [Brazilian High School Exam] (Enem) is a test that includes an essay and four 45-item tests. Its reliability and the impact of fatigue on scores are important considerations, so Computerized Adaptive Testing (CAT) may be a way to address these issues. Therefore, the present study aimed to verify the possibility of reducing the number of items on the Enem, using a CAT. We used tests from the 2009 to 2019 editions of the Enem. We simulated a CAT, which ended when the error was less than 0.30, or when 45 items were applied. On average, the application ranged from 12.0 (Languages and Codes – LC) to 29.2 (Mathematics – MT) items. The results point to the potential of reducing the size of the Enem to 20 items for a proportion that varies from 39.8% (MT) to 94.8% (LC) of the population.

KEYWORDS STUDENT EVALUATION • PSYCHOMETRICS • ITEM RESPONSE THEORY • HIGHER EDUCATION.

PRÓXIMA GENERACIÓN DEL ENEM CON MENOS ÍTEMS Y ALTA CONFIABILIDAD USANDO CAT

RESUMEN

El Exame Nacional do Ensino Médio [Examen Nacional de Bachillerato] (Enem) incluye una redacción y cuatro pruebas con 45 ítems cada una. Su confiabilidad y el impacto de la fatiga en las puntuaciones son consideraciones importantes, y por eso, el Test Adaptativo Computarizado (CAT) puede ser una manera de resolver esas cuestiones. Por lo tanto, el presente estudio tuvo como objetivo verificar, a través de un CAT, la posibilidad de reducción del número de ítems en el Enem. Utilizando pruebas de las ediciones del Enem de 2009 a 2019, simulamos un CAT, que terminó cuando el error fue inferior a 0,30, o cuando 45 ítems fueron aplicados. En promedio, la aplicación osciló entre 12,0 (Idiomas y Códigos – LC) a 29,2 (Matemáticas – MT) ítems. Los resultados apuntan para el potencial de reducción del tamaño del Enem para 20 ítems para una proporción que varía del 39,8% (MT) al 94,8% (LC) de la población.

PALABRAS CLAVE EVALUACIÓN DEL ESTUDIANTE • PSICOMETRÍA • TEORÍA DE RESPUESTA AL ÍTEM • EDUCACIÓN SUPERIOR.

Recebido em: 13 MARÇO 2023
Aprovado para publicação em: 3 JUNHO 2024

INTRODUÇÃO

O objetivo do presente estudo é verificar se o Exame Nacional do Ensino Médio (Enem) pode ser simplificado com o uso do Teste Adaptativo Computadorizado (CAT) (Peres, 2019) sem comprometer a confiabilidade. Desde 2009, o Enem, composto de uma redação e quatro testes com 45 itens de múltipla escolha, tem sido usado como o único exame de admissão em várias instituições de ensino superior. Dado o amplo intervalo de notas de corte para diferentes cursos de graduação, a confiabilidade do exame é crucial. O presente estudo combina uma reflexão sobre a viabilidade técnica da implementação do CAT no Enem, tendo em vista o impacto que tem a posição de um item no desempenho no teste educacional (Debeer & Janssen, 2013; Domingue et al., 2020; Wu et al., 2016), com a consideração do potencial do CAT para melhorar a logística de sua aplicação no Enem. Também busca avançar no conhecimento sobre o uso do CAT em testes educacionais, por meio da realização de uma simulação com bases de dados compostas de mais de seiscentos itens.

ENEM

O Enem foi criado em 1998 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), órgão do Ministério da Educação. Entre seus objetivos estava o de fornecer uma estrutura para autoavaliação com base em indicadores de desempenho e no desenvolvimento de competências e habilidades inerentes à fase de desenvolvimento cognitivo e social ao final da educação básica (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [Inep], 2009).

Em 2009, o Enem passou por mudanças que o deixaram mais próximo de se tornar o único exame de admissão para cursos de graduação em instituições federais de ensino superior brasileiras. No âmbito do Sistema de Seleção Unificada (SiSU), criado pelo Ministério da Educação, as instituições selecionam os candidatos com base em suas pontuações nos testes. Além disso, a participação no Enem tornou-se um requisito para solicitar bolsas de estudo ou financiamento do governo para frequentar uma faculdade ou universidade privada.

O exame compreende uma redação e quatro provas de 45 itens cada, cada uma combinando habilidades de quatro diferentes áreas do conhecimento: Ciências Humanas (CH); Ciências da Natureza (CN); Linguagens e Códigos (LC); e Matemática (MT). Usando a Teoria da Resposta ao Item (TRI) (Pasquali & Primi, 2003), cada teste é aplicado em um formato linear e produz uma medida unidimensional para cada área (Inep, 2012a). Para selecionar os candidatos para admissão em seus cursos, as instituições de ensino podem usar uma média aritmética ou ponderada das quatro provas e da redação.

Implicações do uso do Enem para seleção

Existem vários desafios, em termos de confiabilidade e validade, associados ao uso do Enem como forma de seleção. Primeiro, é essencial demonstrar a validade do teste para prever o desempenho no ensino superior. Poucos estudos exploraram esse aspecto (veja, por exemplo, Ferreira-Rodrigues, 2015). Um segundo problema é a avaliação da equivalência das pontuações das redações, já que os avaliadores diferem em seus níveis de leniência ou rigor (Primi et al., 2019), o que não é levado em consideração na versão atual do teste. Por fim, o erro de medida é uma preocupação quando se usa o teste com candidatos de níveis de proficiência muito diferentes. Não sabemos quão grande é o erro em cada nível da escala, com alguns cursos de graduação tendo pontos de corte em níveis mais baixos da escala e outros em níveis mais altos.

O Inep transforma e padroniza as pontuações da TRI no Enem para que tenham uma média de M = 500 e um desvio padrão de DP = 100, usando as estatísticas da edição de 2009 como referência (Inep, 2012a). Considerando a grande variação nas pontuações de corte, os itens devem abranger um amplo espectro de dificuldades para obter pontuações suficientemente confiáveis para as decisões de seleção. Por exemplo, na edição de 2020 do SiSU, que utilizou o Enem 2019, a menor pontuação média para matrícula no curso de Ciências Sociais da Universidade Federal do Rio de Janeiro, para vagas reservadas para alunos negros com deficiência e para alunos de escolas públicas de baixa renda, foi de 394,10. Por outro lado, a menor pontuação média para ingressar no curso de Medicina na mesma instituição foi de 790,98. As pontuações de corte variaram de 227,78 (Engenharia de Aquicultura) a 928,30 (Medicina). Além disso, para se qualificar para bolsas e financiamentos públicos, é necessária uma média mínima de 450 pontos. Com base nessas pontuações de corte, é evidente que os testes exigem confiabilidade adequada em uma faixa de mais de 7 desvios padrão, dependendo da edição do Enem.[1] O Inep tenta atingir os objetivos do Enem por meio de testes com um número fixo de 45 itens. Claramente, é praticamente impossível atingir o mesmo nível de confiabilidade com 45 itens em um intervalo de 7 desvios padrão. O CAT pode ser uma das poucas maneiras de alcançar esse objetivo.

A fadiga também é uma preocupação. Sabemos que, em testes de desempenho educacional em larga escala e de alto impacto, a posição de uma pergunta pode afetar suas propriedades de dificuldade e discriminação (Domingue et al., 2020). Por exemplo, um item colocado na parte final de um teste terá uma taxa de acerto menor do que se fosse colocado no início do teste. Da mesma forma, o desempenho

1 Essas e outras informações sobre o histórico das edições do Enem e do SiSU estão disponíveis em www.gov.br/inep e www.sisu.mec.gov.br

dos participantes diminui nos itens finais de um teste (Debeer & Janssen, 2013; Wu et al., 2016). Além disso, no Enem de 2016, a posição do item em MT foi associada ao desempenho nesse item (Barichello et al., 2022). Esses achados sugerem que a fadiga pode afetar o 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 um problema de equidade para o teste. Alunos com níveis semelhantes de proficiência podem pontuar de forma diferente em decorrência da fadiga, em vez de competência. A aplicação de um CAT poderia atenuar o problema do tamanho de um teste e, consequentemente, reduzir os possíveis efeitos da fadiga. A promessa do CAT é manter, ou mesmo melhorar, a confiabilidade de um teste com menos itens (Veldkamp & Matteucci, 2013; Weiss, 2011).

Enem de próxima geração usando CAT

Na estrutura da TRI, o erro está inversamente relacionado à quantidade de informação na região do nível de medida estimado (Ayala, 2009). Geralmente, uma pergunta é mais informativa para pessoas cujas habilidades são comparáveis à sua dificuldade. Portanto itens que são muito difíceis ou muito fáceis têm pouco impacto na precisão para avaliar pessoas com habilidades mais baixas ou mais altas. No entanto, devido à natureza do Enem, que deve ser capaz de medir uma grande amplitude com precisão, é solicitado que os participantes com altos níveis de habilidade respondam itens fáceis, mesmo que a informação que eles acrescentem seja mínima. Além disso, os participantes com baixos níveis de habilidade são obrigados a responder a perguntas com um nível muito alto de dificuldade. Como alternativa, o CAT seleciona itens a serem administrados com base nas estimativas provisórias de habilidade do participante, calculadas a partir de respostas anteriores, de modo que os itens que estiverem muito longe do seu nível de habilidade serão evitados.

Seguindo esse procedimento, os participantes responderão a uma grande proporção de itens que são úteis para informar seu nível de proficiência e não perderão tempo respondendo a itens não informativos. Isso poderia resolver um problema com o teste de MT do Enem, que está fora do alvo para as pessoas que o fazem. Esse teste é muito mais difícil do que o nível médio de proficiência da maioria dos participantes que fazem o Enem. Como resultado, a confiabilidade da maioria das pontuações é muito baixa, tornando difícil fazer uma diferenciação entre alunos de baixa habilidade.

Estudos confirmaram que os CATs podem ser usados para reduzir o tamanho dos testes educacionais. Kalender e Berberoglu (2017) simularam um teste, composto

em média de 17 itens, para ingresso no ensino superior na Turquia. Originalmente, esse teste tinha 45 itens. Spenassato et al. (2016) simularam a aplicação dos 45 itens de MT do Enem 2012 no formato de um CAT e concluíram que um conjunto de 33 itens do CAT produziu resultados comparáveis. Mizumoto et al. (2019) demonstraram que testes de vocabulário de língua inglesa compostos de 115, 73 e 56 itens poderiam ser reduzidos a testes com 20, 15 e 10 itens, respectivamente. Considerando que esses são testes de grande importância, os bancos de itens nesses estudos continham relativamente poucos itens para um CAT. No presente estudo, os bancos de itens variaram de 674 a 839 itens.

Além de reduzir o tamanho do teste e o erro de medida, o CAT também pode contribuir significativamente para questões práticas de um teste de larga escala. Por exemplo, é possível pré-testar itens durante a administração regular, reduzir a complexidade logística (sem necessidade de transportar material físico), reduzir as chances de divulgação fraudulenta do conteúdo do teste e fornecer devolutiva imediata aos participantes.

O PRESENTE ESTUDO

O tamanho do teste, um dos pontos citados acima, foi escolhido como ponto de partida para explorar e refletir sobre a viabilidade de implementação do CAT no Enem. Assim, o objetivo do presente estudo é investigar se é possível reduzir o número de itens no Enem, por meio do uso do CAT, sem comprometer a confiabilidade da medida. Nossa pesquisa foi dividida em dois estudos. Primeiramente, objetivamos determinar os parâmetros dos itens aplicados no Enem e equalizá-los a uma única escala, uma vez que o Inep não havia divulgado os parâmetros dos itens até a realização deste estudo. Essas informações já estão disponíveis no *site* do Inep, mas a metodologia discutida no primeiro estudo permanece relevante para pesquisadores interessados em replicar ou desenvolver nosso trabalho. O segundo estudo tem como objetivo simular o CAT utilizando esses itens.

ESTUDO 1

Neste estudo, nos propusemos a criar um banco de itens do Enem, usando uma única métrica, que foi utilizada posteriormente no CAT no Estudo 2. O modelo logístico de três parâmetros da TRI foi utilizado para calibrar os parâmetros de itens de diferentes anos. Em seguida, os convertemos em uma única métrica, usando as pontuações dos alunos do banco de dados oficial do Inep.

Participantes

Este estudo utiliza dados secundários, extraídos dos microdados do Enem de janeiro a junho de 2020 e disponíveis no portal do Inep.[2] Foram excluídos participantes cujos microdados apresentassem inconsistências, como um vetor de respostas com 44 em vez de 45 caracteres, além daqueles que deixaram as 45 respostas em branco.

Selecionamos aleatoriamente amostras de 5.000 participantes de cada aplicação das edições do Enem de 2009 a 2019. As amostras foram estratificadas com base na soma de respostas corretas. Portanto garantimos que os participantes com pontuações altas seriam sorteados e que os itens mais difíceis seriam calibrados adequadamente. Selecionamos 1.250 participantes dos estratos inferior e superior (percentis 25 e 95, respectivamente) e 2.500 do estrato médio. Em aplicações com menos de 25.000 participantes, o estrato superior teve menos de 1.250 indivíduos ($0{,}05 * 25.000 = 1.250$). Nesses casos, a amostra foi suplementada com participantes de outros estratos, mantendo uma proporção de 1:2 (inferior:intermediário). Devido ao arredondamento, algumas amostras compreenderam mais de 5.000 pessoas.

Há pelo menos duas aplicações em cada edição do Enem. Normalmente, a primeira aplicação tem o maior número de participantes. Sempre que havia menos de 5.000 participantes, a calibração era realizada com toda a população, se maior ou igual a 1.000. Esse tamanho de amostra é adequado para usar o modelo TRI de três parâmetros para calibrar itens (Şahin & Anıl, 2017). Portanto excluímos aplicações com menos de 1.000 participantes. O sorteio da amostra foi realizado utilizando a função *strata* do pacote *sampling* (v2.8) (Tillé & Matei, 2016) no ambiente de programação R (R Core Team, 2019). Na Tabela 1, sintetizamos o tamanho da amostra e da população para cada aplicação do Enem, bem como outras informações que serão discutidas na seção Resultados do Estudo 1.

2 www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados

TABELA 1

Informações sobre cada aplicação (participantes, amostra, itens utilizados e correlação)

ÁREA	ANO	APLICAÇÃO	POPULAÇÃO	AMOSTRA	NÚMERO DE ITENS USADOS	CORRELAÇÃO
Ciências Humanas	2009	1	2.552.781	5.000	45	0,996
	2009	2	1.660	1.660	45	0,977
	2010	1	3.369.211	5.000	45	0,997
	2010	2	3.904	3.904	45	0,993
	2011	1	3.981.762	5.000	45	0,995
	2011	2	10.963	5.001	45	0,991
	2012	1	4.217.478	5.000	45	0,995
	2013	1	5.198.617	5.000	45	0,994
	2014	1	6.159.992	5.000	45	0,970
	2015	1	5.747.279	5.000	45	0,987
	2016	1	5.836.551	5.000	45	0,992
	2016	3	159.440	5.000	45	0,990
	2017	1	4.689.506	5.000	45	0,987
	2017	2	1.039	1.039	45	0,989
	2018	1	4.136.361	5.000	45	0,983
	2018	2	1.134	1.134	45	0,995
	2019	1	3.914.432	5.000	45	0,987
				Total	**765**	
Ciências da Natureza	2009	1	2.554.741	5.000	45	0,997
	2009	2	1.668	1.668	45	0,964
	2010	1	3.366.011	5.000	45	0,996
	2010	2	3.890	3.890	45	0,989
	2011	1	3.980.082	5.000	45	0,998
	2011	2	10.953	5.000	45	0,974
	2012	1	4.216.367	5.000	45	0,989
	2013	1	5.197.365	5.000	45	0,986
	2014	1	6.158.548	5.000	45	0,983
	2015	1	5.746.263	5.000	45	0,984
	2016	1	5.835.361	5.000	45	0,984
	2016	3	159.419	5.000	44	0,988
	2017	1	4.433.922	5.000	45	0,972
	2018	1	3.901.771	5.000	45	0,987
	2019	1	3.707.205	5.000	45	0,987
				Total	**674**	

(continua)

(continuação)

ÁREA	ANO	APLICAÇÃO	POPULAÇÃO	AMOSTRA	NÚMERO DE ITENS USADOS	CORRELAÇÃO
Linguagens e Códigos	2009	1	2.434.642	5.000	44	0,999
	2009	2	1.554	1.554	45	0,977
	2010	1	3.246.005	5.000	50	0,998
	2010	2	2.306	2.306	50	0,994
	2011	1	3.866.703	5.000	50	0,997
	2012	1	4.090.691	5.000	50	0,982
	2013	1	5.022.660	5.000	50	0,994
	2014	1	5.971.721	5.000	50	0,996
	2015	1	5.615.054	5.000	50	0,994
	2015	2	1.038	1.038	50	0,995
	2016	1	5.685.125	5.000	50	0,988
	2016	3	155.250	5.000	50	0,986
	2017	1	4.693.808	5.000	50	0,993
	2017	2	1.039	1.039	50	0,996
	2018	1	4.140.393	5.000	50	0,989
	2018	2	1.134	1.134	50	0,990
	2019	1	3.917.238	5.000	50	0,985
				Total	**839**	
Matemática	2009	1	2.433.932	5.000	45	0,984
	2009	2	1.549	1.549	45	0,973
	2010	1	3.244.895	5.000	45	0,997
	2010	2	2.301	2.301	45	0,981
	2011	1	3.865.301	5.000	45	0,995
	2011	2	10.364	5.001	45	0,979
	2012	1	4.088.847	5.000	45	0,997
	2013	1	5.020.489	5.000	45	0,990
	2014	1	5.968.542	5.000	45	0,986
	2015	1	5.612.869	5.000	45	0,988
	2015	2	1.038	1.038	45	0,984
	2016	1	5.683.429	5.000	45	0,992
	2016	3	155.197	5.000	45	0,994
	2017	1	4.433.702	5.000	45	0,987
	2018	1	3.901.617	5.000	44	0,985
	2019	1	3.707.065	5.000	45	0,987
				Total	**719**	

Fonte: Elaboração dos autores.

Nota: Tamanho da população, tamanho da amostra usada para calibração, total de itens e correlação entre as pontuações reestimadas e as pontuações originais (oficiais) da população para cada aplicação de cada área de conhecimento. Todos os coeficientes de correlação foram estatisticamente significativos ($p < 0,001$).

Medidas

Cada um dos quatro testes do Enem contém 45 itens de múltipla escolha com cinco opções, das quais apenas uma está correta. O modelo logístico de TRI de três parâmetros é usado para estimar as pontuações, que são posicionadas em uma escala com uma média de 500 e um desvio padrão de 100. A escala tem como referência os egressos regulares das escolas públicas de 2009 (Inep, 2012a). Desde 2010, o teste de LC continha cinco itens de língua estrangeira. Portanto cada teste nessa área contém 50 perguntas. Os participantes podem selecionar inglês ou espanhol como sua língua estrangeira de escolha. Os outros três testes (CH, CN e MT) têm 45 itens cada. Ocasionalmente, o Inep pode excluir um item por razões pedagógicas (por exemplo, duas respostas corretas). Neste estudo, usamos todas as provas das edições de 2009 a 2019 do Enem que não passaram por adaptação (por exemplo, aquelas para pessoas com deficiência visual). Adicionalmente, foram excluídos os testes com menos de 1.000 participantes, os que não possuíam microdados e a segunda aplicação da LC do Enem 2011, pois as informações referentes ao seu gabarito são inconsistentes. A Tabela 1 apresenta o número de itens utilizados para cada área.

Análise dos dados

A calibração foi realizada usando o pacote mirt (v1.33.2) (Chalmers, 2012). Para a discriminação, usamos uma distribuição prévia log-normal (média de 0 e desvio padrão de 0,5, o que garante valores positivos). Para o pseudochute, usamos distribuições beta com parâmetros 7 e 28 (centralizadas em torno de 0,2, o que é adequado para itens com cinco alternativas). Reestimamos todas as pontuações dos participantes para cada aplicação usando o pacote mirt, com o método EAP (*expected a-posteriori*). A adequação da calibração foi avaliada correlacionando as pontuações reestimadas e as oficiais (publicadas pelo Inep). De acordo com a suposição de invariância dos parâmetros da TRI, esperávamos que esses valores de correlação fossem próximos de 1, com as únicas diferenças sendo o centro da escala e a escala (intercepto e inclinação).

Conforme mencionado, o Inep não havia divulgado os parâmetros do item. As únicas informações que tínhamos sobre a escala de um teste eram as pontuações oficiais dos participantes, que foram equalizadas ao longo dos anos. Usamos o método *sigma-mean* para colocar os itens na mesma métrica do Inep (Hambleton et al., 1991). As pontuações oficiais de cada amostra (métrica Inep) foram transformadas para uma média de 0 e um desvio padrão de 1. Também recalculamos as pontuações dos alunos usando o pacote mirt nas métricas 0 e 1 (pontuações mirt). Portanto obtivemos dois valores teta para cada estudante (oficial e mirt), que deveriam ter sido idênticos. No entanto não foram, pois é improvável que a distribuição de habilidade da amostra selecionada tivesse uma média exata de 0 e um desvio padrão

de 1. Porém o mirt assume esses valores ao calibrar/estimar a identificação métrica. Imagine que Y_i representa a pontuação oficial (métrica na qual queremos calibrar os itens) e X_i representa a pontuação mirt, para o sujeito i. Podemos expressar a igualdade dessas duas pontuações da seguinte forma:

$$\frac{Y_i - \bar{Y}}{DP_y} = \frac{X_i - \bar{X}}{DP_x} \quad \text{(E1)}$$

onde $\bar{Y}$ e DP_y representam a média e o desvio padrão das pontuações oficiais, e $\bar{X}$ e DP_x representam a média e o desvio padrão das pontuações mirt. Nessa equação, consideramos que as pontuações padronizadas das duas métricas são (ou deveriam ser) iguais, pois são provenientes da mesma amostra. Se isolarmos Y_i, teremos (Muñiz, 1997):

$$Y_i = \frac{DP_y}{DP_x} X_i + \left[\bar{Y} - \frac{DP_y}{DP_x}\bar{X}\right] \quad \text{(E2)}$$

As constantes k e d

$$k = \frac{DP_y}{DP_x} \quad \text{(E3)}$$

$$d = \bar{Y} - k\bar{X} \quad \text{(E4)}$$

extraídas dessa equação representam as constantes de equalização (escala e origem, respectivamente) para transformar as pontuações dos participantes e os parâmetros dos itens obtidos via mirt para a métrica do Inep. Como as pontuações dos participantes e os parâmetros dos itens estão na mesma métrica, podemos usar essas constantes para transformar os parâmetros b e a do item j da métrica mirt para a métrica oficial do Inep, usando as seguintes equações lineares:

$$b_{j\,inep} = k b_{j\,mirt} + d \quad \text{(E5)}$$

$$a_{j\,inep} = \frac{a_{j\,mirt}}{k} \quad \text{(E6)}$$

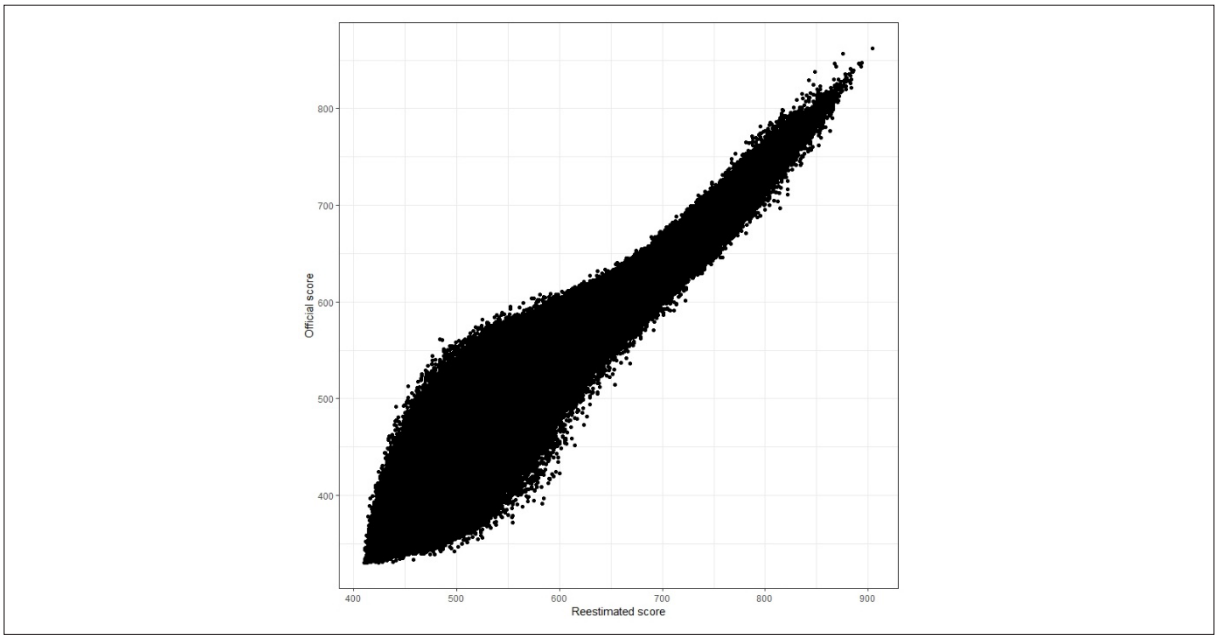
Os itens na mesma área são, então, todos colocados na métrica oficial do Inep após essas transformações. Nossos resultados são, portanto, comparáveis aos resultados oficiais, e podemos estimar a pontuação de um participantes da simulação usando todos os itens de uma determinada área. É como se tivéssemos uma prova composta de centenas de itens para cada uma das áreas de conhecimento do Enem.[3]

3 Os códigos utilizados neste trabalho estão disponíveis em https://github.com/alexandrejaloto/EAE_CAT/

Resultados

As correlações entre as pontuações reestimadas e as oficiais foram estatisticamente significativas ($p < 0{,}001$) e altas. A menor foi de 0,96, observada em CN (2ª aplicação de 2009). As correlações de alta magnitude mostram que as variações nas pontuações reestimadas explicam as variações nas pontuações oficiais. Isso indica que a calibração neste estudo produziu parâmetros de itens comparáveis aos oficiais. Pode-se questionar por que esses valores não chegam a 1 e diferem apenas em escala e origem. Pode haver várias razões para isso, incluindo a questão de estimar um parâmetro de pseudochute. De acordo com nossas observações, as maiores discrepâncias entre nossos resultados e os oficiais foram observadas em alunos com habilidade baixa. Isso pode ser atribuído a diferenças entre o parâmetro de pseudochute. Uma recalibração pode mostrar diferenças no pseudochute mais do que outros parâmetros (Primi et al., 2018). A Figura 1 mostra as pontuações oficiais como uma função das pontuações reestimadas na primeira aplicação de CH em 2014, a menor correlação observada em uma aplicação principal.

FIGURA 1
Relação entre as pontuações oficiais e reestimadas na primeira aplicação do Enem 2014 (CH)



Fonte: Elaboração dos autores.

A Tabela 1 usa todos os valores de correlação entre as pontuações reestimadas e as oficiais. Uma vez obtidas estimativas razoáveis dos parâmetros dos itens aplicados no Enem, passamos para o Estudo 2. Devido às altas correlações, podemos dizer que a simulação do Estudo 2 apresenta resultados comparáveis àqueles publicados pelo Inep.

ESTUDO 2

O objetivo deste estudo foi simular a aplicação do CAT usando itens de todas as edições do Enem. Para tanto, selecionamos uma amostra aleatória de participantes da edição de 2019 para obter um conjunto inicial de valores teta. Em seguida, simulamos as respostas desses participantes para todos os itens do banco de itens. Depois, aplicamos o algoritmo CAT usando o erro de medida e/ou o número de itens aplicados como uma regra de parada.

Participantes

Uma amostra aleatória simples de participantes da edição de 2019 do Enem foi sorteada para cada área. Ao contrário do Estudo 1, não excluímos nenhum participante dessa população, porque os itens já haviam sido calibrados e o insumo dos microdados para o Estudo 2 foi a estimativa de habilidade dos participantes. Usamos todo o grupo de participantes, refletindo um cenário de teste do mundo real. O tamanho da amostra foi suficiente para garantir uma média com erro de amostragem de 3 pontos na escala do Enem, que foi equivalente a 0,03 unidade de desvio padrão. Ao adotar esse procedimento, conseguimos generalizar nossos resultados para a população dessa edição do Enem, o que potencialmente aproxima nossa simulação de situações esperadas para edições futuras com características semelhantes. As estatísticas descritivas dos participantes da edição de 2019 e das amostras retiradas de cada área são apresentadas na Tabela 2.

As respostas aos itens foram simuladas usando a função *generate_pattern* do pacote mirtCAT (v1.10) (Chalmers, 2016); os valores de entrada foram as pontuações oficiais da amostra (que chamamos de pontuações verdadeiras) e os parâmetros do item. Como resultado, obtivemos quatro bancos de respostas, um para cada área de conhecimento. Cada banco de respostas continha o número de linhas correspondente ao tamanho da amostra para cada área de conhecimento, e o número de colunas correspondia ao tamanho do banco de itens para cada área. Produzimos um banco de respostas, como se cada participante tivesse respondido a todos os itens de uma área.

TABELA 2
Estatística descritiva dos participantes do Enem 2019 e de suas amostras

ÁREA	N	MÉDIA	DESVIO PADRÃO	INTERVALO
Ciências Humanas				
População	3.917.245	508,0	80,1	315,9-835,1
Amostra	2.738	509,0	79,9	321,5-771,1

(continua)

(continuação)

ÁREA	N	MÉDIA	DESVIO PADRÃO	INTERVALO
Ciências da Natureza				
População	3.709.827	477,9	75,9	327,9-860,9
Amostra	2.457	478,1	75,1	329,2-737,0
Linguages e Códigos				
População	3.920.058	520,9	62,5	322,0-801,7
Amostra	1.667	522,1	62,0	325,6-688,0
Matemática				
População	3.709.686	523,2	108,8	359,0-985,5
Amostra	5.055	522,4	107,8	359,0-929,2

Fonte: Elaboração dos autores.

Medidas

Foram utilizados os quatro bancos de itens obtidos no estudo anterior. O banco de CH continha 765 itens, o de CN 674 itens, o de LC 839 itens e o de MT 719 itens.

Análise dos dados

A análise de dados para o Estudo 2 consistiu na simulação CAT, que foi realizada usando o pacote mirtCAT. Os itens foram apresentados de acordo com o critério de máxima informação. O método de estimação foi o EAP. Chamamos as pontuações estimadas na simulação CAT de pontuações estimadas. A simulação terminou quando o erro padrão de medida atingiu um valor de 0,30 ou quando 45 itens foram aplicados. Esse valor de erro padrão corresponde a uma confiabilidade de 0,91, uma vez que (Nicewander & Thomasson, 1999):

$$\sigma_{\hat{\theta}} = \sqrt{1 - \rho(\hat{\theta})} \qquad \text{(E7)}$$

onde $\rho(\hat{\theta})$ é a confiabilidade para uma determinada pontuação e $\sigma_{\hat{\theta}}$ é o erro padrão de medida. Substituindo o valor do erro, temos:

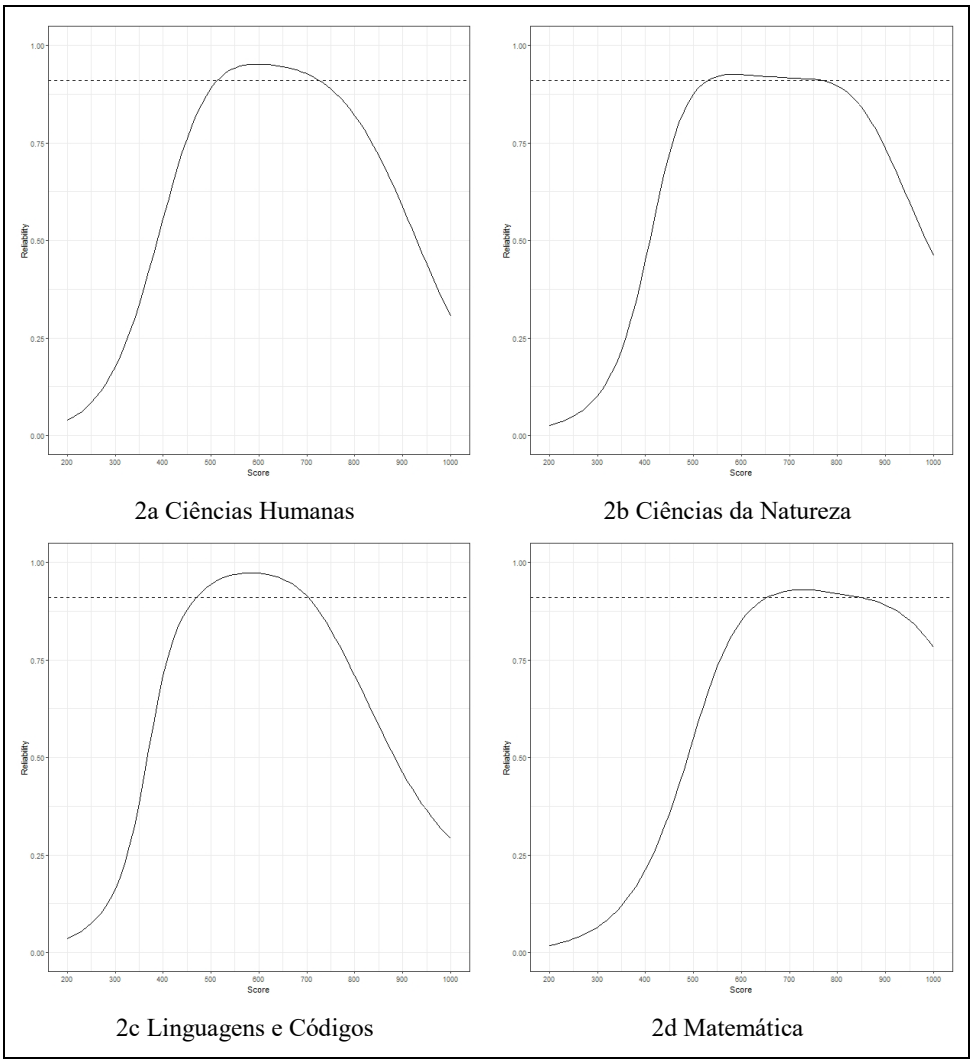
$$0{,}30 = \sqrt{1 - \rho(\hat{\theta})} \therefore \rho(\hat{\theta}) = 0{,}91 \qquad \text{(E8)}$$

O critério utilizado foi mais rigoroso do que a maior confiabilidade marginal encontrada na primeira aplicação do Enem em 2019 (HS = 0,78; NS = 0,74; LC = 0,84; MT = 0,53). A confiabilidade marginal reflete a precisão média em toda a escala (Ayala, 2009). A confiabilidade de 0,91 também excede a precisão na maior parte das escalas desses testes. A precisão em cada ponto da escala $\rho(\theta)$ foi calculada com base na informação $I(\theta)$ do teste para este ponto (Nicewander & Thomasson, 1999):

$$\rho(\theta) = \frac{I(\theta)}{I(\theta)+1} \qquad \text{(E9)}$$

A Figura 2 mostra a confiabilidade em função da pontuação nos testes da primeira aplicação do Enem em 2019.

FIGURA 2
Confiabilidade dos testes da primeira aplicação do Enem 2019



Fonte: Elaboração dos autores.
Nota: A linha horizontal indica a confiabilidade de 0,91.

Resultados

Observamos uma redução de pelo menos 40% no tamanho do teste para mais da metade dos participantes e um erro máximo de 0,30 em todas as áreas. O teste de CH foi reduzido para um máximo de 20 itens para 71,7% dos participantes. A mesma redução nos testes de CN, LC e MT foi possível para 60,4%, 94,8% e 39,8% dos participantes, respectivamente. Com 15 itens, foi possível estimar a pontuação de mais da metade dos participantes nos testes de CH, CN e LC, com erro máximo de 0,30.

Cerca de 30 itens foram necessários no teste de MT. Além disso, 90,3% dos participantes em CH, 79,6% em CN, 99,2% em LC e 65,6% em MT reduziram pelo menos um item em seu teste. Em média, 18,4 itens foram administrados em CH, 22,1 em CN, 12,0 em LC e 29,2 em MT. A Tabela 3 mostra o percentual de participantes em cada área que se submeteram a um máximo de 15, 20, 30 ou 44 itens.

TABELA 3
Resultados da simulação (correlação, erro, itens aplicados e percentual de participantes)

	CIÊNCIAS HUMANAS	CIÊNCIAS DA NATUREZA	LINGUAGENS E CÓDIGOS	MATEMÁTICA
Correlação	0,937	0,922	0,904	0,957
Média do erro padrão	0,29	0,31	0,29	0,32
Maior erro padrão	0,47	0,53	0,43	0,56
Número mínimo de itens aplicados	8	8	7	11
Média de itens aplicados	18,4	22,1	12,0	29,2
No máximo 15 itens	58,7%	52,3%	89,6%	25,8%
No máximo 20 itens	71,7%	60,4%	94,8%	39,8%
No máximo 30 itens	83,1%	70,9%	97,8%	54,0%
No máximo 44 itens	90,3%	79,6%	99,2%	65,6%

Fonte: Elaboração dos autores.

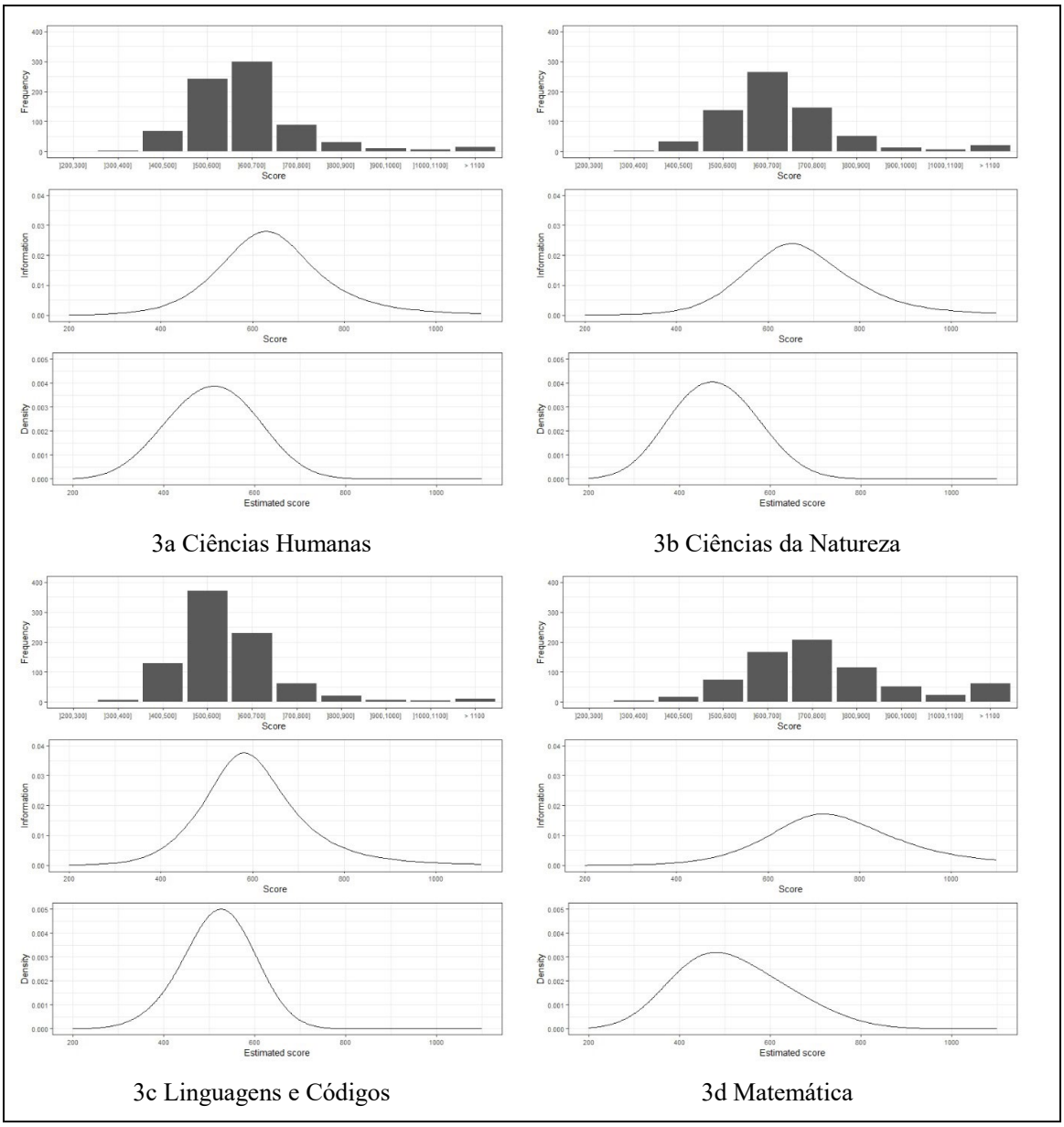
Notas: Correlação entre as pontuações estimadas e as pontuações reais, média do erro padrão de medida, maior erro padrão, menor número de itens aplicados, média de itens aplicados e porcentagem de participantes submetidos a um máximo de 15, 20, 30 ou 44 itens. Todos os coeficientes de correlação foram estatisticamente significativos ($p < 0,001$).

Também encontramos casos em que o erro foi maior que 0,30 mesmo com 45 itens aplicados. O erro padrão médio foi de 0,29 no teste de CH e LC, 0,31 em CN e 0,32 em MT. O maior erro observado no teste de CH foi de 0,47; em CN, 0,53; em LC, 0,43; e em MT, 0,56. Embora nem todos os erros tenham sido inferiores a 0,30, as pontuações estimadas apresentaram correlação significativa com as pontuações reais. O uso de um método otimizado de seleção de itens para criar um teste adaptado a cada participante produziu resultados semelhantes aos que teriam sido alcançados se todos os itens tivessem sido respondidos. A Tabela 3 mostra os valores dos erros e correlações, o número mínimo e o número médio de itens aplicados na simulação para cada área.

A Figura 3 mostra o número de itens no banco em função de sua localização na escala, a distribuição da pontuação estimada e a curva de informação do banco de itens. Considerando a localização dos itens na escala: em LC, a região com maior número de itens está próxima da média; em CH e CN, essa região está próxima de uma unidade de desvio padrão acima da média; e, em MT, essa região está próxima de duas unidades de desvio padrão acima da média. Entre as quatro áreas, essas são as regiões com mais informação. No entanto, uma parte significativa da distribuição

teta cai em áreas com informação limitada para CH, CN e MT. Essa discrepância surge da falta de itens destinados a medir níveis de habilidade mais baixos nessas áreas. Consequentemente, o banco de itens atual tem precisão questionável na medida das habilidades de alunos com menor proficiência.

FIGURA 3
Distribuição de frequência dos itens, curva de informação do banco de itens e distribuição das pontuações estimadas para todas as áreas



Fonte: Elaboração dos autores.

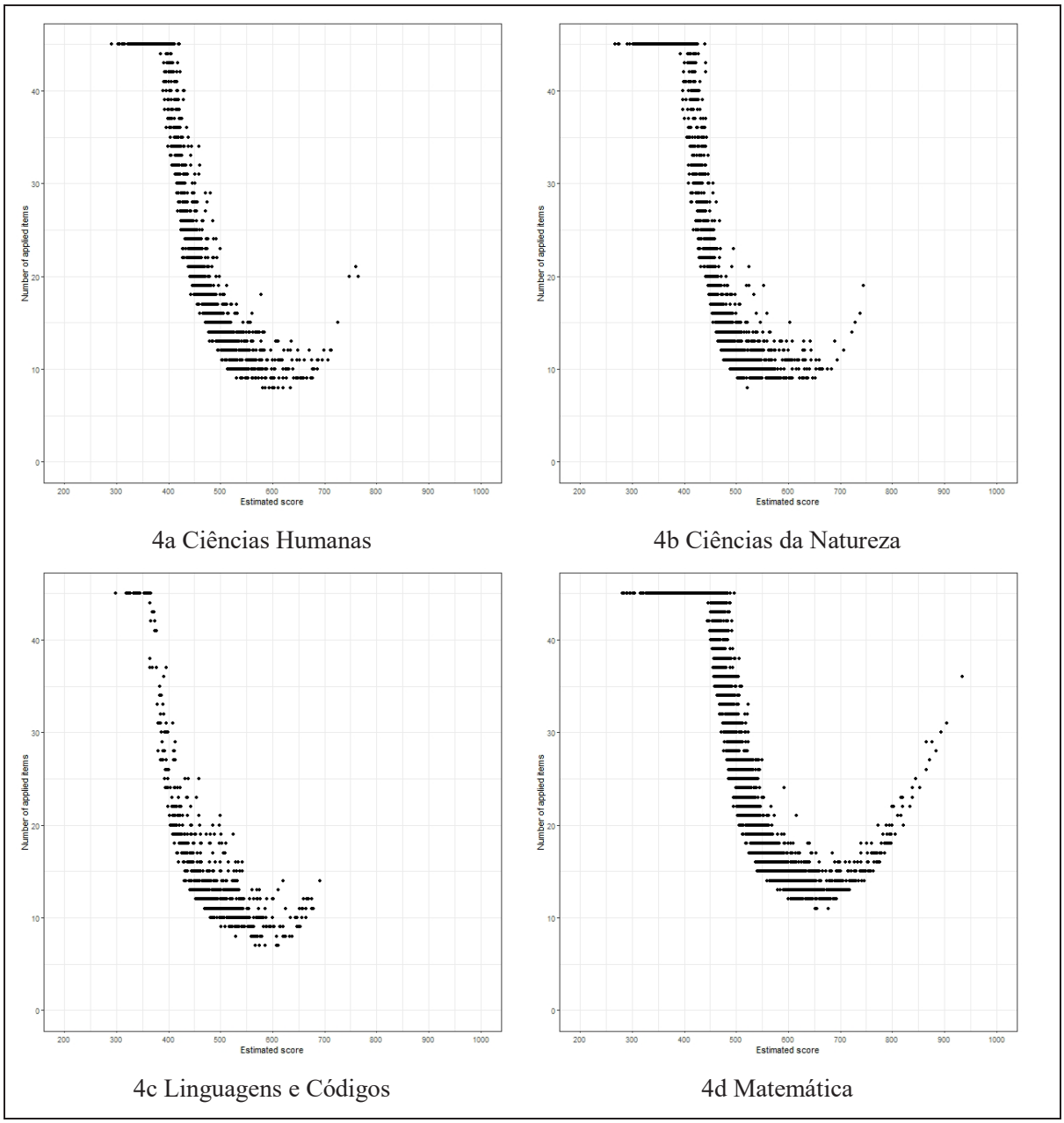
Nota: Em cada quadrante, a figura superior mostra o número de itens em função de sua localização na escala, a figura do meio mostra a curva de informação do banco de itens e a figura inferior mostra a distribuição das pontuações estimadas.

O algoritmo CAT tendeu a dar mais itens aos participantes localizados em regiões com níveis mais baixos de informação. Na Figura 4, o número de itens no

CAT é apresentado em função da nota estimada. É evidente que 20 itens foram suficientes para participantes com teta na faixa de aproximadamente: 500-750 (em CH); 525-750 (em CN); 500-700 (em LC); e 615-800 (em MT).

Nas regiões inferiores da escala, o algoritmo CAT aplicou mais itens. Por exemplo, em CH, todos os participantes com teta menor que 385,1 receberam 45 itens; em CN, 393,6; em LC, 348,0; e em MT, 445,1. O maior teta de participantes a receber 45 itens em CH foi 421,0; em CN, 439,3; em LC, 366,7; e em MT, 495,8. Os tamanhos dos testes de todos os participantes com pontuações superiores a estas foram reduzidos em pelo menos um item.

FIGURA 4
Número de itens apresentados de acordo com a pontuação estimada em cada teste



Fonte: Elaboração dos autores.

DISCUSSÃO

No presente estudo, o objetivo foi determinar se é possível reduzir o número de itens no Enem utilizando um CAT, sem comprometer a confiabilidade. Os resultados mostram que, nas quatro áreas de conhecimento avaliadas no Enem, seria possível uma redução em um amplo intervalo das escalas. A simulação terminou com um máximo de 20 itens em faixas de escala que corresponderam a 2,5 unidades de desvio padrão em CH; 2,2 unidades de desvio padrão em CN; 1,9 unidade de desvio padrão em LC; e 1,8 unidade de desvio padrão em MT. Com uma confiabilidade de 0,91, 20 itens poderiam estimar 71,7% das pontuações dos participantes em CH; 60,4% em CN; 94,8% em LC; e 39,8% em MT. Em todos os testes, para mais da metade dos participantes a aplicação não chegou a 30 itens.

No presente estudo, o tamanho médio da aplicação de CH foi próximo ao da simulação conduzida por Kalender e Berberoglu (2017) – 17 –, que usaram um banco de 45 itens de um teste de admissão ao ensino superior turco, com precisão como critério de parada (erro de 0,30). Eles observaram correlações entre as pontuações baseadas no CAT e as pontuações originais (teste linear) variando de 0,66 a 0,93. Todas as nossas correlações excederam 0,90, o que esperávamos, pois tínhamos um maior número de itens no banco.

Constatamos que nossa simulação foi mais precisa do que a de Spenassato et al. (2016), que simulou um CAT com 45 itens de MT do Enem 2012. Com uma aplicação fixa de 33 itens, eles observaram uma correlação de 0,99 entre o CAT e as pontuações originais (teste linear) e um erro médio de 0,35. Em sua aplicação linear, os autores observaram erros de até 0,84. Em nossa simulação, encontramos um erro máximo de 0,56 com uma média de 0,32 em MT. Em contraste com o estudo deles, nosso limite de itens foi de 45. No entanto, para 54,0% da amostra, 30 itens foram suficientes para um erro de 0,30.

O artigo de Mizumoto et al. (2019) apresentou a possibilidade de reduzir um teste de vocabulário de língua inglesa composto de três partes de 115, 73 e 56 itens para uma aplicação fixa de 20, 15 e 10 itens, respectivamente. A proporção de amostras com erro de medida de até 0,33 variou de 69,61% a 74,34%. CH (71,7%), com até 20 itens, e CN (70,9%), com até 30 itens, apresentam proporções semelhantes. Nossa proporção de aplicações de LC, com 15 itens ou menos, foi maior (89,6%). Em MT, foram necessários 44 itens para atingir um percentual semelhante (65,6%).

É importante destacar que nossos resultados indicam que os dois testes com mais itens (CH e LC) e com itens mais direcionados (ou seja, itens cuja distribuição de dificuldade corresponde à distribuição teta) proporcionaram as maiores reduções. Isso enfatiza a importância de um banco de itens robusto para o CAT reduzir o tamanho do teste de forma eficaz. Um ponto digno de nota é, no entanto, que a variação na amostra de LC e na população foi a menor entre os quatro grupos.

Consequentemente, não é necessário cobrir regiões mais extremas da escala. Concluímos que, nos casos em que a população apresenta baixa variância, o banco de itens pode ser menor do que nos casos em que a população varia muito. Ou seja, quanto mais estreita a área em que a pontuação da população está localizada, menos itens são necessários para estimar sua pontuação com precisão. A menor redução em MT do que em CN reforça esse ponto. A primeira área tinha um banco de itens maior, mas a variância populacional também era maior. Estudos futuros podem verificar o impacto do tamanho da variância na necessidade de aumentar o banco de itens.

No presente estudo, os resultados para todos os testes mostram consistentemente que as curvas de informação atingem o pico à direita da média da escala (500). Isso indica maior precisão para participantes com níveis de habilidade acima da média. Consequentemente, o potencial de redução do teste é limitado principalmente a participantes com notas acima de 500. Essas descobertas sugerem um desafio na redução do Enem no intervalo completo das quatro escalas.

Ressalta-se que o banco de simulação possuía centenas de itens, e o algoritmo selecionou os itens que mais contribuíram para a precisão da pontuação do participante. No entanto, para alguns participantes, mesmo os 45 itens (como nos testes do Enem) não conseguiram atingir um erro de 0,30. Portanto, ainda que o participante tivesse respondido ao teste linear, o erro de sua pontuação provavelmente teria sido maior que 0,30. A falha do CAT em reduzir o tamanho do teste para esses indivíduos não se deve a uma limitação do algoritmo do CAT, mas sim à limitação do banco de itens para cobrir igualmente os níveis mais baixos de teta.

Estudos anteriores também encontraram restrições para limitar o teste a determinadas faixas da escala. Por exemplo, Kalender e Berberoglu (2017) reduziram o teste a 25 itens para intervalos de escala variando de 2,8 a 4,3 desvios padrão. No presente estudo, a redução para 20 itens ocorreu em intervalos que variaram entre 1,8 e 2,5 desvios padrão. Mizumoto et al. (2019) verificaram que a maioria das pontuações com erros acima de 0,33 em sua simulação de CAT estava acima da média da escala. Segundo Spenassato et al. (2016), a região abaixo da média apresentou as maiores médias de erro de medida.

Dado o pequeno número de itens (um máximo de 115) nas simulações descritas acima, é mais fácil entender as dificuldades de alcançar um grau adequado de precisão. Em nosso estudo, contudo, os bancos de itens continham pelo menos 674 itens. No entanto observamos um padrão semelhante. Diante disso, há uma possível limitação dos testes educacionais de larga escala, que apresentam um excesso de itens difíceis para discriminar os alunos de alta habilidade, mas não têm o mesmo número de itens fáceis que diferenciem melhor os alunos de baixa habilidade. Com

relação ao Enem, especificamente, a baixa precisão na região inferior da escala pode afetar a seleção de cursos com baixa relação candidato-vaga. Além disso, também pode ter um impacto negativo na concessão de bolsas ou financiamentos públicos. Consequentemente, o presente estudo reforça a necessidade de desenvolvimento de itens mais fáceis para o Enem.

O Inep desenvolveu diretrizes para a elaboração e revisão dos itens de teste (Inep, 2010, 2012b). No entanto não há diretrizes objetivas sobre como escrever itens com níveis predefinidos de dificuldade. A fim de apoiar a orientação de seu desenvolvimento objetivo, recomendamos investigações sobre os fatores associados à dificuldade do item. Já foram encontradas algumas evidências de que aspectos linguísticos de um item de avaliação educacional em larga escala podem estar relacionados à sua dificuldade (Kan & Bulut, 2015; Masri et al., 2017).

Uma limitação do presente estudo é que a exposição ao item não foi controlada, o que seria crucial em um cenário real de aplicação. Além disso, o trabalho não considerou a representatividade do conteúdo das áreas de conhecimento (por exemplo, tópicos curriculares e língua estrangeira), pois o único critério para apresentar os itens foi sua informação. Por fim, esta pesquisa utilizou respostas simuladas que se encaixam nos modelos utilizados. Portanto nossos resultados podem estar tendenciosos para um cenário excessivamente otimista. Além disso, a simulação se baseou em estimativas do Estudo 1, em vez de parâmetros reais dos itens do Enem, o que poderia introduzir mais diferenças se o CAT fosse implementado. Considerando isso, recomendamos que trabalhos futuros preencham essas lacunas e incluam outros critérios para interromper a aplicação e selecionar itens em simulações.

Uma possível agenda de implementação de CAT no Enem precisa incluir estudos que identifiquem como a fadiga reduz a precisão da medida. Vários estudos encontraram uma relação entre a dificuldade de um item e sua posição no teste (Setzer et al., 2013; Ulitzsch et al., 2020). Não há estudos utilizando testes com características semelhantes às do Enem (ou seja, um teste de grande importância e composto de quatro medidas independentes). Primeiramente, embora este estudo tenha verificado a possibilidade de redução do Enem, não sabemos se a diminuição é suficiente para minimizar o efeito da fadiga nas respostas. Em segundo lugar, a agenda deve abordar a viabilidade da implementação de CAT em larga escala no Brasil, considerando fatores como a disponibilidade de equipamentos de informática apropriados e infraestrutura de internet confiável em todo o país.

CONCLUSÃO

O presente estudo indica que o Enem poderia ser reduzido para 20 itens, para proporções que variam de 39,8% a 94,8% dos participantes, usando o CAT. A redução foi mais eficaz nos domínios em que mais itens correspondiam à distribuição de habilidades de proficiência na população. Ao produzir itens mais fáceis em cada uma das quatro áreas, a faixa de escala em que essa redução foi possível pode ser ampliada. Recomendamos que as diretrizes para a elaboração de itens incluam aspectos objetivos dos itens relacionados à sua dificuldade. Esperamos que este estudo contribua para a melhoria das avaliações educacionais em larga escala, particularmente o Enem. Além disso, esperamos que o processo de seleção para admissão nas universidades se torne cada vez mais barato, mais eficiente e mais justo, e inclua pessoas cujo direito de acesso à educação é atualmente negado.

AGRADECIMENTOS

Agradecemos ao pesquisador do Inep Giordano Sereno pelo esclarecimento de dúvidas sobre os microdados do Enem.

DECLARAÇÃO DE CONFLITO DE INTERESSES

As opiniões expressas nesta publicação são de exclusiva e integral responsabilidade dos autores, não necessariamente expressando o ponto de vista do Inep ou do Ministério da Educação do Brasil.

REFERÊNCIAS

Ayala, R. J. de. (2009). *The theory and practice of item response theory*. The Guilford Press.

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

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. https://doi.org/10.18637/jss.v071.i05

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164-185. https://doi.org/10.1111/jedm.12009

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

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

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

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

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2010). *Guia de elaboração e revisão de itens*. MEC/Inep. http://download.inep.gov.br/outras_acoes/bni/guia/guia_elaboracao_revisao_itens_2012.pdf

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2012a). *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 (Inep). (2012b). *Guia de elaboração de itens Provinha Brasil*. MEC/Inep. http://download.inep.gov.br/educacao_basica/provinha_brasil/documentos/2012/guia_elaboracao_itens_provinha_brasil.pdf

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

Kan, A., & Bulut, O. (2015). Examining the language factor in mathematics assessments. *Journal of Education and Human Development*, *4*(1), 133-146. https://doi.org/10.15640/jehd.v4n1a13

Masri, Y. H. E., Ferrara, S., Foltz, P. W., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: The case of key stage 2 assessments. *The Curriculum Journal*, *28*(1), 59-82. https://doi.org/10.1080/09585176.2016.1232201

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. https://doi.org/10.1177/0265532217725776

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

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

Pasquali, L., & Primi, R. (2003). Fundamentos da teoria da resposta ao item – TRI. *Avaliação Psicológica*, *2*(2), 99-110.

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., Nakano, T. de C., & Wechsler, S. M. (2018). Using four-parameter item response theory to model human figure drawings. *Avaliação Psicológica*, *17*(4), 473-483. https://doi.org/10.15689/ap.2018.1704.7.07

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/

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. https://files.eric.ed.gov/fulltext/EJ1130806.pdf

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34-49. https://dx.doi.org/10.1080/08957347.2013.739453

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*(2), 1-12. http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/6416

Tillé, Y., & Matei, A. (2016). *Sampling: Survey Sampling*. https://CRAN.R-project.org/package=sampling

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A multiprocess item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement*, *80*(3), 522-547. https://doi.org/10.1177/0013164419878241

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

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

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers Theory into practice*. Springer.