Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 24/Feb/2022; 1st round notif.: 16/May/2022; New version: 26/Jun/2022; 2nd round notif.: 03/Oct/2022
Camera ready: 04/Dec/2022; Edition review: 24/Jan/2023; Available online: 09/Feb/2023; Published: 09/Feb/2023

Teste Adaptativo Multiestágio para o ENEM

Multistage Adaptive Testing in ENEM

Gabriel Couto Tabak
Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação - ICMC
ORCID: 0000-0002-8578-3686
gabrielctabak@usp.br

Jean Piton-Gonçalves
Universidade Federal de São Carlos
Departamento de Matemática
ORCID: 0000-0002-7392-2001
jpiton@ufscar.br

Thales Akira Matsumoto Ricarte
Fundação Cesgranrio
ORCID: 0000-0002-0830-5138
thalesamr@gmail.com

Mariana Curi
Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação - ICMC
ORCID: 0000-0002-7651-1064
mcuri@icmc.usp.br

Resumo

Testes internacionais de avaliação de alunos, em anos recentes, alteraram suas estruturas para implementar o formato adaptativo. O ENEM digital torna possível uma reestruturação da prova aderindo também aos testes adaptativos. Este artigo propõe um teste adaptativo multiestágio (TAM) para a prova do ENEM baseado na edição de 2019 na área de Matemática. Analisaram-se os itens do ENEM através da Teoria de Resposta ao Item, os quais foram utilizados para construir os módulos e estágios do TAM. O roteamento entre os módulos foi definido por um estudo do ponto de corte ótimo para o traço latente estimado, testado exaustivamente para encontrar o que trouxesse o melhor resultado comparado com a prova completa. Ao final, propôs-se uma arquitetura do teste multiestágio com valores de corte específicos para o roteamento. Constatou-se que o teste adaptativo reduziu em 44,4% o número de itens da prova e a estimação das habilidades foi mantida próxima da estimação com o teste completo. Notou-se também que o exame do Enem é voltado para a avaliação de níveis mais altos na escala de habilidade, tornando a estimação das habilidades prejudicada para indivíduos menos proficientes.

Palavras-Chave: *Teste Adaptativo Multiestágio; Teoria de Resposta ao Item; Teste Adaptativo Computadorizado*

Abstract

International student assessment tests in recent years have changed their structures to implement the adaptive format. The digital ENEM makes it possible to restructure the test by adhering to adaptive tests. This article proposes an adaptive multistage test (MST) for the ENEM test based on the 2019 edition in Mathematics. The ENEM items were analyzed through the Item Response Theory, and those items were used to build the MST modules and stages. The routing between the modules was defined by a study of the optimal cutoff point for the estimated latent trait, exhaustively tested to find the one that brought the best result compared to the complete exam. Ultimately, a multistage test architecture with specific cutoff values for routing was proposed. It was found that the adaptive test reduced the number of test items by 44.4%, and the estimation of skills was kept close to the estimation with the complete test. It was also noted that the ENEM is aimed at evaluating higher levels on the skill scale, making the estimation of skills impaired for less proficient individuals.

Keywords: *Multistage Adaptive Testing; Item Response Theory; Computerized Adaptive Testing*

Cite as: Tabak, G. C., Piton-Gonçalves, J., Ricarte, T. A. M. & Curi, M. (2023). Teste Adaptativo Multiestágio para o ENEM. Revista Brasileira de Informática na Educação, 31, 60-86. DOI: 10.5753/rbie.2023.2529

1 Introdução

O uso em potencial da tecnologia digital para apoiar avaliações educacionais tem crescido rapidamente nos últimos tempos (Bennett, 2015). O desenvolvimento de itens inovativos (que usam a tecnologia multimídia em seu conteúdo), a avaliação de novos construtos, a geração automática de itens (Gierl & Lai, 2013), a correção automática de textos (Dikli, 2006), a integração de avaliação e instrução são apenas exemplos dessa realidade. A evolução tecnológica também pode aprimorar avaliações educacionais em larga escala. Neste trabalho, discutem-se o estado da arte de testes adaptativos.

Historicamente, os denominados Testes Lineares (do inglês *Linear Tests*) vem sendo aplicados ao longo do tempo enquanto instrumento padrão para a avaliação educacional, sendo compostos por itens dissertativos ou de resposta objetiva (por exemplo, itens de múltipla escolha com apenas uma resposta correta). De acordo com von Davier (2017), "um teste linear é usualmente administrado no papel e comumente referenciado como testes de *lápis e papel*". Nele, todos os examinados respondem aos mesmos itens, geralmente abrangendo diferentes níveis de dificuldade. A maioria dos testes educacionais é enquadrada nesse modelo.

A evolução tecnológica levou ao surgimento dos Testes Baseados em Computador ou, simplesmente, Testes Computadorizados (TC) (do inglês *Computer-Based Testing*), cujos itens são administrados de maneira automatizada. O nível de automatização de um TC dependerá dos objetivos educacionais, conteúdos avaliados e tecnologias disponíveis para a sua implementação.

Tal avanço abriu espaço para maior dinâmica do formato dos itens, viabilizando itens multimídia ou mesmo simuladores de objetos virtuais, expandindo o nível de interação do examinado com o teste (Bartram, 2006). Vale notar que testes compostos por itens de múltipla escolha possibilitam maior autonomia do sistema em relação àqueles compostos por itens dissertativos. As formas de administrar e pontuar um TC podem seguir desde abordagens lineares até abordagens adaptativas.

Num extremo, faz-se apenas transposição do modelo linear (lápis e papel) para o computador, em que a nota final do teste é computada pela soma das pontuações de cada item. No outro extremo, os métodos adaptativos adotam ferramentas computacionais e metodologias Psicométricas modernas para o desenvolvimento do processo avaliativo, em suas diversas etapas (de criação, administração, pontuação, interpretação e divulgação dos resultados). Em casos que se tem um banco de itens bons, os métodos adaptativos podem se mostrar muito mais eficientes (Wainer, Kaplan, & Lewis 1992).

Nesse contexto, o principal destaque são os Testes Adaptativos Computadorizados (TAC) (do inglês *Computerized Adaptive Test*), que utilizam as respostas fornecidas pelo estudante em itens apresentados anteriormente no teste para selecionar e apresentar um novo item. A adaptabilidade pode ser baseada na Teoria de Resposta ao Item (TRI) (do inglês *Item Response Theory*), uma metodologia psicométrica bastante adotada em cenários nacionais e internacionais, mesmo em avaliações "lápis e papel". A TRI propõe uma modelagem estatística para a estimação da habilidade do examinado (Andrade et al., 2000), diferenciando os itens respondidos de acordo com sua dificuldade e capacidade de discriminação, por exemplo.

O principal objetivo do TAC é produzir um teste eficiente, no sentido de estimar bem a habilidade do examinado, com um número menor de itens do que os testes lineares (ou convencionais). Para tanto, deve contar com um banco de itens diversificado em níveis de dificuldade e de qualidade para avaliar o construto de interesse. Assim, o algoritmo matemático, baseado na TRI, tem ampla possibilidade para seleção de itens adequados ao nível de conhecimento demonstrado pelo examinado durante a avaliação (Yan, 2014). Recentemente, para

diferenciar da proposta apresentada no parágrafo a seguir, este formato tem sido referido como Teste Adaptativo em nível de item (TAC-I) (em inglês *item-level Adaptive Tests*). Neste trabalho, adota-se a nomenclatura TAC representante de testes adaptativos em geral, seja a nível de item, seja a nível de módulo, apresentado a seguir.

Outro formato de avaliação adaptativa é o Teste Adaptativo Multiestágio (TAM) (do inglês *Multistage Adaptive Test*). Surgiu como uma abordagem mais viável do ponto de vista prático, apesar de perder ligeiramente na precisão das estimativas (Yan, 2014). O TAM apresenta conjuntos de itens (chamados de módulo) para os examinados. Uma vez que um examinado termina de responder um módulo ele passa para o próximo estágio, em que responderá um novo conjunto de itens até atingir um critério de parada.

Diferentemente do TAC-I, no TAM, os examinados podem navegar entre os itens de um determinado módulo (blocos de 2 ou mais itens), diminuindo substancialmente a ansiedade e o estresse (Zheng, 2012), uma vez que se pode revisar as respostas enquanto um novo módulo não for apresentado. Essencialmente, em um TAM, a adaptabilidade se dá a nível de módulo e não de item, ou seja, dentro de um determinado módulo, a avaliação funciona como um teste linear. Além disso, o balanceamento de conteúdo no teste é mais controlável, e não tão automatizada quanto nos TAC-I, facilitando um aspecto importante do ponto de vista pedagógico.

Na prática, desde 2011, a *Educational Testing Service* (ETS) possui uma versão do *Graduate Record Examination* (GRE) computadorizada baseada em TAM (Yan, 2014). Um ponto importante para o examinado é a possibilidade de revisão das respostas dos itens dentro do módulo corrente, antes de submetê-lo e prosseguir/terminar com o teste. Porém, após a submissão não é possível retornar a módulos anteriores.

Na direção do TAM, o *Programme for International Student Assessment* (PISA) é um exame que avalia estudantes de 15 anos em 79 países e economias, incluindo o Brasil. Em 2015, o exame do PISA migrou do formato "lápis e papel" para um TAC, focando nas estimativas populacionais e na evolução do desempenho dos estudantes ao longo do tempo. Em 2018, houve uma nova mudança significativa, pois passou a ser um TAM. Detalhes metodológicos são encontrados em Yamamoto (2018) e OCDE (2019). Além do PISA, a OCDE aplica o *Programme for International Assessment of Adult Competencies* (PIAAC) que avalia adultos entre 16 a 65 anos de idade, em 40 países (o Brasil não participa atualmente), nos domínios de leitura, escrita e aritmética, bem como suas capacidades na resolução de problemas em contextos tecnológicos. Tal exame possui uma versão computadorizada baseada em TAM (Kirsch, 2017).

No Brasil, o Exame Nacional do Ensino Médio (ENEM) mostra-se como o mais importante e a maior avaliação em larga escala nacional, sendo o segundo maior do mundo e perde apenas para o Gaokao na China. Tradicionalmente, o ENEM vem sendo aplicado via lápis e papel. Porém, em 20 de abril de 2020, tornou-se pública a realização do ENEM Digital, por meio do edital n° 34 do Diário Oficial da União, com aplicação de sua segunda edição em novembro de 2021 (INEP, 2019). Num cenário internacional com importantes avaliações no formato TAM somada à realidade de um ENEM digital, cabe estudar a viabilidade metodológica de um Teste Adaptativo Multiestágio em provas do ENEM Digital.

Na literatura, alguns artigos exploram Testes Adaptativos Computadorizados para as provas do ENEM (Spenassato, et al. 2016; Jatobá, et al. 2018; 2019; Dias, 2019), entretanto, ainda não foi abordado o teste multiestágio. Esta lacuna é importante de ser preenchida, aplicando ao ENEM as metodologias mais atuais da Psicometria, sob a luz das grandes avaliações internacionais. Este artigo tem como objetivo propor um teste adaptativo multiestágio (TAM) para o ENEM. Apresenta-se um método para desenhar um TAM e avalia-se por meio de simulações a sua precisão para estimar as habilidades latentes dos respondentes de uma das provas do Enem. A

questão de pesquisa é se o TAM estima adequadamente as habilidades de candidatos ao Enem e concomitantemente reduz o número de questões, tornando o teste mais curto e tão eficaz quanto o tradicional, em lápis e papel. Os resultados obtidos corroboram o uso do TAM como uma alternativa relevante ao Enem tradicional.

Com o TAM, é possível reduzir o número de itens respondidos pelos indivíduos, concentrando-se em itens mais condizentes com a habilidade de cada examinado, com total controle sobre o conteúdo cobrado na prova. Mantendo o tempo de prova igual ao do teste linear, cada examinado pode dedicar mais tempo a itens adequados a seu nível de habilidade. Além de que itens muito fáceis ou muito difíceis para o examinado não aparecem em seu teste.

Esse artigo encontra-se estruturado em 6 Seções. A Seção 2 traz os trabalhos relacionados da literatura que abordam o tema de testes adaptativos no ENEM. Na Seção 3, apresentam-se a Fundamentação Teórica, com uma introdução sobre o ENEM, e os temas: TAC, TRI, TAC-I e TAM. Na Seção 4, introduz-se os Materiais e Métodos utilizados na pesquisa, trazendo informações sobre os tipos de prova do ENEM, o processo de amostragem, as decisões de montagem dos módulos e o processo de roteamento. Na Seção 5, descrevem-se os Resultados, análise de itens, o TAM proposto e a comparação entre o TAM e o teste completo. Na Seção 6, tem-se as considerações finais.

2. Trabalhos Relacionados

O trabalho de Spenassato et al. (2016) apresentou as vantagens de um teste adaptativo do teste de Matemática e suas Tecnologias do Enem edição 2012, quando comparado com a versão via lápis e papel. Partindo de um banco de itens composto por 45 itens calibrados, os resultados simulados mostram que o teste adaptativo poderia ser reduzido, no mínimo, 26,6% sem perda significativa de precisão das habilidades estimadas.

Os trabalhos de Jatobá et al. (2018) e Jatobá (2019), essencialmente, desenvolveram a abordagem personALIzed Computerized Adaptive Testing (ALICAT), que personaliza o processo de seleção de itens em CAT. Um estudo de caso foi aplicado na prova de Matemática e suas tecnologias do ENEM de 2012 e indicou que a regra de seleção de Kullback-Leibler, com distribuição a posteriori, apresentou melhores resultados na estimativa das habilidades dos respondentes em relação a outros métodos de seleção. Neste estudo de caso, o teste foi reduzido em 53,3% em relação ao teste completo (45 itens), sem perda significativa na estimativa das habilidades. O desempenho do ALICAT mostrou que foram necessários apenas 21 itens para o teste adaptativo, sem perda significativa na estimativa das habilidades.

A dissertação de Dias (2019) focou na calibração de itens das provas de Matemática do ENEM referente a 2017, propondo uma abordagem Bayesiana para estimar a habilidade considerando a incerteza quanto aos parâmetros dos itens calibrados. Neste contexto, a autora propôs uma extensão dessa abordagem para o formato de testes adaptativos. De acordo com os estudos de simulação, o teste adaptativo pode estimar satisfatória e eficientemente as habilidades necessitando menos itens respondidos na prova do ENEM.

Publicações e reportagens sobre o ENEM (de Macedo, 2021, Avellar, 2012) apontam a testagem adaptativa como um futuro do exame, seja para a avaliação em si, seja para simulados da prova, evidenciando a importância de mais estudos e avanços nesse sentido.

Tais trabalhos elucidam e atestam as vantagens descritas na literatura de testes adaptativos, a nível de item, para o contexto da prova do ENEM. O TAM é o que há de mais atual na literatura de testes adaptativos informatizados, equilibrando as vantagens dos testes adaptativos a nível de item com a praticidade dos testes lineares (Hendrickson, 2007). Este trabalho objetiva avançar a

pesquisa sobre o ENEM nas metodologias mais modernas de testes adaptativos, atualmente o TAM.

3. Fundamentação Teórica

3.1 Exame Nacional do Ensino Médio

O ENEM foi instituído em 1998 e a partir de 2009 passou a ser utilizado como exame de seleção para o ingresso no ensino superior do Brasil, permitindo o acesso às instituições públicas e privadas por meio do Sistema de Seleção Unificada (Sisu). Além disso, os estudantes podem pleitear financiamento estudantil em programas do governo, como o Fundo de Financiamento Estudantil (Fies), por exemplo. As notas do ENEM estão sendo utilizadas internacionalmente, como é o caso de 50 instituições de ensino de Portugal (INEP, n.d.).

Indicado àqueles que concluíram o Ensino Médio, na edição de 2019 foram 5 milhões de inscritos interessados em cerca de 500 instituições de ensino. A nota do ENEM pode ser utilizada na totalidade ou complementar, dependendo dos objetivos da instituição.

Do ponto de vista de estruturação do exame, são 180 itens de múltipla escolha distribuídos em 4 áreas do conhecimento, além de uma redação. É um exame aplicado via papel e caneta, com um tempo mínimo de permanência na sala de prova de duas horas. A resposta ao item é preenchida em uma ficha nominal, que posteriormente é lida oticamente.

Em termos de aplicação, ele ocorre em dois domingos subsequentes e as provas são distribuídas da seguinte forma:

- **Primeiro Domingo**: Com cinco horas e meia de duração, a prova é composta por (i) uma redação (que não é calculada pela TRI, de correção manual, realizada por cerca de 5000 avaliadores), (ii) 45 itens de múltipla escolha de Ciências Humanas e suas tecnologias e (iii) 45 itens de múltipla escolha de Linguagens, Códigos e suas tecnologias.
- **Segundo Domingo**: Com cinco horas de duração, a prova é composta por (i) 45 itens de múltipla escolha de Ciências da Natureza e suas tecnologias e (ii) 45 itens de múltipla escolha de Matemática e suas tecnologias.

3.2 Testes Adaptativos Computadorizados

O psicólogo e educador francês Alfred Binet (1857 -- 1911) desenvolveu o teste adaptativo de Binet (em inglês *Binet-type adaptive test*), que é um teste de inteligência com diferentes níveis de dificuldade e parte do princípio que, se os itens de um certo nível de dificuldade forem respondidos corretamente, eleva-se o nível; caso contrário, desce-se o nível (Weiss, 1985).

Após a introdução do teste de Binet em 1905, a abordagem adaptativa foi retomada na década de 50, com o Teste Adaptativo de Dois Estágios (em inglês *Two-Stage Adaptive Test*), por meio dos estudos de Angoff (1958), que compararam um teste de dois estágios com testes convencionais em habilidades verbais e matemática. Tal teste divide-se em dois sub-testes de menor dificuldade e maior dificuldade. Segundo as respostas obtidas no primeiro sub-teste, selecionam-se os itens do segundo (Weiss, 1985). Um teste multiestágio é uma múltipla extensão de dois estágios em versão computacional.

Com os avanços em *software* e *hardware* na década de 60, surgem trabalhos como de (Reckase, 1974), que implementa computacionalmente um teste adaptativo. Nessa direção, os TACs permitem a aplicação de testes mais flexíveis e adaptáveis, com o objetivo de maximizar a

acurácia do teste (Weiss, 1985) e com a vantagem de administrar um número menor de itens do que os testes via lápis e papel.

Estruturalmente um TAC seleciona os itens dinamicamente conforme o examinado responde um item ou um conjunto deles, resultando em um teste adaptável e individualizado. Ou seja, cada um poderá responder a um conjunto diferente de itens em quantidade e grau de dificuldade. Os principais componentes de um TAC são: (i) um banco de itens pré-calibrado, (ii) um critério para iniciar o teste, (iii) um método de seleção de itens ou de blocos de itens e (iv) um critério de parada do teste.

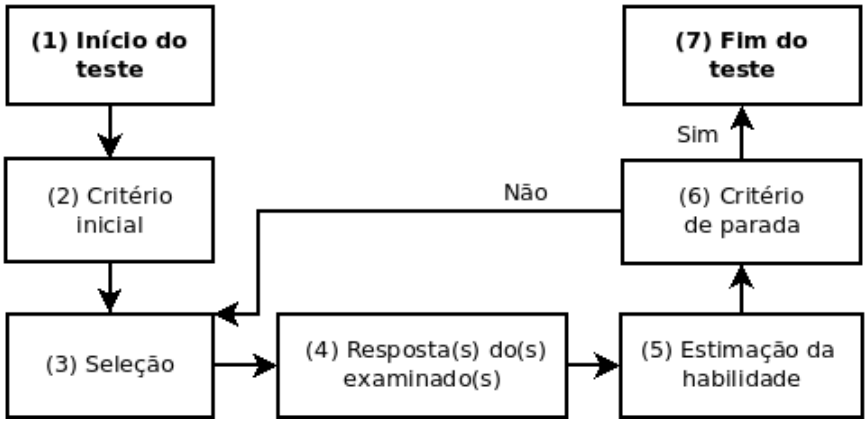

Figura 1: Fluxograma funcionamento TAC.

O fluxograma da Figura 1 mostra o esquema geral de um TAC baseado na TRI, que é executado da seguinte forma:

(1) Inicia-se o teste aplicando um *setup* inicial (uma configuração de início de teste, por exemplo, dificuldade(s) dos itens iniciais, frequência de exposição dos itens, escolha do modelo, banco de itens, etc.).
(2) Aplica-se um critério para seleção para a escolha do primeiro item a ser administrado ou um bloco de itens.
(3) Seleciona-se o próximo o item ou um bloco deles, de acordo com algum critério preestabelecido.
(4) São obtidas a(s) resposta(s) ao(s) item/itens do examinando.
(5) Estima-se a habilidade com base nas respostas dadas.
(6) Se algum critério de parada for satisfeito, vá para (7). Senão, volte para (3).
(7) Fim de teste.

Quanto ao nível de adaptação, um TAC pode ser categorizado em:

- **Teste Adaptativo em nível de item.** No TAC-I a adaptação ocorre item a item, ou seja, quando o examinado responde a um item, o próximo é selecionado imediatamente (Yamamoto, 2018), produzindo um teste personalizado, acurado, válido e fidedigno, com a vantagem de ser mais curto do que um teste linear (Piton, 2020). Um TAC-I pode manter níveis adequados de precisão e administrando um número de itens menor que *50%* em relação à um teste linear (Weiss, 1984; Piton, 2020), diminuindo substancialmente a fadiga causada por longos testes.

- **Teste Adaptativo Multiestágio.** No TAM a adaptação ocorre em blocos de itens, denominados de módulos (Yamamoto, 2018). O examinado responde a um teste linear em cada módulo, porém cada módulo é selecionado de forma adaptativa. De acordo com von Davier (2017), para se adaptar ao nível de habilidade do examinado durante o teste, utiliza-se uma estrutura modular baseada em regras previamente definidas por especialistas humanos.

3.3 Teoria de Resposta ao Item

A TRI surgiu na década de 50, com o trabalho de Lord (1953) e caracteriza-se por um conjunto de modelos que descrevem a relação de um (ou mais) traço latente (nível de conhecimento, proficiência, habilidade, intensidade de um processo psíquico ou gravidade de uma doença) e as respostas de um indivíduo a itens de múltipla escolha. Vários modelos da TRI são propostos na literatura, dependendo do número de traços latentes, do tipo de item (com 2 ou mais de 2 categorias de resposta, ordinais ou nominais) e do número de características dos itens que serão levadas em consideração no teste.

Duas vantagens da TRI quando aplicadas em conjunto com os TAC seriam: (i) o conhecimento dos indivíduos estimados sob seus modelos é comparável, mesmo que estes tenham realizado provas com itens diferentes e (ii) o traço latente do indivíduo e os níveis de dificuldade dos itens estão na mesma escala de medida.

A seguir, será descrito o modelo de 3 Parâmetros da TRI. Este modelo pode ser usado para a implementação de um TAC e é adequado para itens dicotômicos (resposta correta ou incorreta, por exemplo) e que considera apenas um traço latente, ou seja, unidimensional (conhecimento em Matemática, por exemplo).

O modelo é definido pela função que relaciona a probabilidade de resposta correta a um determinado item e o traço latente, que corresponde ao conhecimento do indivíduo. Nesse modelo, essa função é dada pela Equação (1) e representada pela curva característica do item (CCI), na Figura 2. X_i representa a resposta observada ao item i, que pode ser igual a 0 (se o indivíduo erra a resposta) ou 1 (se acerta). θ representa o traço latente do indivíduo (seu nível de conhecimento, por exemplo). a_i é o parâmetro de discriminação do item i. b_i é o parâmetro de dificuldade do item i. c_i é o parâmetro de acerto ao acaso do item i. A função é crescente com o traço latente, tende a zero quando o traço latente tende a menos infinito e a um, quando o traço latente tende a mais infinito.

A inflexão da curva (ponto em que a curvatura se altera) está exatamente na probabilidade de resposta correta igual a $(1 + c)/2$. Este ponto tem valor na abscissa (eixo x, que representa o traço latente) exatamente igual ao parâmetro b_i do modelo. A interpretação desse parâmetro é "o valor do traço latente necessário para que o indivíduo tenha probabilidade $(1 + c)/2$ de acertar a resposta do respectivo item". Essa peculiaridade do modelo faz com que a escala de dificuldade dos itens e a escala de conhecimento do indivíduo sejam as mesmas, característica importante num TAC.

No TAC, a proximidade entre o valor estimado de θ e a dificuldade do item (b_i) é usada para a seleção do item a ser respondido pelo indivíduo. A característica do modelo de ter parâmetros distintos representando o conhecimento do indivíduo e as dificuldades dos itens possibilita que estas sejam consideradas para a estimação de seu traço latente (conhecimento).

O parâmetro de discriminação, a_i, é considerado para cada item. A CCI resultante permite que a velocidade de crescimento da probabilidade de acerto mude para cada item, dependendo do valor de a_i. Itens com valores maiores de a_i tem um crescimento maior na probabilidade de acerto, diferenciando mais os níveis de conhecimento em torno da dificuldade b_i.

Neste modelo, a probabilidade de acerto do item quando o indivíduo tem traço latente baixo não tende a zero, mas sim a uma assíntota inferior com valor igual ao parâmetro c_i. Na área educacional, para itens de múltipla escolha, mesmo que o indivíduo não tenha nenhum conhecimento do construto em questão, a possibilidade de acertar a resposta escolhendo-se aleatoriamente uma das alternativas de resposta existe. Essa interpretação justifica a nomenclatura do parâmetro c. Vale destacar que este é o modelo adotado na avaliação do ENEM e este foi o modelo utilizado neste artigo.

$$P(X_i|\theta, a_i, b_i, c_i) = c_i + \frac{1-c_i}{1+e^{-a_i(\theta-b_i)}} \qquad (1)$$

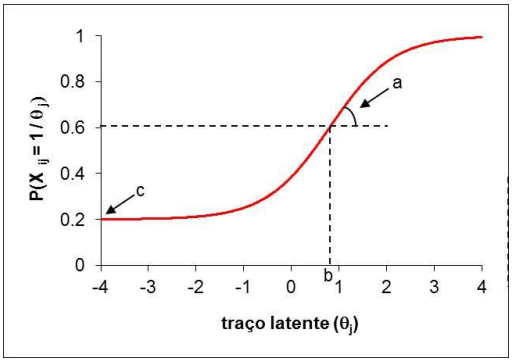

Figura 2: CCI de três itens sob o modelo logístico de 3 parâmetros. (a=1,5; b=0; c=0,2).

3.4 Teste Adaptativo em nível de item

O TAC-I é um teste adaptativo computadorizado que usa um algoritmo para administrar os itens do teste (Yan, 2014). O indivíduo recebe o teste, e responde os itens conforme eles aparecem. A cada item respondido, estima-se a habilidade do examinado, e então o próximo item é escolhido. Uma vez que algum critério de parada é atingido, o teste é finalizado e a habilidade final do examinado é calculada.

O TAC-I, quando tem um banco de itens adequado e diverso, consegue ser mais eficiente que o teste linear. Testes lineares muitas vezes podem ressaltar algum nível de habilidade em particular, isso implica que os examinados com habilidades próximas de tal nível serão melhores estimados, em detrimento do restante da população. Enquanto isso, no TAC-I, a estimação da habilidade é adaptada de acordo com o indivíduo, retornando resultados mais uniformemente precisos ao longo das possíveis habilidades da população (Yan, 2014).

Existem vários formatos possíveis para a implementação do TAC-I, em que várias decisões devem ser tomadas.

- Escolha do item inicial (pode ser o mesmo para todos, ou podem ser diferentes)
- Método de Estimação das habilidades.
- Calibração dos itens.
- Critérios de seleção dos itens.
- Critério de parada do teste.

No TAC-I, os itens são escolhidos adaptados ao traço latente do indivíduo. Entretanto, isso pode causar um problema no balanceamento do conteúdo dos itens, conforme o examinado os responde, pode ser que ele tenha um comportamento que leve a responder itens repetidos do mesmo conteúdo, ou pode ser que um determinado conteúdo sequer seja apresentado em sua prova. Para contornar esse problema, seria ideal ter um banco de itens que satisfaça o propósito do teste, com conteúdo variado e com itens com níveis de dificuldades abrangentes, além de

critérios e algoritmos de seleção que incorporem restrições baseadas em especificações pedagógicas (como conteúdo e exposição dos itens) (Silva et al., 2019).

No TAC-I, os itens não podem ser revisados uma vez que respondidos. Caso houvesse a possibilidade de revisar os itens, o examinado poderia alterar o traço latente estimado em determinado ponto, o que atrapalha na ordem de aparição dos itens. O examinado não tem conhecimento dos próximos itens que podem aparecer, e ele pode não saber a quantidade de itens que responderão ao longo do teste. Os fatores apresentados podem gerar níveis mais altos de estresse e ansiedade, quando se trata de avaliações educacionais formais, principalmente quando ligados às crianças e jovens (Fritts et al., 2010).

O TAC-I é uma possibilidade a ser explorada, entretanto, neste artigo, optou-se por não utilizar esse método dado suas desvantagens. No interesse de um aprofundamento maior em TAC-I, recomenda-se a leitura do trabalho do Van der Linden (2010).

3.5 Testes Adaptativos Multiestágio

Análogo ao TAC-I, tem-se o TAM, que também é um teste adaptativo. Entretanto, no TAM, os itens são agrupados em módulos, e a etapa adaptativa ocorre após cada módulo ser respondido, ou seja, a cada módulo e não a cada item (Ricarte et al., 2018).

O TAM, assim como no TAC-I, contém uma série de decisões para sua montagem. Neste artigo, fez-se um TAM fixo, em que os módulos são montados antes da aplicação do teste. O fato de ter módulos construídos antes da aplicação do teste ajuda no controle do balanceamento do conteúdo. Pode-se adicionar itens que envolvam o conteúdo esperado em todos os módulos, adequando o item à dificuldade do módulo (Yan, 2014).

Um possível design para o TAM é a utilização de estágios. Nesse caso, o teste é dividido em um número específico de estágios e cada estágio contém módulos de dificuldade diferentes. Os examinados passarão por todos os estágios, e em cada um terão apenas um módulo para responder. O primeiro estágio serve para o roteamento dos examinados, nele têm-se itens de dificuldades médias, em que se tenta fazer uma estimação aproximada da habilidade do aluno. Caso o aluno tenha um bom (mal) resultado, ele será apresentado a um módulo mais difícil (fácil) no estágio seguinte. E mantém-se essa lógica até o último estágio.

Dentro de um estágio, o examinado tem liberdade para revisar os itens e alterar as alternativas. Entretanto, quando passa para o estágio seguinte, os itens dos estágios anteriores deixam de ser acessíveis.

Quando se tem um banco de itens grande, e variado em dificuldade, abre-se a oportunidade de criar vários painéis, em que cada um contém uma montagem do TAM diferente, considerando designs diferentes. A vantagem de se ter vários painéis é que essa prova pode ser aplicada em diferentes contextos, com objetivos diferentes, bastaria mudar o painel utilizado.

Na Figura 3 tem-se um exemplo de TAM, com apenas um painel. Nesse caso, são 3 estágios, em que o primeiro estágio contém um módulo de roteamento, o segundo estágio contém 3 módulos de dificuldades fácil, média e difícil, e o terceiro estágio contém 3 módulos também com as três dificuldades. O examinado ao passar pelo módulo de roteamento tem sua habilidade estimada e então recebe um módulo condizente no segundo estágio, e sua habilidade é estimada novamente e o módulo condizente no terceiro estágio é apresentado.

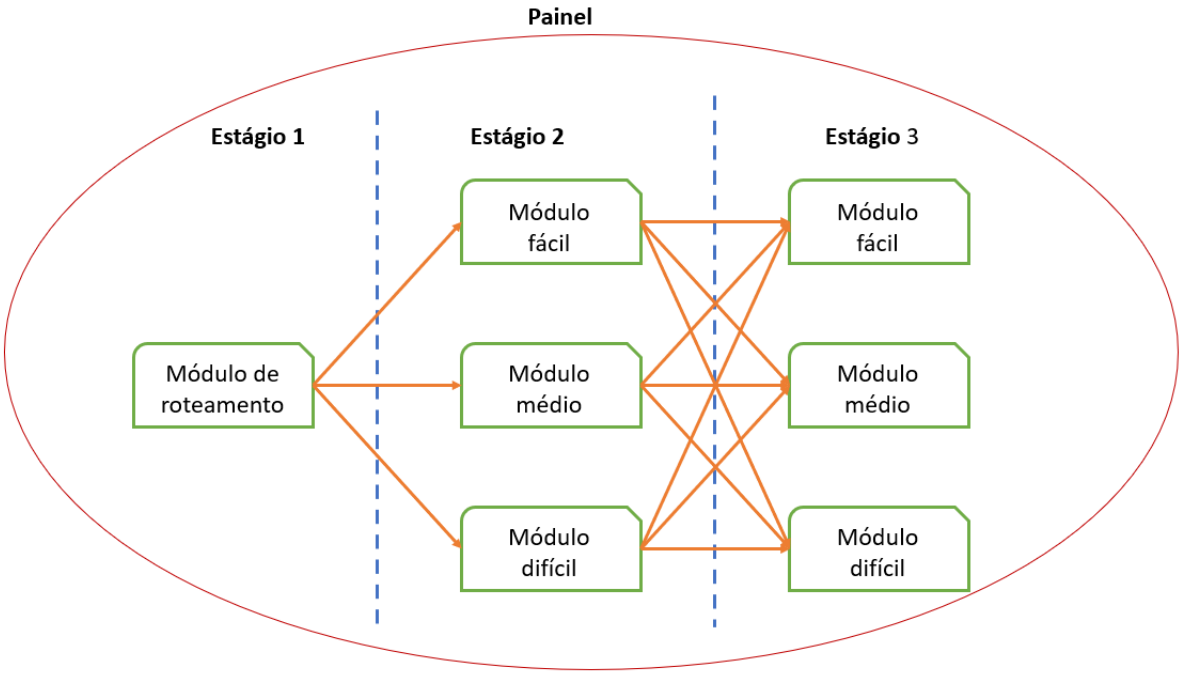

Figura 3: Exemplo de um Painel de um TAM.

Algumas decisões que precisam ser tomadas para a elaboração de um TAM, estão descritas a seguir.

- Número de estágios, número de módulos por estágios (arquitetura do TAM).
- Número de painéis.
- Dificuldade de cada módulo.
- Separação dos itens dentro dos módulos.
- Calibração dos itens.
- Método de estimação das habilidades.
- Método de seleção dos módulos (Roteamento).

O roteamento de um estágio para o outro pode ser feito de diversas formas. Uma vez determinada a habilidade do examinado em um módulo, o módulo seguinte a ser selecionado precisa ser adequado ao indivíduo. Explorou-se uma metodologia de habilidade de corte, em que se os indivíduos alcançam uma habilidade corte tal, são apresentados a um módulo mais difícil, caso não consigam, são apresentados a um módulo mais fácil.

4. Materiais e Métodos

Computacionalmente, utilizou-se a linguagem R, na versão 4.0.3, no Sistema Operacional Windows. As principais bibliotecas utilizadas para a análise da TRI dos itens foram a mirt (Chalmers, 2012) e a mirtcat (Chalmers, 2016). Algumas rotinas da biblioteca CTT (Willse et al., 2008) foram incluídas para análise dos itens. E a biblioteca ggplot2 (Wickham, 2016) foi a principal usada na construção dos gráficos.

A simulação do TAM para o ENEM baseou-se na prova de 2019, que seria a prova mais recente com os microdados disponíveis. Sendo a versão mais atualizada dos microdados, durante o período do estudo, feita em 20/05/2021 (esta foi a considerada no presente artigo) (INEP, n.d.).

Neste trabalho, optou-se por usar apenas uma área da prova do ENEM, a de Matemática. A escolha de Matemática se deve ao fato de outros autores também utilizarem tal disciplina em testes adaptativos, como Jatobá (2019), Spenassato (2016). Além disso, outros estudos (Palermo, 2014; Barbosa, 2005) indicam que Matemática é a disciplina mais afetada pelas características das escolas. Portanto, a prova de Matemática seria mais pertinente na análise.

No ENEM de 2019, teve-se um pouco mais de 5 milhões de candidatos, entretanto, nem todos participaram dos dois dias da prova. Na análise, removeu-se os participantes que não estiveram presente no dia da prova de matemática, dado que tais participantes não auxiliam nas estimativas dos itens nem na estimativa de sua habilidade em matemática (não possuem um vetor de resposta). Dessa forma, 3.707.811 participantes foram considerados.

As versões (tipos) de provas são divididas em cores (e códigos), e em cada tipo de prova os itens são dispostos em ordens diferentes. Na Tabela 1, tem-se o código do tipo de prova, sua descrição e a quantidade de participantes que a fizeram. Percebe-se que as provas de código 515 a 518 estão concentradas a maior parte dos examinados. E nelas, têm-se os mesmos 45 itens, apenas embaralhados em ordem diferentes.

Tabela 1: Descrição tipos de prova de matemática.

Código	Descrição	Aplicações
515	Azul	924477
516	Amarelo	925550
517	Rosa	924231
518	Cinza	933553
522	Laranja - Adaptada Ledor	507
526	Verde - Videoprova - Libras	2089
555	Amarela (Reaplicação)	6
556	Cinza (Reaplicação)	5
557	Azul (Reaplicação)	6
558	Rosa (Reaplicação)	9

Já em relação as provas de código 555 a 558 (provas de reaplicação), têm-se apenas 26 examinados. A reaplicação do ENEM é uma prova com itens distintos dos itens originalmente aplicados. Por ter um número muito pequeno de examinados, e itens diferentes da prova original, não seria possível uma estimação boa e precisa desses novos itens. Portanto, optou-se por excluir as reaplicações da análise.

Por fim, têm-se as provas de código 522 e 526, essas são as provas adaptadas (INEP, n.d.). Optou-se por excluí-las também da análise. Existem poucos examinados que se encaixaram nessas categorias em relação ao total de examinados. Observou-se que a prova laranja possuía três itens diferentes das demais, o que, novamente, tornaria inviável a estimação desses itens. Dessa maneira, a análise foi feita com apenas quatro tipos de provas, cada uma com sua disposição dos itens, e com mais de três milhões de respondentes.

4.1 Amostra

Dado o número expressivo de examinados, optou-se por retirar amostras representativas da população. Para a amostragem ser possível, foi necessário fazer a normalização da base. Isto é, colocar todos os itens dos tipos de provas escolhidos numa mesma ordem.

Cada item presente na prova contém um código único para identificá-lo. Caso esse item apareça em mais de um tipo de prova, ele sempre terá o mesmo código. O código de item é um número inteiro. A informação sobre os códigos de itens e os tipos de prova está contida na planilha sobre as provas dos microdados.

A normalização da base consistiu em agrupar os itens conforme seu tipo de prova (nos 4 tipos escolhidos), e, em seguida, ordenar crescentemente os itens de acordo com seu código. Dessa forma, têm-se 4 sequências de números indicando qual a ordenação que cada tipo de prova deve ter. A classificação dos itens (como primeiro, segundo, e assim por diante) utilizada no trabalho foi pela ordem crescente do código de item. Ou seja, o primeiro item nas análises será o item de menor código, e o último item será o item de maior código.

Na Tabela 2 tem-se a ordenação dos 6 primeiros itens da prova de matemática do ENEM de 2019. Pode-se perceber que o item 160 na prova Azul é equivalente ao item 165 na prova Amarela e seu código é 8386, sendo este item o correspondente ao primeiro item da análise do trabalho.

Tabela 2: Ordenação dos itens.

Prova Azul	Prova Amarela	Prova Rosa	Prova Cinza	Código Item	Número do item
160	165	171	155	8368	1
168	173	179	163	8401	2
172	138	147	167	8442	3
137	140	143	152	9779	4
166	171	177	161	10360	5
161	166	172	156	13303	6

Realizaram-se três amostragens distintas com o auxílio da função *sample* do R. A primeira amostra possuía 50 mil examinados e foi utilizada na calibração (estimação dos parâmetros dos) dos itens e montagem do TAM. A segunda amostra possuía 10 mil examinados, também foi gerada a partir da função *sample,* considerando todas as observações. Essa amostra foi utilizada para escolher os valores de corte do roteamento e comparar o teste completo com o TAM. Por fim, a terceira amostra foi selecionada com 20 mil examinados. Dessa amostra, selecionou-se os examinados com as *25%* maiores habilidades, reduzindo-se para 5 mil examinados. Tal amostra foi importante para comparar o teste completo com o TAM. O processo de amostragem está disponível no GitHub do primeiro autor[1], por terem sido especificados sementes no processo de amostragem, todas as amostras podem ser reproduzidas.

Para garantir que os resultados de calibração dos itens, feitos com a primeira amostra, são confiáveis, foram feitas replicações desse estudo com diferentes amostragens. Vinte testes com amostras de tamanho diferentes (15 delas com 50 mil examinados, e as outras com 10 mil) foram repetidos e as estimativas dos parâmetros dos itens foram calculadas em cada teste. Também se fez um teste considerando apenas os examinados que fizeram a prova Azul. Os testes consistiram em calibrar os itens. Desses resultados, foi possível perceber que as estimativas foram concordantes e similares entre si.

4.2 Construção dos módulos

A primeira amostra então foi utilizada para a calibração dos itens de acordo com a TRI. Pelo método da Máxima Verossimilhança Marginal, com o algoritmo iterativo EM, foi feita a estimação dos parâmetros dos itens (dificuldade, discriminação e acerto ao acaso). Esse procedimento já está implementado na biblioteca mirt (Chalmers, 2012). Os parâmetros dos itens são importantes para a construção dos módulos e estágios.

[1] https://github.com/GabrielTabak/TAM-ENEM

Na literatura atual, um dos métodos mais importantes e utilizados para a construção dos módulos e estágios de um TAM é o *Automated Test Assembly* (ATA). Em que esse software segue algoritmos e heurísticas para satisfazer condições impostas pelos programadores e alcançar objetivos estatísticos definidos (Van Der Linden et al., 2010). Entretanto, neste trabalho apenas um painel será montado, sendo o uso do ATA descartado.

A arquitetura do MST foi definida com 3 estágios. Sendo o primeiro estágio com apenas um módulo, o segundo estágio com dois módulos e o terceiro estágio com três módulos.

O critério de seleção dos itens adotado para a formação dos módulos foi a informação do Item (Andrade et al., 2000). Itens com maiores informações para um dado traço latente serão escolhidos para entrarem nos módulos. A informação do Item tem uma forte influência da discriminação do item (é diretamente proporcional ao *a*). Isso implica que os itens ruins de discriminação baixa, praticamente não serão escolhidos nos módulos (sua informação sempre será menor que a de outros itens).

O processo de montagem dos módulos segue o seguinte algoritmo:

- Encontra-se qual traço latente que maximiza a informação de cada item.
- Faz-se uma ordenação desses traços latentes e os percentis *25%, 33%, 50%, 67%* e *75%* são encontrados.
- Os módulos do terceiro estágio são divididos em fácil, médio e difícil, com, respectivamente, os percentis *25%, 50%* e *75%* representando o traço latente desses módulos.
- Encontram-se os 5 itens com maior informação para cada traço latente dos módulos do terceiro estágio. Esses itens passam a formar cada um desses módulos. Respeitou-se um máximo de dois itens em comum entre esses módulos.
- Os itens escolhidos são removidos para a próxima etapa.
- Os módulos do segundo estágio seguem uma lógica análoga ao do terceiro estágio. Porém, utilizam os traços latentes equivalentes aos percentis *33%* e *67%*.
- Encontram-se os 10 itens com maiores informações para os traços latentes dos módulos do segundo estágio, permitindo um máximo de 5 itens em comum entre os módulos.
- Os itens escolhidos são removidos para a próxima etapa.
- Por fim, monta-se o módulo do estágio 1, o módulo de roteamento. Foram escolhidos os dois itens mais informativos para cada um dos 5 percentis.

A Figura 4 mostra um quadro esquemático das etapas de construção do módulo. Os itens escolhidos são aqueles que maximizam a informação do Item. Para garantir o limite de itens em comum, escolhe-se o item em comum menos informativo de um módulo e substitui-se tal item pelo próximo mais informativo dentro do banco de itens. Em seguida, faz-se o mesmo procedimento para o outro módulo. Esse procedimento se repete até atingir o limite de itens em comum.

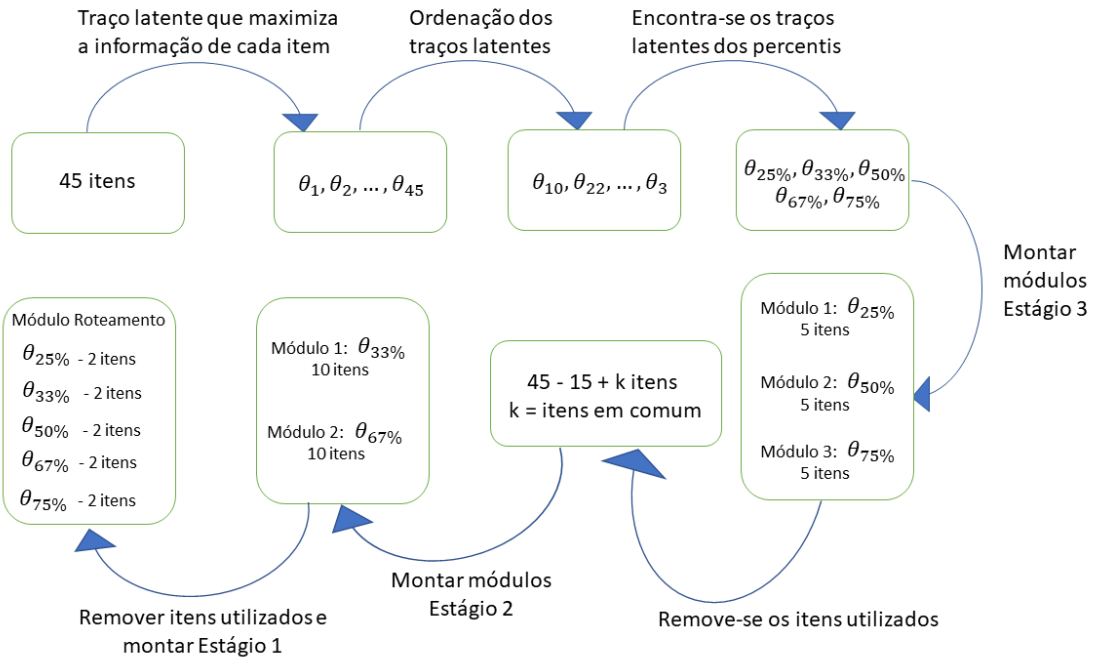

Figura 4: Fluxograma construção dos módulos.

4.3 Decisão do Roteamento do Teste Adaptativo Multiestágio

A próxima decisão do TAM é em relação ao roteamento dos examinados. Como decidir qual examinado é direcionado para cada módulo? Para isso, fez-se uma simulação dos examinados respondendo o TAM (seguindo algum método de roteamento) e uma estimação das habilidades dos mesmos examinados quando apresentados a todos os itens do ENEM.

Utilizou-se o método EAP nas estimativas das habilidades dos examinados. Esse método já está implementado na biblioteca mirt (Chalmers, 2012). A métrica utilizada na estimação das habilidades é *(0,1)*.

A segunda amostra gerada foi utilizada nesta fase. Na primeira etapa, estimam-se as habilidades dos examinados considerando as respostas dos 45 itens da prova. Os parâmetros dos itens utilizados são os mesmos calculados anteriormente. Na segunda etapa, faz-se a estimação das habilidades supondo que os examinados tenham respondido apenas os itens indicados para ele no TAM.

A realização de um TAM segue uma sequência de etapas. Primeiro, supôs-se que todos os examinados da amostra responderam o módulo de roteamento do Estágio 1. Em seguida, estimou-se a habilidade dos examinados, utilizando o método EAP e os parâmetros de itens já calculados. Após isso, fez-se um roteamento para determinar quais examinados seguem para o módulo 1 do Estágio 2, e quais seguem para o módulo 2 do Estágio 2.

Esse roteamento consistiu em testar exaustivamente várias possibilidades de corte. Os examinados com habilidade acima de um valor de corte são direcionados para o módulo 1, e o restante para o módulo 2 do Estágio 2. Os valores de corte testados estão no intervalo de habilidades encontradas no Estágio 1, inicia-se no menor valor de habilidade, e soma-se 0.1 até alcançar o maior valor de habilidade.

Uma vez no Estágio 2, estima-se novamente a habilidade dos examinados, considerando os itens referentes ao caminho percorrido. Faz-se o roteamento novamente, seguindo a mesma lógica explicada no Estágio 2, e cada examinado responde o módulo apresentado a si no estágio 3. E, por fim, estima-se a habilidade final do examinado, considerando todos os itens que tal examinado teria feito nessa simulação de TAM.

Para decidir quais pontos de corte das habilidades mais se adequaram ao TAM, escolheu-se os pontos que atingiram o menor Erro Quadrático Médio, este dado pela fórmula:

$$\frac{\sum_{i=1}^{10000}\left(EstimaçãoTAM\right| \quad \left|i - EstimaçãoCompleta_i\right)^2}{10000} \tag{2}$$

5. Resultados

Com a primeira amostra separada, foi possível fazer a estimação dos parâmetros dos itens, assim como montar os estágios e módulos. Já com a segunda amostra, foi possível comparar as duas abordagens possíveis: com um TAM e com o teste completo. E por fim, com a terceira amostra, consegue-se analisar o TAM quando aplicado a uma população de habilidade maior. Esta última análise mostrou-se interessante de ser realizada pelo fato observado de que o ENEM é composto de itens mais difíceis, adequados a um intervalo de habilidades mais altas.

5.1 Análise dos itens

A partir da amostra de 50 mil participantes, fez-se a calibração dos itens. A Figura 5 apresenta a CCI de cada um dos 45 itens.

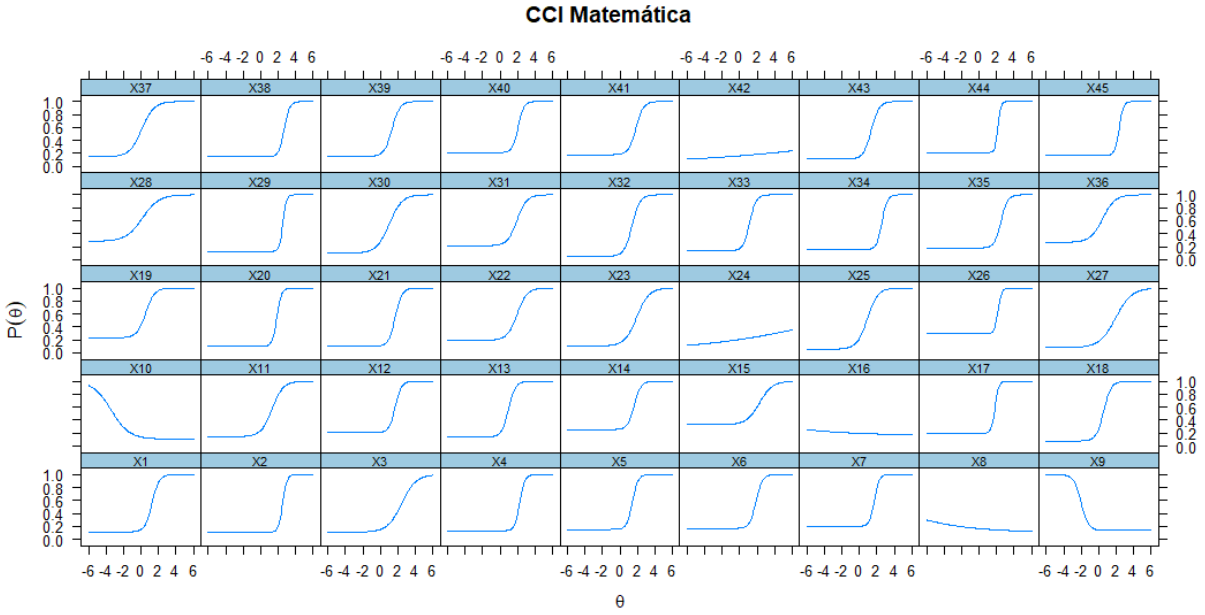

Figura 5: CCI prova de matemática.

Itens com discriminação próxima de zero ou com discriminação negativa são caracterizados como itens ruins. Na prova, encontraram-se 6 itens ruins, sendo estes os itens *8*, *9*, *10* e *16* (discriminação negativa), *24* e *42* (discriminação próxima de zero). A existência de itens

ruins na prova, incentivou a fazer uma exploração das dificuldades dos itens, utilizando a porcentagem de acerto destas.

Com o pacote CTT (Willse et al., 2008) do R, encontrou-se a porcentagem de acerto de cada item. A Tabela 3 mostra o número do item, sua classificação e a porcentagem de acertos, para os itens ruins e mais dois itens bons, para efeito de comparação.

Tabela 3: Análise Clássica de alguns itens.

Item	Classificação	Porcentagem de Acertos
8	Ruim	0.17
9	Ruim	0.19
10	Ruim	0.16
16	Ruim	0.20
24	Ruim	0.21
42	Ruim	0.16
37	Bom	0.55
39	Bom	0.28

Desses resultados, é possível perceber que os itens tidos como ruins possuem uma porcentagem de acerto muito baixa. Esses itens podem prejudicar a montagem do TAM, entretanto foram mantidos na análise. Na seção seguinte, mostrar-se-á a bem-sucedida estrutura do TAM, uma vez que tais itens ruins não passaram pelo critério se seleção automaticamente aqui proposto.

Em seguida, explorou-se a distribuição dos parâmetros de itens no contexto da TRI e a informação de tais itens. É importante ter uma ideia de como a dificuldade dos itens está distribuída para verificar se o banco de itens é suficientemente diversificado para montar um TAM adequado à população de examinados. Adicionalmente, pode-se observar as informações contidas nos itens ao longo da escala de habilidade e se os parâmetros estão de acordo com o esperado para um modelo da TRI.

Na Figura 6, tem-se a distribuição das estimativas dos parâmetros dos itens. O parâmetro *c*, do acerto ao acaso, está dentro do esperado, com valores entre 0 e 0.3, em que a maioria está concentrada em 0.15. Na Figura 7, apresenta-se a informação de todos os 45 itens.

O parâmetro *b* tem a maioria de valores concentrado entre 0 e 5, isto implica que tem uma variabilidade boa para examinados com traço latente nesse entorno. Alguns valores da dificuldade estão fora do intervalo *(0,5)*, e são justamente os itens classificados como ruins. O intervalo de valores da dificuldade não é muito abrangente considerando que a população foi imaginada com métrica *(0,1)*. Apesar disso, a montagem do TAM ainda é possível, com a ressalva que esse TAM não será tão abrangente quanto poderia ser.

O parâmetro *a* está muito concentrado entre 1 e 4, que são valores esperados em um modelo de TRI. Os valores próximos a 0 e menores que zero são inadequados, e já foram discutidos.

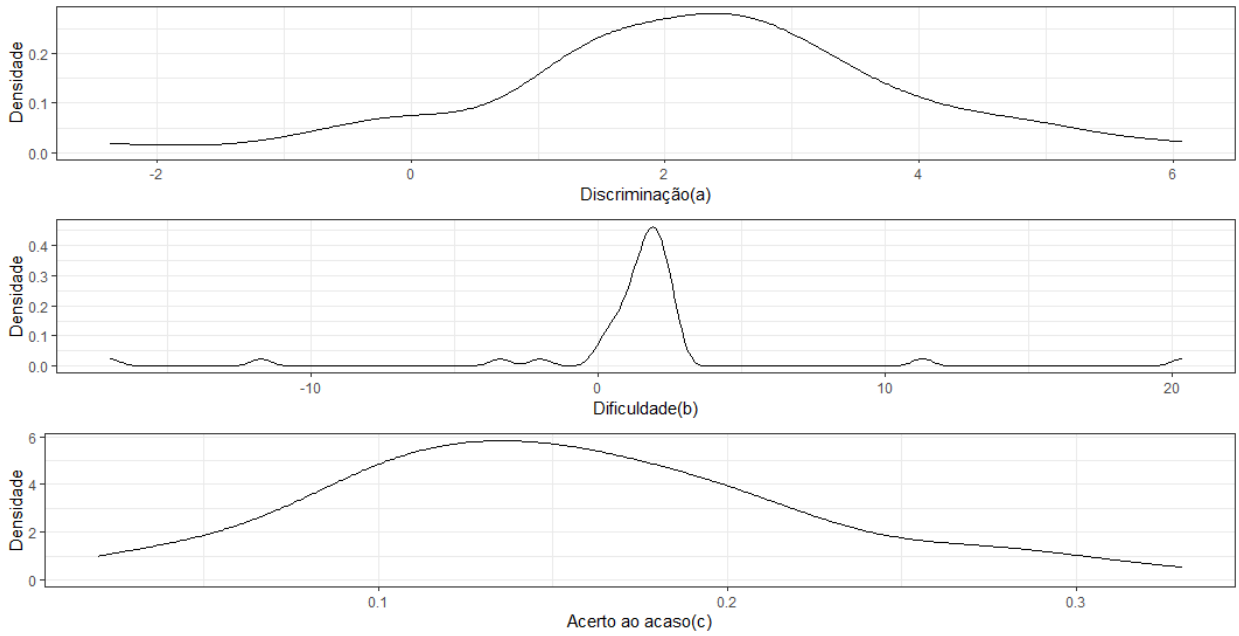

Figura 6: Distribuição das estimativas dos parâmetros dos itens.

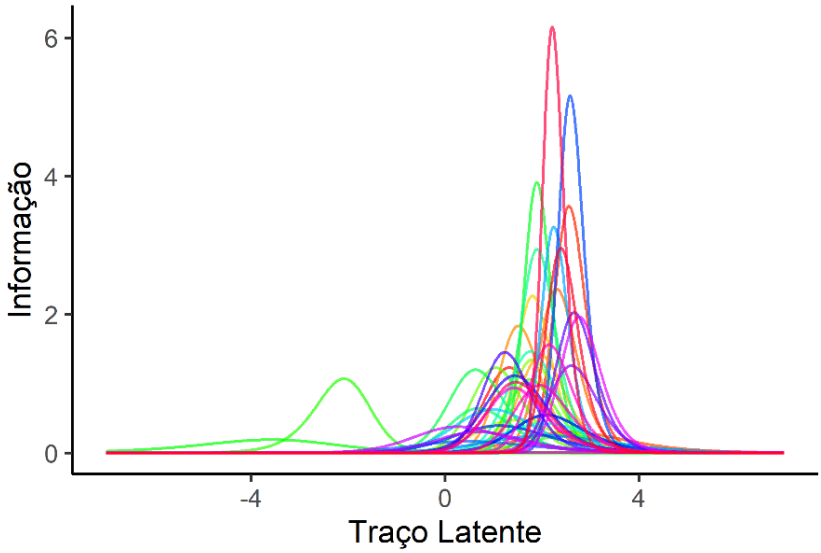

Figura 7: Informação dos itens.

5.2 Arquitetura Proposta

Nesta seção, propôs-se uma arquitetura de TAM para a prova do ENEM de 2019. A Figura 8 contém o TAM estruturado, com os 3 estágios e com as questões distribuídas nos módulos. Os itens classificados como ruins não foram selecionados para formarem os módulos, como já dito na seção anterior.

Os estágios 2 e 3 tiveram restrições para limitar os itens em comum. Um mesmo item pode ser muito informativo para dois traços latentes diferentes, possibilitando a entrada dele em dois módulos com habilidades diferentes. O estágio 2 permitiu um máximo de 5 itens em comum, no TAM montado, que foram os itens 43, 12, 14, 6, 41. O estágio 3 permitiu um máximo de 2 itens em comum. Nota-se que entre o módulo 1 e módulo 2 apenas o item 5 ficou em comum. Entre o

módulo 2 e módulo 3, apenas o item 17 ficou em comum. Entre o módulo 1 e módulo 3 não tiveram itens em comum.

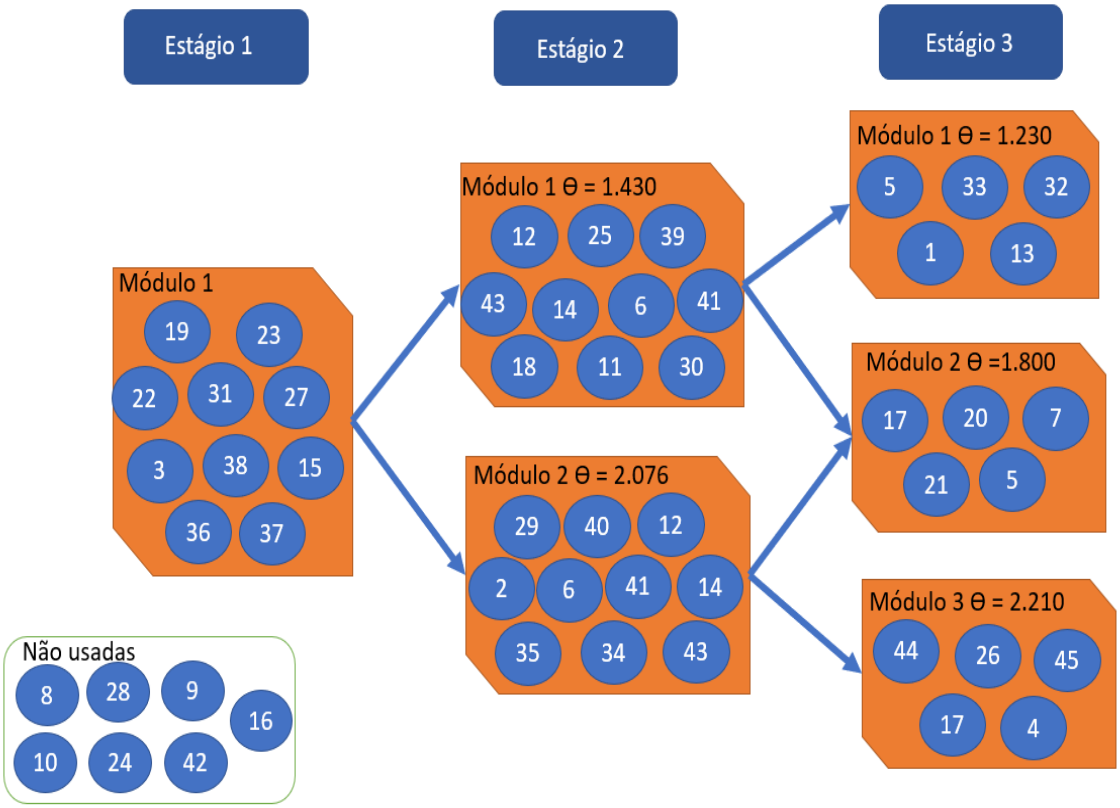

Figura 8: Formato proposto para o teste.

A partir do formato proposto, é possível encontrar características particulares desse teste. A Figura 9 mostra a informação de cada módulo, e tem-se uma ideia de quão precisas serão as habilidades estimadas em cada módulo. A Figura 10 mostra a dificuldade em cada módulo. O módulo de roteamento (do estágio 1) possui itens abrangendo uma maior variedade de dificuldades (o que é desejado), enquanto itens dos estágios seguintes têm dificuldades mais específicas para cada módulo.

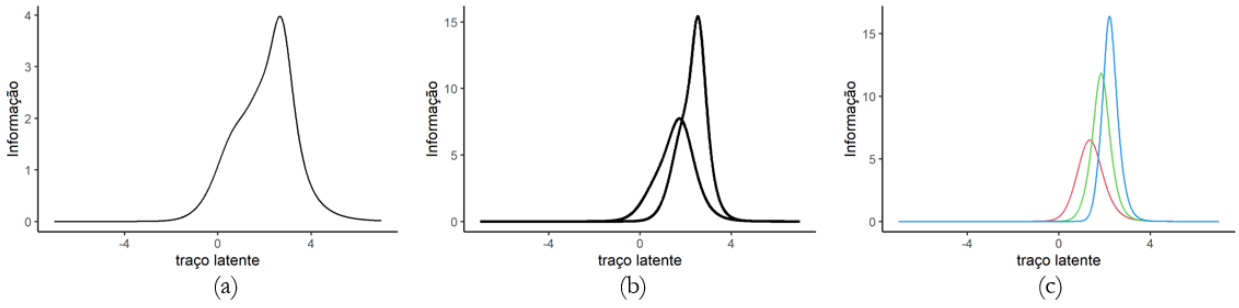

Figura 9 - Informação dos Estágios. (a) Informação do Estágio 1, apenas 1 módulo. (b) Informação do Estágio 2, com 2 módulos. (c) Informação do Estágio 3, com 3 módulos.

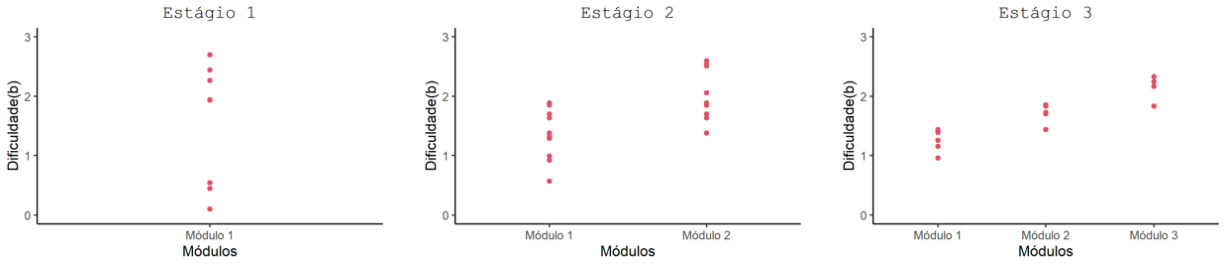

Figura 10: Dificuldade dos itens dos módulos divididas pelos Estágios.

Após essas análises, é importante ressaltar a falta de itens voltados para pessoas com habilidades menores. Apesar da divisão dos estágios, os módulos estão voltados para pessoas com habilidades maiores que 0. Ou seja, verifica-se que a prova é difícil em todos os seus itens, e isso dificulta a diferenciação dos indivíduos de baixa habilidade. Essa hipótese é corroborada ao analisar a porcentagem de acerto dos itens: o item com maior média de acertos possui *62%* de acerto. A Figura 11 mostra como essa porcentagem está distribuída. A maior parte dos itens tiveram média de acertos menor que *40%*.

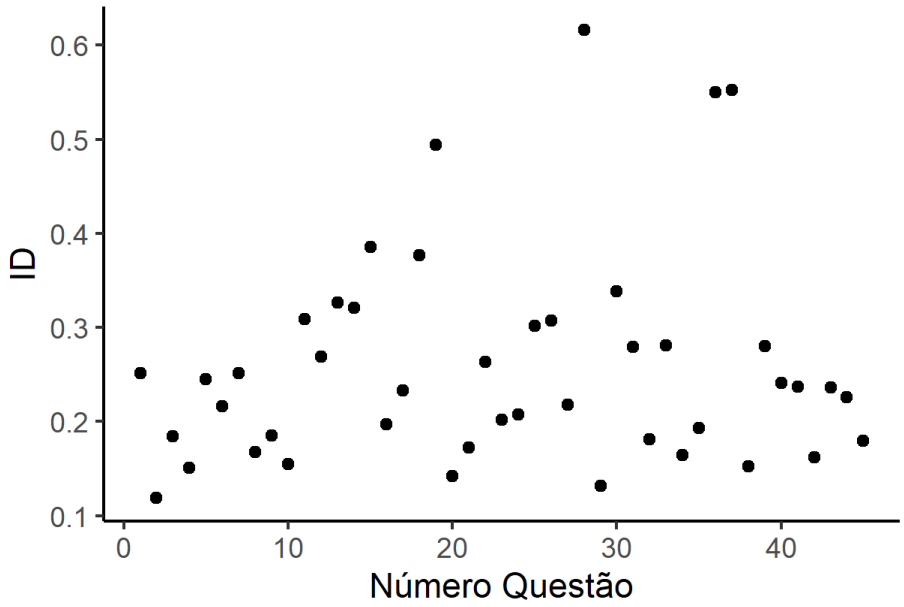

Figura 11: Média de acertos de cada item.

Esse fenômeno ocorre considerando o banco de itens utilizado. Seria importante ter um banco de itens maior para aperfeiçoar a construção do TAM, além de itens mais adequados para todos os participantes da prova.

5.3 Roteamento

Complementando a arquitetura proposta, o TAM necessita de uma forma para determinar quais examinados serão direcionados para cada módulo. Então, a partir da segunda amostra, testou-se exaustivamente várias possibilidades de corte para as habilidades dos indivíduos e encontrou-se que o menor valor do Erro Quadrático Médio foi de 0.035. Os pontos de corte foram: 1.7 no Estágio 2 e 1.39 e 2.18 no Estágio 3.

A Figura 12 esquematiza o roteamento feito, e indica quantos examinados seguiram para cada módulo. Dessa maneira, o TAM proposto consiste na arquitetura apresentada na Figura 8, juntamente com os valores de corte mostrados na Figura 12.

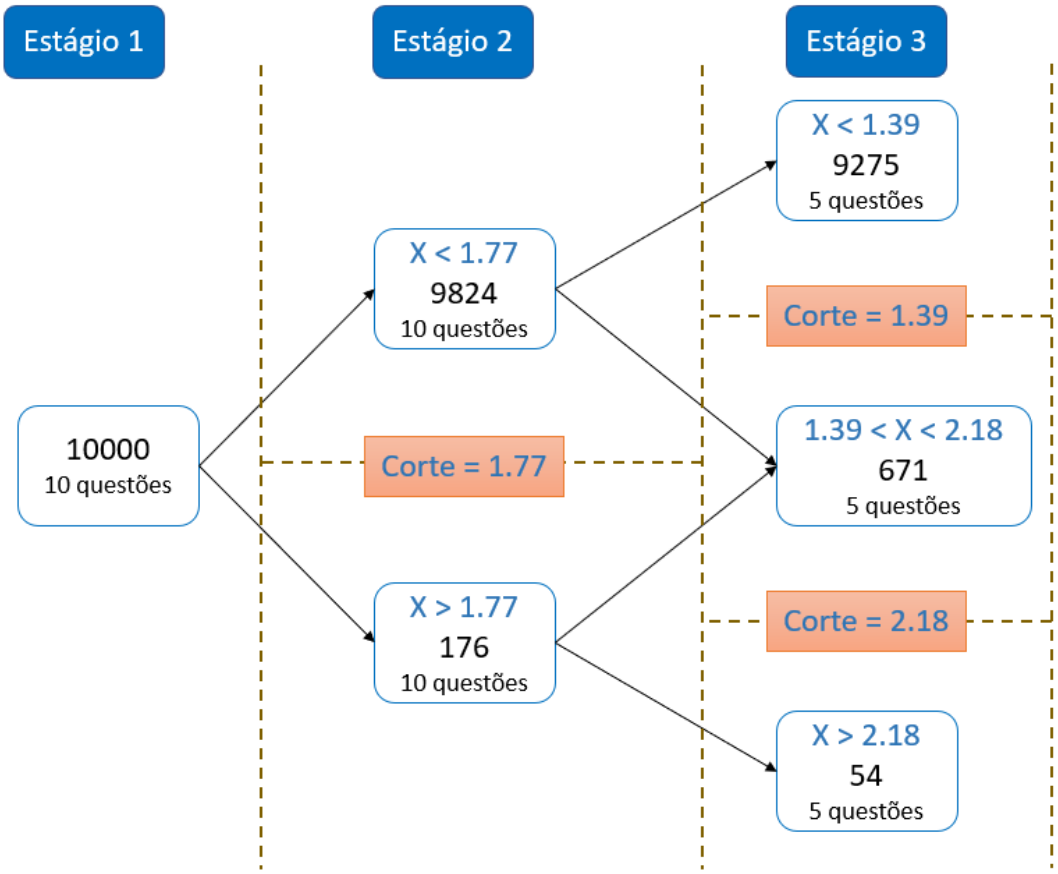

Figura 12: Roteamento dos examinados, segundo os níveis de habilidade estimados.

5.4 Verificação das habilidades: teste completo e Teste Adaptativo Multiestágio

Uma vez que se tem o TAM sugerido, e os valores de corte para realizar o roteamento dos examinados, é possível fazer uma comparação entre o teste completo e simulando o TAM. Para o caso do TAM, alguns problemas precisam ser ressaltados. A verdadeira prova feita como TAM não seria possível mensurar atualmente, pois tem-se apenas as respostas dos examinados em seu teste linear. Se o examinado tivesse feito o TAM verdadeiramente, ele teria um resultado diferente, pois precisaria fazer menos itens e poderia dar mais atenção (e tempo) para os itens. O formato da prova, por ser inédito para o examinado no ENEM, também poderia afetar seu desempenho. Essas são questões futuras, que devem ser levadas em consideração em estudos subsequentes.

Neste trabalho, para comparar os resultados do teste linear com o TAM, optou-se por fazer uma simulação de um TAM utilizando as respostas do teste linear, que é o disponível atualmente. Ou seja, considerou-se que a amostra de examinados responderam apenas os itens do TAM, com cada examinado respondendo os módulos condizentes a ele. A segunda amostra gerada foi utilizada para as simulações.

A Figura 13 mostra um gráfico de dispersão entre as estimativas das habilidades das duas simulações. A Figura 14 mostra a diferença entre as duas estimativas. A Figura 13, explicita a ideia de que os valores maiores de estimação ficaram muito próximos, enquanto os valores menores têm uma dispersão maior nas diferenças. Nas Figuras 13 e 14 percebeu-se que não existe um padrão claro na distribuição dos valores de habilidades maiores, as estimações por um formato não ficaram sistematicamente maiores que o outro. Apenas quando os valores são menores tem-se a prova TAM com estimações maiores do que no teste completo.

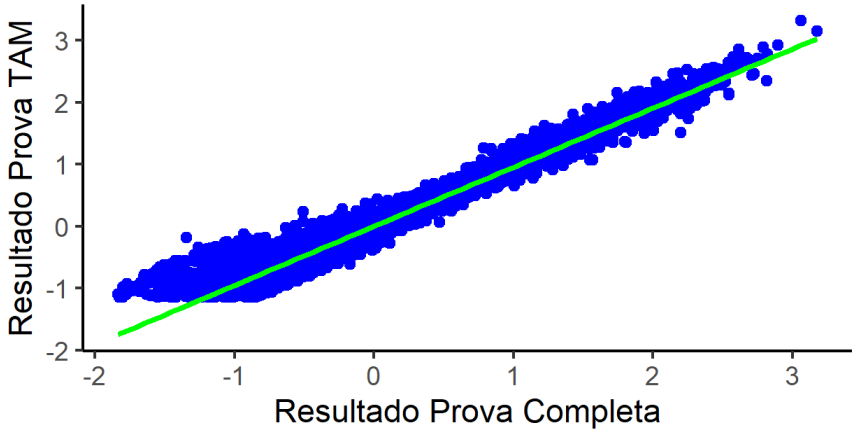

Figura 13: Gráfico de dispersão entre prova completa e TAM.

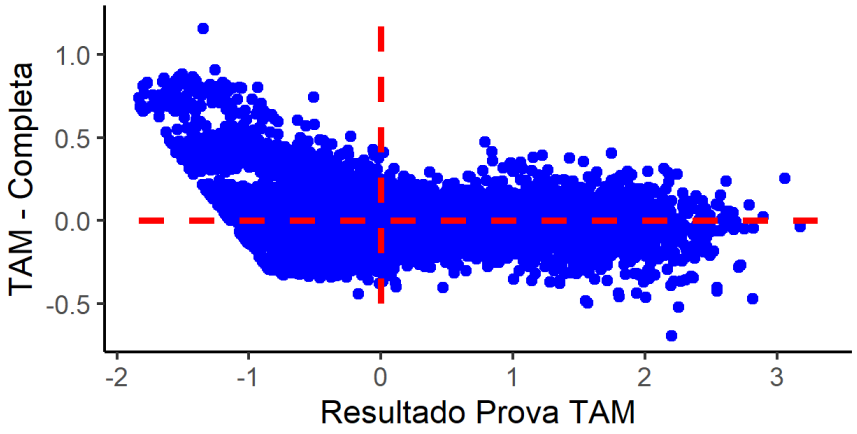

Figura 14: Diferença entre TAM e Teste Completo pelo TAM.

Nos procedimentos anteriores, foi possível perceber que a prova não é adequada para uma parte da população. A maioria dos itens tem uma dificuldade e uma informação adequadas a examinados com habilidades maiores. As Figuras 13 e 14 explicitam que o TAM teve uma estimação muito parecida para tais examinados com habilidades maiores, enquanto os examinados de habilidades menores ficaram bem distintos. Isso pode ter ocorrido pelo fato da estimação com a prova completa considerar os itens classificados como ruins, entre esses existem alguns que apresentam uma probabilidade maior de acerto para indivíduos com habilidade menor. Tais itens não apareceram no TAM, criando uma diferença na estimação dos examinados de menor habilidade. Outra possível justificativa seria a falta de itens adequados a indivíduos com menor habilidade, dificultando a estimação mais precisa deles.

A inadequação da prova para uma parte da população motivou a necessidade da terceira amostra. Onde o TAM seria simulado considerando apenas examinados com habilidades adequadas a prova. E para essa amostra, explorou-se se com uma redução de 45 para 25 itens seria possível encontrar estimativas de habilidade parecidas com o teste com os 45 itens (indicando que muitos itens respondidos pelos examinados não agregam tanta informação para a habilidade de tal aluno).

Considerando o caso do teste completo, a Figura 15 apresenta a distribuição de notas. Algumas estatísticas descritivas podem ser encontradas na Tabela 4.

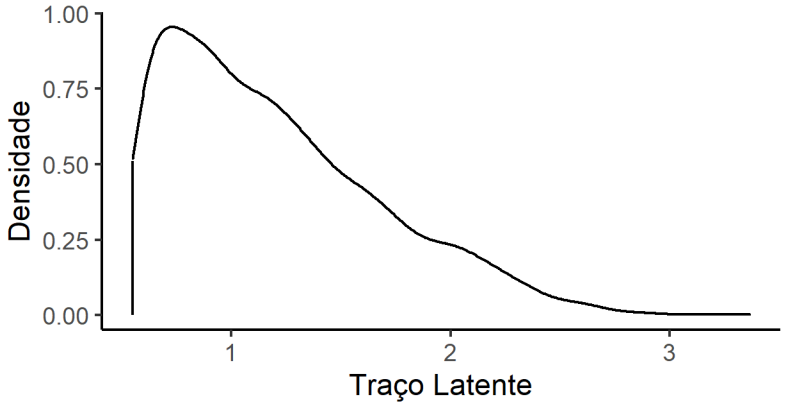

Figura 15: Distribuição dos traços latentes de 10 mil examinados.

Tabela 4: Estatísticas descritivas habilidades dos examinados para prova completa.

Mínimo	Mediana	Média	Máximo	Desvio Padrão	Média do Erro Padrão
0.55	1.1	1.196	3.37	0.49	0.27

Considerando apenas o TAM, a Figura 16 apresenta a distribuição de notas. Algumas estatísticas descritivas podem ser encontradas na Tabela 5.

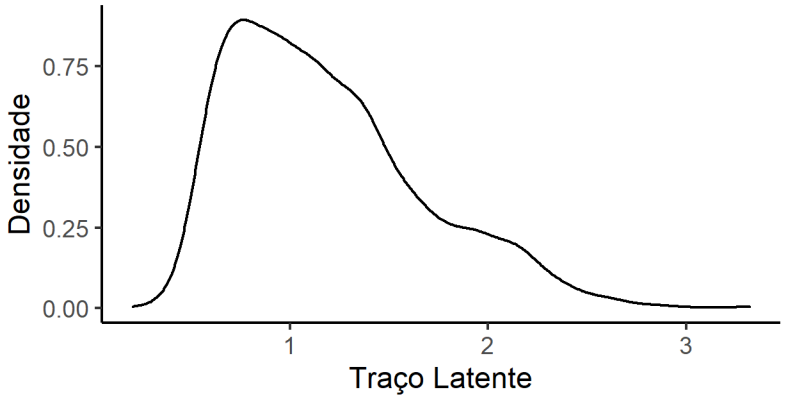

Figura 16: Distribuição das habilidades pelo TAM.

Tabela 5: Estatísticas descritivas habilidades dos examinados para o TAM.

Mínimo	Mediana	Média	Máximo	Desvio Padrão	Média do Erro Padrão
0.21	1.1	1.19	3.32	0.49	0.31

Pelas Tabelas 4 e 5, o valor mínimo no caso do teste completo é de 0.55, enquanto no TAM é de 0.21, essa diferença pode ser explicada pelos itens classificados como ruins. A média e a mediana de ambos ficaram em valores próximos.

A média do Erro Padrão nos dois casos ficou bem próxima. Isto é, a estimação de cada traço latente dos 5 mil examinados, possui um erro padrão, o qual foi estimado com o auxílio da biblioteca mirt, e tirou-se a média desses erros. Percebeu-se que essa média foi muito próxima em ambos os casos, indicando que a estimação está com uma precisão parecida nos dois tipos de prova.

Fez-se também uma análise quantitativa da diferença entre ambos os casos. Encontrou-se a correlação de Pearson (Rodgers et al., 1988) entre as estimativas dos dois casos, a correlação de Spearman e a Raiz do Erro Quadrático Médio (Tabela 6). Também foi feito um gráfico de dispersão entre as estimativas (Figura 17) e um gráfico mostrando a diferença entre as duas estimativas (Figura 18).

Tabela 6: Medidas de comparação entre o Teste Completo e o TAM.

Correlação Spearman	Correlação Pearson	Raiz do Erro Quadrático Médio
0.967	0.968	0.12

As estimativas de ambos os testes ficaram muito próximas, e a Raiz do Erro Quadrático Médio bem pequeno. Isso indica que as estimativas usando a TAM ficaram muito próximas das originais. Ou seja, com uma redução de aproximadamente *45%* dos itens, conseguiram-se estimativas muito próximas das habilidades. Os valores de correlação encontrados foram bem altos, indicando que a relação das estimativas é forte e linear, e enquanto uma cresce a outra também cresce. É possível perceber que a Raiz do Erro Quadrático Médio é menor que a Média do Erro Padrão do teste completo. Isto é um indicativo que as discrepâncias entre as habilidades estimadas com o TAM e o teste completo, são, em média, menores do que a imprecisão das mesmas. Assim sendo, é intuitivo concluir que os dois testes levam a estimativas equivalentes das habilidades individuais.

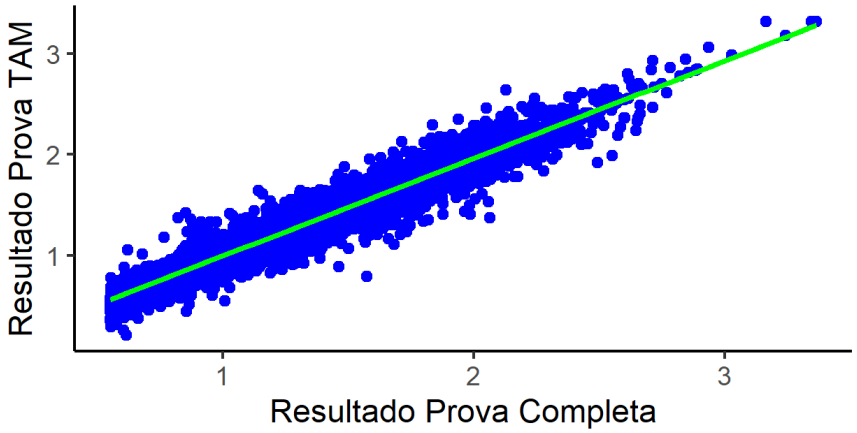

Figura 17: Gráfico de dispersão entre prova completa e TAM para os examinados selecionados.

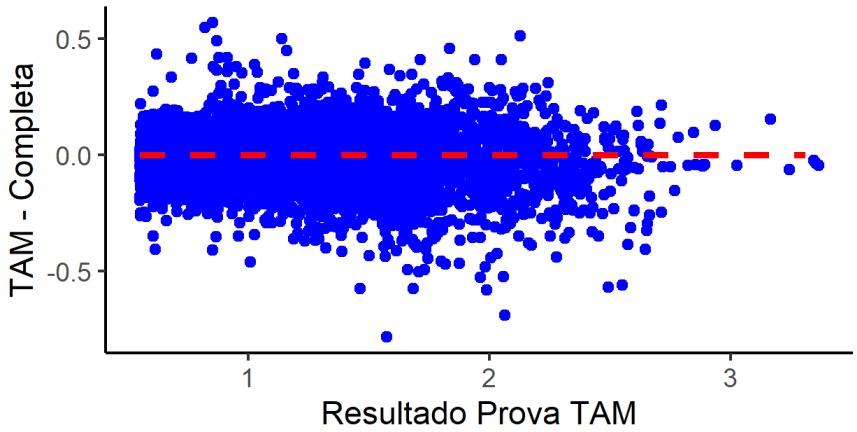

Figura 18: Diferença entre TAM e Teste Completo pelo TAM para os examinados selecionados.

6. Conclusão

Em 2020 ocorreu a primeira edição do ENEM Digital e em 2021 a segunda edição, apontando a continuidade da modalidade digital da prova. Uma vez que tal exame é aplicado no computador, tem-se a possibilidade (em um futuro próximo) de se implementar um ENEM adaptativo, proporcionando um teste mais curto para o participante do que o aplicado atualmente.

Nesse contexto, este artigo trouxe de forma inédita, no âmbito nacional, os aspectos teórico-metodológicos do TAM, instanciado no banco de itens de Matemática e suas Tecnologias da edição do ENEM de 2019. O processo de criação do TAM e uma simulação de aplicação foram detalhados em três etapas fundamentais: (i) a distribuição dos itens em módulos e estágios, (ii) as decisões de roteamento e (iii) as estimações das habilidades e respectivas comparações com as estimadas no teste completo.

Os resultados desse trabalho responderam positivamente à questão de pesquisa, evidenciando similaridade entre as estimações das habilidades do TAM e do teste completo, acompanhado de uma considerável diminuição no número de itens da avaliação. Os experimentos apontaram que a correlação de Spearman entre as habilidades estimadas do TAM (25 itens) e do teste completo (45 itens) foi alta (0.967) e a Raiz do Erro Quadrático Médio foi baixa (0.12), ou seja, o TAM estimou adequadamente as habilidades dos participantes, confirmando ser um teste eficaz quando comparado ao teste completo.

Além da eficácia confirmada, o TAM proposto foi capaz de reduzir o comprimento em 44,4% em relação ao teste completo. No que se refere à plausibilidade dessa porcentagem da redução de teste adaptativos em nível de item (aplicados ao ENEM), Spenassato et al. (2016) mostrou que seria possível reduzir em 26,6% o tamanho da prova de matemática do ENEM 2012 sem perda significativa de precisão das habilidades. Jatobá et. al (2018) e Jatobá (2019), utilizando o ALICAT e o ENEM 2012, conseguiram reduzir em 53,3% em relação ao teste completo sem perda significativa na estimação das habilidades. Vale lembrar que, em ambos os casos, os autores não usaram a metodologia TAM.

A calibração do banco de itens da edição de 2019 apontou para (i) uma baixa quantidade de itens “fáceis” e, consequentemente, (ii) uma prova caracterizada por itens mais “difíceis”, que agregam pouca informação para as habilidades estimadas de grande parte dos indivíduos (Figura 7).

Uma limitação do estudo foi em relação ao banco de itens utilizados, considerando apenas as questões de 2019. Em trabalhos futuros, sugere-se considerar uma quantidade maior de itens, contemplando as provas do ENEM de outros anos (desde que estejam na mesma escala) e garantir a presença de itens adequados para avaliar níveis de proficiência mais baixos na escala de habilidade, assim como um adequado balanceamento do conteúdo pedagógico entre os módulos. O banco de itens ser pequeno também afeta possíveis discussões acerca do conteúdo avaliado. Para assegurar que todos os examinados passem pelos conteúdos esperados, seria necessário aumentar o banco de itens com itens de mesmo conteúdo e distintas habilidades. Conseguindo-se um banco de itens nessas condições, o desenvolvimento de um sistema de TAM funcional passível de aplicação a usuários reais também pode ser uma interessante sequência deste tema de estudo.

Outro ponto limitante foi o fato de utilizar-se apenas a área de matemática. Novas pesquisas podem aprofundar a discussão incluindo outras disciplinas. Também não foram consideradas provas adaptadas para portadores de necessidades especiais, que é essencial para uma maior inclusão de todos os candidatos no sistema de avaliação. A superação dessas limitações é importante e esbarra em dificuldades técnicas relevantes que podem ser objetos de trabalhos futuros.

O TAM é promissor para mensurar habilidades latentes de candidatos ao ENEM, uma prova que tem um número muito grande de examinandos. Estimar de forma precisa as habilidades dos respondentes traz justiça a um sistema de avaliação extremamente importante que permite que os estudantes ingressem no ensino superior.

Uma vez que o ENEM se torne completamente digital, o TAM é uma alternativa a ser considerada para a realização da prova, visto que se pode criar diferentes painéis de acordo com o cenário educacional, atendendo às mais variadas necessidades e contextos de instituições de ensino e avaliação.

Referências

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). Teoria da resposta ao item: conceitos e aplicações. *ABE*, São Paulo. [GS Search]

Angoff, W.H. (1957). The "Equating" of non-parallel tests. *Journal of Experimental Education*, 25(3), pp. 241–247. doi: 10.1080/00220973.1957.11010574. [GS Search]

Avellar, Simone. Enem pode ser aplicado via computador a partir de 2016. O Globo, Rio de Janeiro, 16/12/2012. Seção Educação. Disponível em [Link].

Barbosa, M. E. F., & Fernandes, C. (2001). A escola brasileira faz diferença? Uma investigação dos efeitos da escola na proficiência em Matemática dos alunos da 4ª série. *Em C. Franco (org), Promoção, ciclos e avaliação educacional*. ArtMed, Curitiba. [GS Search]

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education*, 39. doi: 10.3102/0091732X14554179. [GS Search]

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In: Lord, F.M. and Novick, M.R., Eds., *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, pp. 397-479. [GS Search]

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

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software,* 71(5), pp. 1-38. doi: 10.18637/jss.v071.i05. [GS Search]

de Macedo, E. P. N. (2021). As diferentes fases do Enem: olhar o passado para pensar o futuro. *Em Aberto*, 34(112). doi: 10.24109/2176-6673.emaberto.34i112.4999. [GS Search]

Dias, A. C. F. (2019). Uma solução Bayesiana para se considerar a incerteza associada à calibração de itens na teoria de resposta ao item. Dissertação de Mestrado. Programa de Pós-Graduação em Estatística. Universidade Federal de Minas Gerais. Belo Horizonte. Defesa: 28/06/2019. [GS Search]

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1), pp. 1–35. [GS Search]

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education: An International Journal*, 13(3), 441–458.

Gierl, M.J., & Lai, H. (2013). Instructional Topics in Educational Measurement (ITEMS) Module: Using Automated Processes to Generate Test Items. *Educational Measurement: Issues and Practice*, 32(3), pp. 36–50. doi: 10.1111/emip.12018. [GS Search]

Grégoire, J., & Laveault, D. (2002). Introdução às Teorias dos Testes em Ciências Humanas. Porto, Portugal: Porto.

Hendrickson, A. (2007). An NCME Instructional Module on Multistage Testing. *Educ. Meas. Issues Pract*. 26 (2), pp. 44–52. doi: 10.1111/j.1745-3992.2007.00093.x. [GS Search]

INEP. (2019). Edital ENEM 2019. Diário Oficial da União – Seção 3. Disponível em [Link].

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (n.d.). Exame Nacional do Ensino Médio (ENEM). Disponível em [Link].

Jatobá, V. M. G., Delgado, K. V., Farias, J. S., & Freire, V. (2018). Comparação de Regras de Seleção de Itens em Testes Adaptativos Computadorizados: um estudo de caso no ENEM. *Anais do XXIX Simpósio Brasileiro de Informática na Educação*. Pp. 1453-1462. doi: 10.5753/cbie.sbie.2018.1453. [GS Search]

Jatobá, V. M. G. (2019). Uma abordagem personalizada no processo de seleção de itens em Testes Adaptativos Computadorizados. Dissertação de Mestrado. Programa de Pós-Graduação em Sistemas de Informação. Universidade de São Paulo. São Paulo. Defesa: 08/10/2018. [GS Search]

Kirsch, I., & Lennon, M.L. (2017). PIAAC: a new design for a new era. *Large-Scale Assessments in Education*, 5(1), 11. doi: 10.1186/s40536-017-0046-6. [GS Search]

Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13(4), pp. 517–549. doi: 10.1177/001316445301300401. [GS Search]

Palermo, G. A., Silva, D. B. N., & Novellino, M. S. F. (2014). Fatores associados ao desempenho escolar: uma análise da proficiência em matemática dos alunos do 5º ano do ensino fundamental da rede municipal do Rio de Janeiro. *Revista Brasileira de Estudos de População*, 31(2). Pp. 367-394. doi: 10.1590/S0102-30982014000200007. [GS Search]

Piton-Gonçalves, J. (2020). Testes Adaptativos para o Enade: uma aplicação metodológica. *Revista Meta: Avaliação*. 12(36). doi: 10.22347/2175-2753v12i36.2735. [GS Search]

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. *Danish Institute for Educational Research*. [GS Search]

Reckase, M.D. (1974). An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods & Instrumentation*, 6(2), pp. 208–212. doi: 10.3758/BF03200330. [GS Search]

Ricarte, T. A. M., Curi, M., & von Davier, A. (2018). Modeling Accidental Mistakes in Multistage Testing: A Simulation Study. *Springer Proceedings in Mathematics & Statistics*, 233, pp. 55-65. doi: 10.1007/978-3-319-77249-3_5. [GS Search]

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1), pp. 59-66. doi: 10.1080/00031305.1988.10475524. [GS Search]

Silva, V. R., Curi, M. (2019). Academic English proficiency assessment using a computerized adaptive test. *TEMA: Tendências em Matemática Aplicada e Computacional*, 20(2). doi: 10.5540/tema.2019.020.02.0381. [GS Search]

Spenassato, D., Trierweiller, A. C., Andrade, D. F., & Bornia, A. C. (2016). Testes Adaptativos Computadorizados Aplicados em Avaliações Educacionais. *Revista Brasileira de Informática na Educação*, 24(2), pp. 1. ISSN 2317-6121. doi: 10.5753/rbie.2016.24.02.1. [GS Search]

Van Der Linden, W. J., & Glas, C. A. W. (2010). Elements of Adaptative Testing. Springer. Capítulo 18. Disponível em [Link].

Von Davier, A. (2017). Computational Psychometrics in Support of Collaborative Educational. *Assessments Journal of Educational Measurement*, 54(1), pp. 3–11. [GS Search]

Wainer, H., Kaplan, B., & Lewis, C. (1992). A Comparison of the Performance of Simulated Hierarchical and Linear Testlets. *Journal of Educational Measurement*, 29(3), pp. 243–251. doi: 10.1111/j.1745-3984.1992.tb00376.x. [GS Search]

Weiss, D.J. (1985). Adaptive Testing by Computer. *Journal of Consulting and Clinical Psychology*, 53(6), pp. 774–789. doi: 10.1037/0022-006X.53.6.774.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. doi: 10.1080/15366367.2019.1565254.

Willse, J. T., & Shu, Z. (2008). CTT: Classical Test Theory Functions. R package version 1.0.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). Introduction of multistage adaptive testing design in PISA 2018. *OECD Education Working Paper Número 209*. doi: 10.1787/19939019.

Yan, D., Lewis, C., & von Davier, A. (2014). Computerized multistage testing: Theory and applications. Capítulo 1. CRC. [GS Search]

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. (2012). Multistage adaptive testing for a large-scale classification test: The designs, automated heuristic assembly, and comparison with other testing modes. Report number: ACT Research Reports 2012-6Affiliation: ACT, Inc. [GS Search]