UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Departamento de Sistemas de Computação

Teste Adaptativo Multiestágio para o ENEM 2019

Gabriel Couto Tabak

São Carlos - SP

Teste Adaptativo Multiestágio para o ENEM 2019

Gabriel Couto Tabak

Orientadora: Mariana Cúri

Monografia referente ao projeto de conclusão de curso dentro do escopo da disciplina SSC0670 - Projeto de Formatura I do Departamento de Sistemas de Computação do Instituto de Ciências Matemáticas e de Computação – ICMC-USP para obtenção do título de Engenheiro de Computação.

Área de Concentração: Estatística

USP – São Carlos
12 de junho de 2020

“É no problema da educação que assenta o grande segredo do aperfeiçoamento da humanidade.”

(Immanuel Kant)

Agradecimentos

Agradeço aos meus pais, Benjamin e Claudia, que me apoiaram muito durante a minha graduação, e durante a realização deste trabalho.

Agradeço à Mariana Cúri, minha orientadora, desde que comecei os estudos de TRI com ela, me guiou e me ajudou muito. Agradeço a ela por ter acreditado em mim, e apoiado as minhas ambições.

Agradeço a minha tia, Vilma, minha segunda mãe, que apesar de ter se graduado em uma área completamente diferente desta monografia, tirou um tempo para poder ler e me ajudar em muitas coisas.

Agradeço aos meus primos, Pedro e Leo, que fizeram uma enorme diferença para mim durante a graduação.

Agradeço aos meus colegas de São Carlos, especialmente o Daniel, que fez sua monografia concomitantemente à minha, e acabamos nos ajudando e nos incentivando. Também fica um agradecimento ao Benine, Camo, Felipe, Matheus, Paroni, Paulo, Eiji e Afonso, que foram colegas importantíssimos durante essa jornada na engenharia.

Agradeço à Universidade de São Paulo (USP), que me proporcionou muito conhecimento e muitas oportunidades. E também agradeço à USP, pela bolsa concedida (Programa Santander - USP de Política Públicas) na qual meus estudos sobre TRI se iniciaram.

Resumo

Avaliações educacionais têm um impacto enorme na formação de um estudante. No Brasil, o Exame Nacional do Ensino Médio (Enem) é uma avaliação em larga escala que promove a entrada no ensino superior. Em 2020, implementou-se esse exame de forma digital, criando a oportunidade de alterar o formato do exame. Exames computadorizados permitem a elaboração de provas adaptativas. Neste trabalho, explorou-se a possibilidade de construir um teste adaptativo multiestágio com a prova do Enem de 2019 na área de Matemática. Discutiram-se algumas vantagens do teste multiestágio em comparação com o teste adaptativo por item e em comparação com o teste linear. Dessa prova, analisaram-se os itens com a Teoria de Resposta ao Item e com a Teoria Clássica dos Testes. Construiu-se um teste com 3 estágios. Com esse teste construído, as habilidades de um examinado considerando o teste completo foram comparadas com o mesmo examinado fazendo o teste adaptativo. Constatou-se que o teste adaptativo reduziu em 45% o tamanho do teste e a estimação das habilidades foi mantida próxima da estimação com o teste completo. Notou-se também que esse exame é voltado para pessoas com habilidades maiores, tornando a estimação das habilidades prejudicada para indivíduos com habilidades menores.

Sumário

Lista de Abreviaturas/Siglas

TCT	Teoria Clássica dos Testes
TRI	Teoria de Resposta ao Item
ENEM	Exame Nacional do Ensino Médio
ENADE	Exame Nacional de Desempenho dos Estudantes
ID	Índice de Dificuldade
IDS	Índice de Discriminação
MEC	Ministério da Educação
Sisu	Sistema de Seleção Unificada
TC	Testes Computadorizados
TAC-I	Testes Adaptativos Computadorizados em nível de Item
TAM	Testes Adaptativos Multiestágio
CCI	Curva Característica do Item
EM	Expectativa-Maximização
EAP	Estimação pela Média da Posteriori
GRE	*Graduate Record Examination*
ETS	*Educational Testing Service*
PISA	*Programme for international Student Assessment*
ATA	*Automated Test Assembly*

Lista de Tabelas

Lista de Figuras

CAPÍTULO 1: INTRODUÇÃO

1.1. Contextualização e Motivação

Em 2019, o Ministério da Educação (MEC) anunciou a primeira aplicação digital do Exame Nacional do Ensino Médio (ENEM) para o ano de 2020. Em seu planejamento, o MEC propôs a consolidação da prova digital do ENEM até 2026[1] (LOPES, 2019).

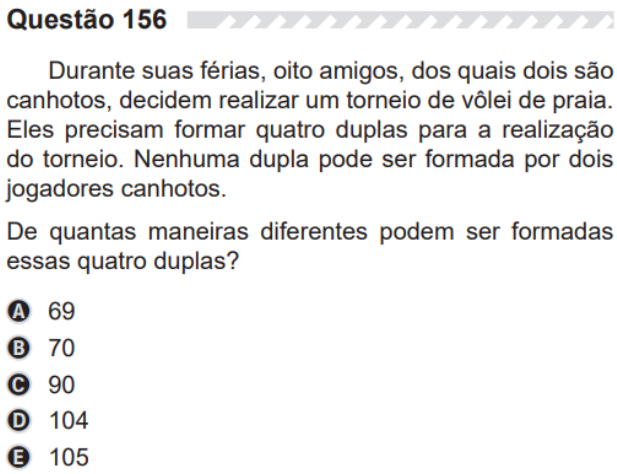
Existem algumas vantagens que podem ser observadas com a digitalização do ENEM (LOPES, 2019). Essas vantagens devem ocorrer em relação ao barateamento dos custos e aprimoramento das questões: i) não seria necessário imprimir as provas para todos os aplicantes, e; ii) a prova seria mais versátil - permitiria a utilização de questões com vídeos, infográficos, lógica de jogos.

O ENEM foi estabelecido em 1998, cujo objetivo era avaliar a performance escolar dos estudantes ao término da educação básica (INEP 2021). A partir de 2009, passou a ser utilizado como exame de acesso ao ensino superior do Brasil, em que alunos que já concluíram o ensino médio, ou estão concluindo, podem pleitear vagas em instituições públicas e privadas por meio do Sistema de Seleção Unificada (Sisu).

Em relação à estruturação do exame, este consiste em 180 itens de múltipla escolha (com cinco alternativas - Exemplo de questão Figura 1) distribuídos em quatro áreas do conhecimento (Matemática, Ciências Humanas, Ciências da Natureza e Linguagens e Códigos), além de uma redação (INEP 2021).

[1] Uma questão relevante é a da necessidade de computadores em todas as unidades nas quais o Enem Digital seria aplicado e das ações que devem ser desenvolvidas para que isso seja possível e permitir que o Enem Digital se torne uma realidade em todas as regiões do país. Não se discutiu esta questão neste trabalho.

Figura 1 -Questão 156 da prova azul de Matemática e suas tecnologias da edição de 2019. A resposta correta é a opção C.



Questão 156

Durante suas férias, oito amigos, dos quais dois são canhotos, decidem realizar um torneio de vôlei de praia. Eles precisam formar quatro duplas para a realização do torneio. Nenhuma dupla pode ser formada por dois jogadores canhotos.

De quantas maneiras diferentes podem ser formadas essas quatro duplas?

A) 69
B) 70
C) 90
D) 104
E) 105

Fonte: Microdados ENEM (INEP, 2021)

A aplicação da prova ocorre em dois domingos. No primeiro domingo do Exame, são aplicadas as provas de Linguagens, Códigos e suas Tecnologias (45 questões), Redação e Ciências Humanas e suas Tecnologias (45 questões). Com duração de 5 horas e 30 minutos. No segundo domingo do Exame, são aplicadas as provas de Ciências da Natureza e suas Tecnologias (45 questões) e Matemática e suas Tecnologias (45 questões). Com duração de 5 horas (ENEM, 2012).

Em 2019, cada uma das áreas do conhecimento era dividida em 6 tipos de provas separadas entre 7 cores: Azul, Amarela, Branca, Cinza, Rosa, Verde e Laranja (INEP, 2021). Dentre os tipos de prova, tem-se o formato adaptado (MEC, 2021) para auxiliar participantes com algum tipo de deficiência[2].

Os tipos de prova de cores Azul, Amarela, Branca, Cinza e Rosa possuem o formato convencional. O tipo de prova Laranja, é adaptada com um ledor, em que existem adaptações nos textos e descrições das imagens presentes na prova convencional. O tipo de

[2] Importante ressaltar que o ENEM Digital também irá exigir formatos adaptados para pessoas com variados tipos de deficiência, de modo a permitir que tenham seus direitos garantidos. Não se trata desta questão neste trabalho.

prova Verde também é adaptado, em que se utiliza da plataforma de Videoprova em Libras. Caso a adaptação de algum item da prova se mostrar inadequada, tal item é substituído por outro equivalente em termos pedagógicos e psicométricos.

O ENEM também permite a reaplicação (INEP, 2019) da prova para participantes que justifiquem sua ausência seguindo os critérios exigidos. A reaplicação conta com uma prova nova com itens inéditos.

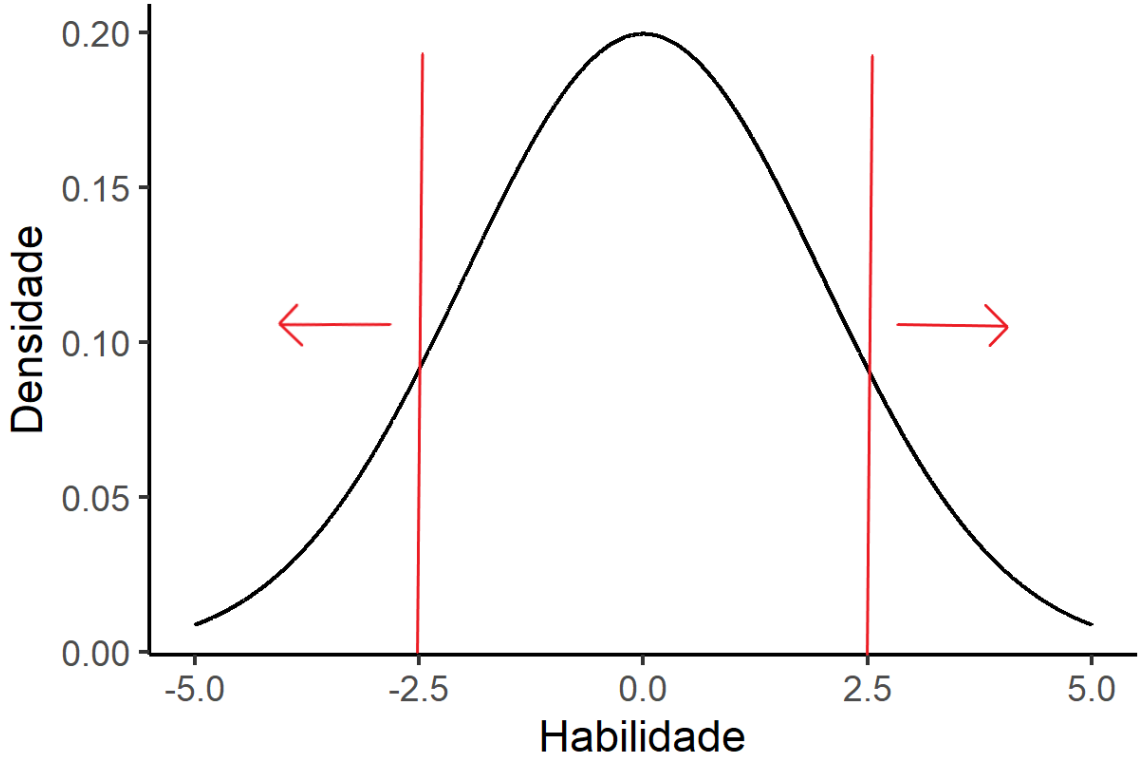
A aplicação do ENEM Digital em 2020 contou com 93.079 inscritos, porém, apenas *31.5%* de presença (ARÊA, 2021). Como os elaboradores do ENEM têm como meta a elaboração de uma prova totalmente computadorizada, abrem-se oportunidades de exploração de outros métodos para aplicação da prova.

Tradicionalmente, a aplicação de provas seguiu a metodologia de testes lineares (YAN et al., 2014). Testes lineares são aqueles no qual os participantes fazem todas as questões. E as questões são apresentadas a todos os participantes. Geralmente esses testes são feitos em papel e lápis, sendo essa a metodologia adotada pelo ENEM.

É razoável assumir que as notas (habilidades) dos examinados em um teste seguem uma determinada distribuição de probabilidade. Na Figura 2, tem-se um exemplo em que as notas de uma população seguem uma distribuição normal.

Percebe-se que existe um intervalo de notas possíveis para a população. E quando se estima a nota de tal população, é importante conseguir distinguir bem as notas dentro de toda a distribuição. Ou seja, é importante existirem questões que sejam adequadas a todos os níveis de habilidades.

Figura 2 – Exemplo de distribuição das habilidades.



Fonte: Elaborado pelo autor.

Observe que questões com grau de dificuldade alto, atende a parte da direita da população demarcada na Figura 2. E a probabilidade dos examinados na parte à esquerda de acertar tais questões é muito baixa. Analogamente, questões com grau de dificuldade muito baixo ajudam a estimar a nota dos examinados a esquerda da distribuição. Enquanto essas questões têm probabilidade muito alta de acerto para a parte à direita da distribuição. Nas avaliações educacionais, espera-se que seja possível estimar as habilidades ao longo de toda a distribuição.

Dessa forma, um teste linear precisa de questões que atendam a toda a distribuição dos alunos. Precisa-se de um número grande de itens para obter uma precisão uniformemente boa das habilidades. Como nos testes lineares todas as questões são apresentadas aos alunos, é necessário provas mais extensas (YAN et al., 2014).

Uma prova extensa dura mais tempo e precisa-se de mais tempo para que os examinados possam realizá-la. Tal prova também possui itens que não auxiliam na estimação da habilidade de certo examinado (probabilidade de acertar muito baixa ou muito alta), apesar de auxiliar em outro. Ou seja, a prova deve possuir muitos itens para atingir toda a distribuição de habilidades, porém cada examinado tem uma prova inflada de itens não condizentes com sua habilidade. Uma alternativa proposta para tal, são os testes adaptativos (YAN et al., 2014).

Com o advento dos testes baseados em computadores, ou, simplesmente, testes computadorizados, abre-se a oportunidade de outros métodos de aplicação da prova. O Teste Computadorizado (TC) pode ser linear, adaptativo em nível de item ou adaptativo multiestágio.

Os testes, quando adaptativos, estimam a habilidade durante o exame, antes de apresentarem os próximos itens. Portanto, necessitam do auxílio do computador para facilitar essa estimação. O TC linear, segue a mesma lógica do teste linear, porém, agora, aplicado no computador (sendo esse o caso do ENEM Digital de 2020).

Para o caso adaptativo, têm-se os Testes Adaptativos Computadorizados em Nível de Item (TAC-I), em que um algoritmo determina quais itens são enviados para o examinado. Esse algoritmo é feito para adaptar os itens conforme a habilidade do respondente. A cada item respondido, a habilidade é estimada e o próximo item é escolhido (YAN et al., 2014).

Ou seja, os examinados respondem os itens condizentes com sua habilidade. Não se tem mais uma prova inflada, cada examinado responde menos questões. O banco de itens ainda precisa de variadas questões que se adequem a toda a distribuição de habilidade (YAN et al., 2014).

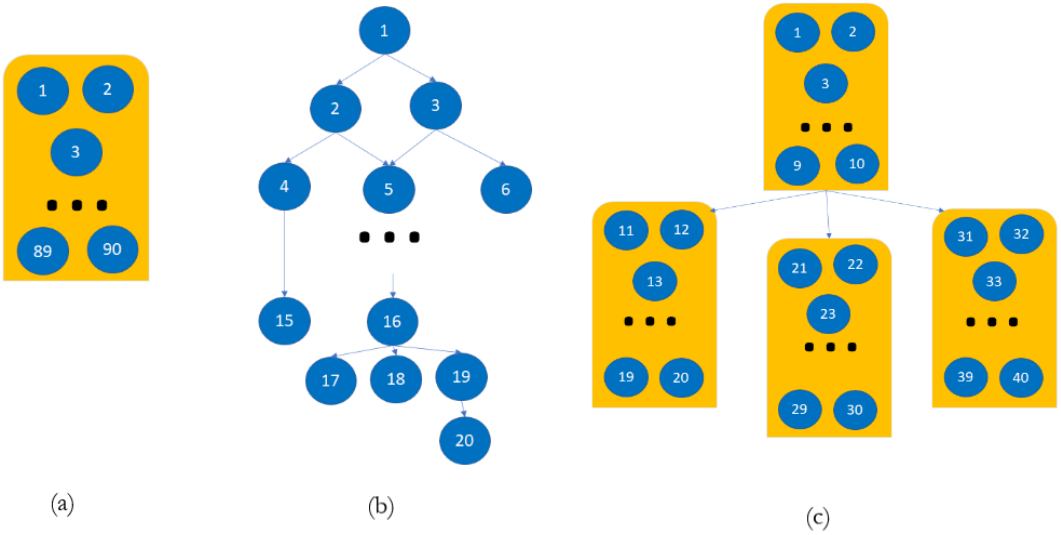
O TAC-I possui algumas desvantagens (YAN et al., 2014). Quando implementado, o respondente não terá o direito de revisar suas questões (a próxima questão a aparecer depende da resposta da questão atual). O TAC-I não é facilmente aplicado para alguns modelos de questões (por exemplo redações), e exigem algoritmos complexos para

satisfazer um controle de conteúdo apresentado (o conteúdo das questões apresentadas deve seguir diretrizes pedagógicas para todos os examinados).

Dessa maneira, uma alternativa aos testes lineares e ao TAC-I, são os testes adaptativos multiestágio. Sendo este, o método explorado neste trabalho para o ENEM. O Teste Adaptativo Multiestágio (TAM) consiste em uma prova formada por estágios, em que em cada estágio o respondente recebe um conjunto de itens pré-determinados (esse conjunto é chamado de módulo). O examinado passará por todos os estágios, em que cada um tem um conjunto de questões adequadas a sua habilidade. Ao fim de um estágio, a habilidade do examinado é estimada e decide-se qual módulo do estágio seguinte é mais adequado para tal habilidade (YAN et al., 2014).

Na Figura 3, tem-se um quadro esquemático mostrando as diferenças entre as possíveis metodologias. Os círculos representam as questões, e as setas representam o caminho que determinado examinado irá seguir. O TAC-I, a cada item tem-se um caminho possível, no TAM após o módulo, tem-se um novo caminho, no teste linear tem-se que responder toda as questões.

Figura 3 - Tipos de TC: (a) Teste Linear. (b) TAC-I. (c) TAM de dois estágios.



Fonte: Elaborado pelo autor.

O TAM aproveita as vantagens dos testes lineares e dos testes adaptativos, enquanto tenta minimizar suas desvantagens. Os módulos de um estágio são montados conforme a habilidade esperada do indivíduo. Pode-se ter um estágio com um módulo fácil, outro médio e outro difícil, sendo possível manter uma boa precisão na estimação de habilidades distintas.

Em comparação com os testes lineares, tem-se uma previsão melhor da habilidade dos examinados com menos itens (as questões respondidas por tal, são voltadas para sua habilidade). Entretanto, como são conjuntos de questões, não se tem um direcionamento tão bom quanto no TAC-I (onde cada questão escolhida é direcionada para a habilidade estimada) (YAN et al., 2014).

A prova de um indivíduo não precisa ser tão extensa como nos testes lineares. Assim como no TAC-I, questões que tenham uma probabilidade muito alta de o aluno errar, ou acertar, não são apresentadas para tal aluno.

O TAM permite que seus módulos sejam montados antes da administração do teste (YAN et al., 2014). Os desenvolvedores do teste conseguem montar módulos balanceados em conteúdo, sem a necessidade de algoritmos muito complexos (como ocorreria no TAC-I). O TAM admite que os alunos revisem os itens dentro de um módulo, antes de prosseguirem para o próximo estágio.

Neste trabalho, propõe-se um TAM para a prova de Matemática do ENEM, em que o banco de itens é a prova de 2019. Acredita-se que o TAM pode trazer diversos benefícios para o ENEM. O ENEM segue a metodologia dos testes lineares, portanto, para se ter uma boa precisão de todo o intervalo de habilidades dos examinados seria necessário um número muito grande de questões.

Em 2019, a Universidade Federal do Ceará (UFC, 2019) disponibilizou as notas de cortes do Sisu para entrada em seus cursos no ensino superior. Observam-se notas muito diferentes, como exemplo as encontradas no curso de Medicina (801) e de Ciências Sociais (443). Tal diferença, justifica a necessidade de uma prova que consiga estimar bem as habilidades dentro de toda a distribuição.

Com o TAM, seria possível reduzir o número de questões respondidas pelos indivíduos, e tornar as questões mais condizentes com as habilidades dos examinados. Mantendo o tempo de prova igual ao do teste linear, cada examinado poderia dedicar mais tempo as questões adequadas a seu nível de habilidade. Além de que questões não condizentes com a habilidade do examinado não apareceriam (uma questão muito fácil ou muito difícil para um examinado não apareceria).

1.2. Objetivos

Este estudo tem como objetivo geral montar um teste adaptativo multiestágio com a prova do ENEM de 2019, na área de Matemática. Como objetivos específicos pretende-se:

1. Entender a estrutura do banco de questões do ENEM e conseguir uma amostra de examinados.

2. Analisar e avaliar os itens da prova de Matemática, com os artifícios da Teoria de Resposta ao Item e da Teoria Clássica dos Testes.

3. Examinar a dificuldade dos módulos do teste adaptativo e procurar quais habilidades cada módulo se adequa.

4. Apresentar uma simulação de um TAM e o roteamento dos indivíduos nos estágios do teste.

5. Verificar o ajuste da estimativa das habilidades dos examinados, comparando estimativas de um indivíduo que respondeu o teste completo, com o mesmo indivíduo respondendo a simulação do TAM.

6. Comparar as notas simuladas pelo TAM com as notas atribuídas no ENEM.

1.3. Organização do Trabalho

O Capítulo 2 apresenta as referências teóricas utilizadas neste trabalho e alguns trabalhos correlatos sobre testes adaptativos e sobre o ENEM. O Capítulo 3 descreve a metodologia utilizada e as etapas consideradas para a montagem do Teste Adaptativo Multiestágio (TAM). No Capítulo 4 são apresentados os resultados, como o teste foi elaborado, e, também, faz-se uma comparação do teste linear com o TAM e com as notas do ENEM de 2019. Por fim, no Capítulo 5, são apresentadas as considerações finais, as discussões dos resultados, as limitações do trabalho e possíveis contribuições futuras, bem como algumas considerações sobre o curso de Engenharia da Computação.

CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA

2.1. Considerações Iniciais

Neste capítulo é apresentada a fundamentação teórica para mensurar a habilidade de examinados em uma prova. Apresenta-se também os conceitos dos testes adaptativos, por item ou por estágios. Destacam-se alguns tipos de provas aplicadas em larga escala que utilizam a metodologia dos testes adaptativos.

2.2. Conceitos e Técnicas Relevantes

2.2.1. Teoria Clássica dos Testes e Análise de Itens

No cenário de avaliações educacionais, procura-se uma forma de medir a aptidão de um indivíduo em um certo lócus do conhecimento. Historicamente, usam-se os escores brutos ou padronizados de uma prova (quantidade de itens que o examinado acertou). E esses resultados são dependentes do conjunto de itens que compõe a prova, implicando que os julgamentos e as avaliações estão associados a tal prova (ANDRADE et al., 2000). Sendo esses tipos de análises relacionadas a Teoria Clássica dos Testes (TCT).

A TCT consiste de conceitos e técnicas cujo objetivo fundamental é encontrar o resultado obtido por um examinado em um teste (SARTES et al., 2013). A TCT possui sete postulados. Como o estudo não utilizou a TCT para a montagem do TAM, não é necessário um aprofundamento dos postulados da TCT. No entanto, é válido ressaltar o primeiro postulado:

$$X = V + E, \tag{1}$$

em que X representa o resultado observado no teste, V o resultado verdadeiro do examinado e E um erro associado a medida. Ou seja, na TCT imagina-se que o escore bruto do examinado (quantos itens acertou) é uma função da verdadeira habilidade do aluno (seu escore real) somada com um erro. Um aprofundamento maior da TCT pode ser visto no trabalho de Grégoire e Laveault (SARTES et al., 2013 e GRÉGOIRE et al., 2002).

A despeito da TCT não ser utilizada para montar os módulos, uma parte importante deste trabalho, envolveu a análise dos itens da prova. E na verificação de adequação dos itens, a TCT proporciona alguns índices de itens que auxiliam nesse processo.

2.2.1.1. Índice de dificuldade

O Índice de Dificuldade (ID) consiste na proporção de examinados que acertaram um item (PITON-GONÇALVES et al., 2018).

$$ID = \frac{A}{n}, \tag{2}$$

em que A representa o total de indivíduos que acertaram o item, e n o total de participantes que responderam o item. Itens que tenham baixa quantidade de acertos são considerados mais difíceis se comparados com os itens com maior número de acertos. Ou seja, o ID com valor baixo indica um item difícil, e com valor alto, um item fácil.

É interessante ter um TAM que seja adequado a todos os níveis de habilidades presentes em uma população. Fazendo um paralelo com o ID, um banco de questões no qual se tem uma boa quantidade de ID diferentes traria uma diversidade de questões para as diferentes habilidades.

2.2.1.2. Índice de discriminação

O Índice de Discriminação (IDS) é uma medida para determinar a capacidade do item de diferenciar os examinados que tiveram um alto ou baixo escore (PITON-GONÇALVES et al., 2018). Um exemplo seria um item em que uma parte dos alunos de escores maiores acertaram tal item e pouquíssimos alunos de escore baixo acertaram-no. Para quantificar esse índice, pode-se usar a correlação Bisserial por Pontos.

$$r_{pb} = \frac{M_1 - M_0}{S_n}\sqrt{pq}, \tag{3}$$

em que $M1$ é a média do escore bruto entre os examinados que acertaram o item, $M0$ é a média entre os examinados que erraram o item, Sn é o desvio padrão populacional do teste

considerando todos os examinados, p é o percentual de acertos do item e q o percentual de erros (LE, 2012). Espera-se um valor positivo e distante de zero.

Um item com discriminação alta indica que a média do escore bruto entre os respondentes que acertaram tal item é alta, e a média do escore bruto entre os indivíduos que erraram tal item é baixa.

Casos em que o IDS se aproxima muito de zero, indica que o item não tem uma padronização entre os indivíduos de maiores escores e de menores escores. Como se a média do escore entre os examinados que acertaram e erraram é um número muito próximo.

Casos extremos, em que o IDS é negativo, tem-se que indivíduos com escores baixo acertam tal item, enquanto indivíduos com escores maiores erram, indicando que o item não foi bem elaborado.

2.2.2. Teoria de Resposta ao Item

A Teoria de Resposta ao Item (TRI) propõe modelos para a estimação da habilidade de um aluno. No contexto da TRI, denomina-se a habilidade do aluno como seu traço latente (ANDRADE et al., 2000). Nesse estudo, a TRI foi utilizada para estimar as habilidades dos examinados e montar o TAM.

Uma das vantagens da TRI perante o TCT é a possibilidade de comparar a habilidade de populações díspares, desde que expostas a uma prova que tenha alguns itens iguais. Ou até comparar indivíduos da mesma população que fizeram provas diferentes (ANDRADE et al., 2000). O TCT tem suas medidas considerando a prova em específico analisada, se apresentada outra prova, os resultados serão diferentes e não serão comparáveis.

A TRI é a teoria usado para calcular o traço latente dos alunos no ENEM (JATOBÁ et al., 2018). A cada ano tem-se uma prova com todos os itens diferentes, mas a população que realiza a prova é similar (alunos que terminaram o ensino médio, ou estão concluindo). Dessa forma, o resultado do ENEM é comparável em anos diferentes.

A TRI pode ser definida como: *"um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e da habilidade (ou habilidades) do respondente. Essa relação é sempre expressa de tal forma que quanto maior a habilidade, maior a probabilidade de acerto no item."* (ANDRADE et al., 2000, p: 7)

Existem vários modelos discutidos na literatura que dependem da natureza do item (dicotômico ou não), das populações envolvidas na prova (podendo ser uma ou mais), das habilidades estimadas (podendo ser uma ou mais). Neste estudo, considerou-se um modelo que tenha uma população única realizando a prova, que se estima apenas um traço latente, e que os itens foram avaliados como certo ou errado (dicotômicos).

O modelo utilizado foi o modelo de 3 parâmetros, no qual a probabilidade de um respondente acertar um item é uma função de seu traço latente, e dos parâmetros desse item.

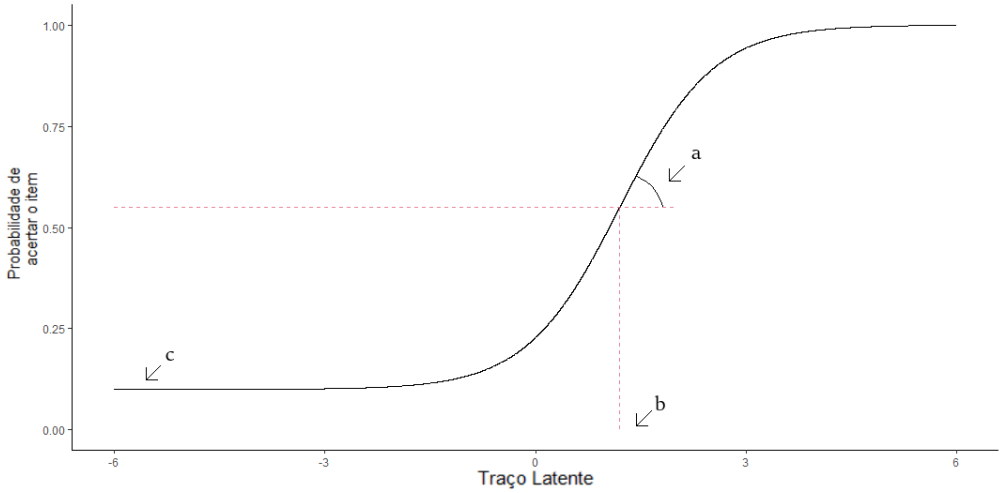
$$P(U_{ij}=1|\theta_j)=c_i+(1-c_i)\frac{1}{1+e^{-a_i(\theta_j-b_i)}}, \quad (4)$$

em que U_{ij} é uma variável dicotômica, que representa o resultado do examinado *j* para o item *i*, podendo assumir os valores *0* (quando o indivíduo erra o item) e *1* (quando o indivíduo acerta o item). E θ*j* representa o traço latente do indivíduo *j*, ou seja, sua habilidade. *P (Uij = 1 /* θ*j)* representa a probabilidade do examinado *j* acertar ao item *i*. 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 do item *i* que representa a probabilidade de um indivíduo com baixa habilidade encontrar o item correto.

Na Figura 4 tem-se um exemplo da Curva Característica do Item (CCI), que é a função de probabilidade de acertar um item pelo traço latente. O modelo de 3 parâmetros baseia-se no fato de examinados com maior habilidade possuem maior probabilidade de acertar o item, com esta relação sendo não linear (ANDRADE et al., 2000). Sendo esse o

caso na figura, o gráfico mostra que quanto maior a habilidade, maior a probabilidade de acerto, em que a curva é crescente em formato de "S".

Figura 4 - CCI.



Fonte: Elaborado pelo autor.

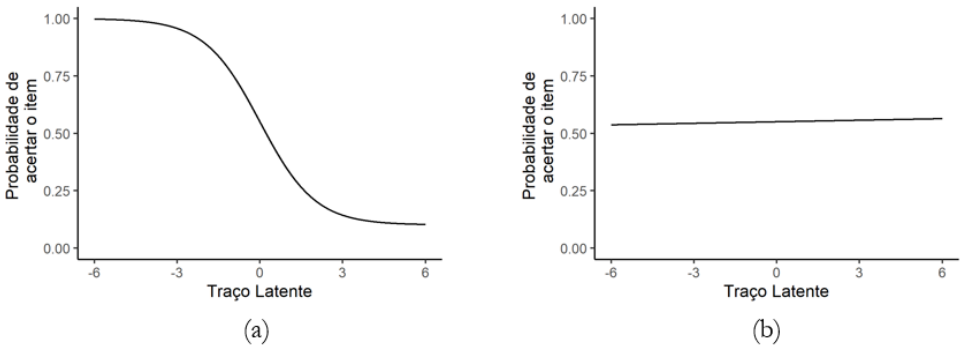
Os parâmetros do item (*a*, *b* e *c*) modelam a CCI. Em particular, o parâmetro *a* é proporcional a derivada da tangente da curva no ponto de inflexão. Dessa forma, esse parâmetro indica qual a inclinação da curva, indica o quão rápido a curva cresce.

Valores negativos de *a* (como mostrado na Figura 5) indicam que traços latentes menores, tem maior probabilidade de acertar um item do que traços latentes maiores. Sendo isso, o oposto do proposto pelo modelo. Caracterizando, então, o item como ruim.

Valores próximos de zero (como mostrado na Figura 5), indicam que a inclinação é pequena, ou seja, o crescimento da curva é devagar. A probabilidade de acertar o item é parecida para o intervalo de habilidade. Indivíduos com habilidades maiores possuem a probabilidade de acerto muito próxima de indivíduos com habilidades menores. Caracterizando, novamente, o item como ruim.

O parâmetro *a* é determinado como discriminação. Diferentemente da TCT, a discriminação não é uma correlação (valor entre -1 e 1), esperam-se valores entre 0 e +2, sendo mais apropriados aqueles maiores que 1 (ANDRADE et al., 2000).

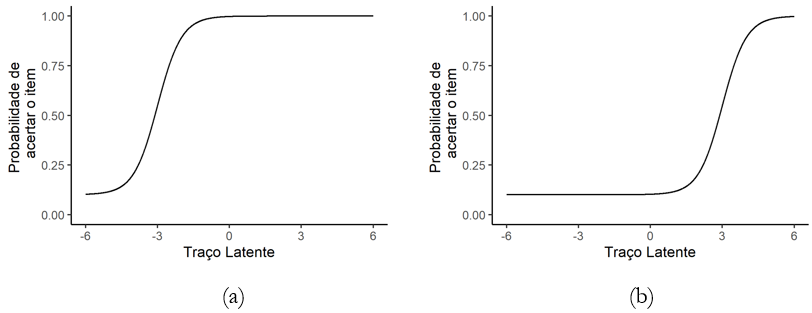
Figura 5 – CCI com valores de a inadequados. (a) Valor Negativo. (b) Valor Próximo de Zero.



Fonte: Elaborado pelo autor.

O parâmetro *b*, indica a dificuldade de um item. Ele está na mesma escala do traço latente do indivíduo, e representa qual a habilidade necessária de um examinado para ter uma probabilidade de acerto igual a $(1 + c)/2$. Apesar de *b* ter a mesma conotação da TCT, seu significado é diferente e não representa uma proporção. Quando *b* tem valores altos, precisa-se de um traço latente elevado para uma probabilidade alta de acertar o item (ou seja, um item difícil). Para o caso de *b* pequeno, com traços latentes menores, a probabilidade de acertar o item já é alta (ou seja, um item mais fácil). A Figura 6 ilustra um item com o *b* baixo e alto.

Figura 6 – CCI com b diferente. (a) Valor baixo. (b) Valor alto.



Fonte: Elaborado pelo autor.

Por fim, tem-se o parâmetro *c*, que é conhecido como a chance de acerto ao acaso (ou "chute"). Isso é, indivíduos com traço latente baixo podem escolher uma resposta ao acaso e ainda assim possuem uma probabilidade de acertar o item (ANDRADE et al., 2000). Intuitivamente, imagina-se que em um teste de K alternativas a probabilidade de acertar ao acaso é $1/K$ (No caso do ENEM $1/5$). Entretanto, na prática, o acerto ao acaso é mais complexo do que o examinado simplesmente escolher uma das alternativas aleatoriamente. Empiricamente, estimativas desse parâmetro tendem a ser um pouco menores que $1/K$ (VAN DER LINDEN et al., 2015).

Diferentemente do escore bruto apresentado na TCT, em que seu valor é igual a quantidade de itens certos por um examinado, na TRI a habilidade pode assumir qualquer valor no intervalo dos reais. Na TRI, precisa-se estabelecer uma origem e uma unidade de medida para definir uma escala, sendo estes representados, respectivamente, pela média e desvio-padrão das habilidades (ANDRADE et al., 2000). Neste estudo, utilizou-se a métrica $(0,1)$ com média 0 e desvio-padrão 1. No ENEM utiliza-se a escala $(500,100)$, média 500 e desvio-padrão 100. Na prática, não faz diferença qual a métrica escolhida, o relevante são as relações de ordem existentes entre seus valores (ANDRADE et al., 2000).

O modelo de 3 Parâmetros da TRI parte de alguns pressupostos que são importantes para a estimação dos parâmetros (traços latentes e parâmetros dos itens). É necessário que o postulado da unidimensionalidade seja satisfeito (ANDRADE et al., 2000). Tal postulado pressupõe a unidimensionalidade do teste. Imagina-se que o teste esteja medindo uma única habilidade (Por exemplo, habilidade em matemática, não habilidade em matemática e português). Para satisfazê-lo, é suficiente existir um fator dominante (habilidade) que seja responsável pelas respostas dos indivíduos. O teste estaria mensurando esse fator (individual).

A outra suposição usual é de independência local, assume-se que os indivíduos ao fazerem o item não mudam sua habilidade. Os itens são independentes entre si, ao responder um item, o examinado não adquire conhecimento, nem perde conhecimento para responder o próximo item (LAROS, 2021).

2.2.3. Informação do Item

Uma métrica importante no contexto da TRI, é a Informação do Item. Procura-se saber o quão sensível o modelo é para pequenas mudanças no traço latente. Isto é, quanto que a probabilidade de acertar o item altera quando o traço latente altera. Uma maneira de se mensurar isso é usando a informação de Fisher.

$$I_X(\theta) = \sum_{x \in X} \left(\frac{d}{d\theta} logf(x \mid \theta)\right)^2 f(x \mid \theta), \tag{5}$$

em que I_X é a informação de Fisher, X é uma Variável Aleatória Discreta, θ é um parâmetro da função f e $f(x \mid \theta)$ é uma função de probabilidade (LY et al., 2017).

A derivada $\frac{d}{d\theta} logf(x \mid \theta)$ está relacionada com a sensibilidade do modelo (como que mudanças de θ afetam a função f). Dessa forma, a informação de Fisher mede a sensitividade geral de f, fazendo uma média ponderada da sensitividade em cada possível resultado da variável X(LY et al., 2017).

Dessa maneira, a informação de Fisher no contexto da TRI pode ser entendida como a sensitividade geral de um item quando se muda o traço latente, em que os parâmetros do item são conhecidos. Na TRI, X é a variável aleatória dicotômica do acerto ou não do item (0 ou 1), θ é o traço latente, $f(x \mid \theta)$ é a função de probabilidade de acertar o item (com a, b e c conhecidos).

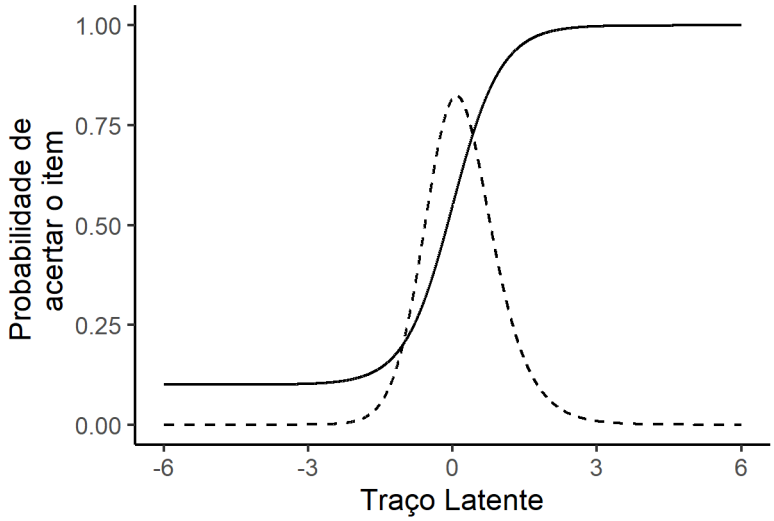
No modelo de 3 Parâmetros, a função de informação de Fisher pode ser reescrita como (ANDRADE et al., 2000):

$$I_i(\theta) = a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i}\right]^2, \tag{6}$$

em que, $P_i(\theta)$ é a probabilidade de acertar o item i em função de θ. $Q_i(\theta)$ é a probabilidade de errar o item i, isto é, $(1 - P_i(\theta))$.

A Figura 7 mostra um exemplo de um item, com sua informação ao longo do traço latente. A informação de Fisher também pode ser entendida como a recíproca da precisão de um parâmetro (BAKER, 2001). No contexto da TRI, se a informação de um item for grande para um traço latente determinado, significa que os examinados com habilidades próximas de tal traço latente terão estimativas mais precisas. Ou seja, se um item tiver uma informação alta para uma habilidade de um candidato, esse item seria bom ser apresentado, melhorando ainda mais a estimação da habilidade do candidato.

Figura 7 - CCI e Informação do Item. Em pontilhado tem-se a informação do item.



Fonte: Elaborado pelo autor.

A informação de um determinado teste é dada pela soma das informações dos itens contidos no teste. Para o caso do TAM, pode-se encontrar a informação de cada módulo como a soma das informações dos itens pertencentes a esse módulo. E a informação de um teste do TAM seria a soma da informação dos módulos que um determinado examinado respondeu.

2.2.4. Métodos de Estimação

O modelo de 3 Parâmetros da TRI, consiste de um conjunto de parâmetros de itens (*a*, *b* e *c*) e um conjunto de parâmetros dos examinados (os traços latentes). Em geral, tais parâmetros são desconhecidos, e apenas as respostas dos respondentes aos itens são conhecidas.

No caso deste trabalho, têm-se dois cenários diferentes. O primeiro cenário se refere a montagem dos módulos. Precisa-se distinguir a dificuldade dos itens, e a informação que eles têm para os traços latentes. Ou seja, precisa-se estimar os parâmetros dos itens. Nesse caso, os traços latentes são desconhecidos.

No segundo cenário, procura-se encontrar a habilidade dos examinados quando apresentados aos módulos do TAM. Ou seja, precisa-se estimar os traços latentes. Nessa etapa, os parâmetros de itens já foram estimados e serão considerados como parâmetros conhecidos (diferente do caso usual).

Tanto na estimativa dos parâmetros de itens, quanto na estimativa dos traços latentes, pode-se utilizar o Método da Máxima Verossimilhança. Muitas vezes esse método não possui solução exata, o que leva a utilização de processos iterativos como *Newton Raphson* ou *Scoring de Fisher*. No entanto, quando o número de parâmetros a serem estimados for muito grande, a Máxima Verossimilhança leva a uma enorme exigência computacional, então, optou-se por usar outro método. (ANDRADE et al., 2000)

Neste trabalho, a estimação dos parâmetros dos itens seguiu o Método da Máxima Verossimilhança Marginal. Encontra-se a função de Verossimilhança, a qual é marginalizada em relação ao traço latente, e procura-se o máximo dessa nova função. O processo de estimação dos parâmetros dos itens, também é conhecido como calibração dos itens.

O modelo de 3 parâmetros informa a probabilidade de um examinado *j* acertar o item *i*. Ou seja, *1* menos a equação do modelo informa a probabilidade de o examinado errar o item. Quando se tem o vetor de respostas dos examinados, a verossimilhança é a probabilidade de tal vetor acontecer. Com a suposição de independência local, e

independência entre indivíduos, a Verossimilhança se torna o produtório da probabilidade do examinado *j* ter acertado ou errado o item *i*, de acordo com o vetor de respostas.

$$L(\theta_j, a_i, b_i, c_i) = \prod_{i=1}^{I} \prod_{j=1}^{N} P(U_{ij} = u_{ij} \mid \theta_j, a_i, b_i, c_i), \tag{7}$$

em que I é a quantidade de questões no teste, N é a quantidade de indivíduos que fizeram a prova, u_{ij} é a resposta observada do indivíduo j ao item i e $P(U_{ij} = u_{ij} \mid \theta_j, a_i, b_i, c_i)$ é a probabilidade da variável U_{ij} ser igual a resposta observada u_{ij}. Ou seja:

$$P(U_{ij} = u_{ij} \mid \theta_j, \zeta_i) = P(U_{ij} = 1 \mid \theta_j, \zeta_i)^{u_{ij}} (1 - P(U_{ij} = 1 \mid \theta_j, \zeta_i))^{1-u_{ij}} \tag{8}$$

O vetor ζ_i se refere aos parâmetros do item *i*.

No processo de Máxima Verossimilhança Marginal, as habilidades dos examinados não são conhecidas, então usa-se algum artifício para a Verossimilhança não depender mais das habilidades. Determina-se uma distribuição para tais habilidades e marginaliza-se a Verossimilhança (integra-se a função de verossimilhança com relação a distribuição do traço latente). Neste estudo, a distribuição das habilidades foi considerada como a distribuição Normal, e a métrica utilizada foi *(0,1)*. A escolha de uma métrica é possível (e necessária), pois, no caso de parâmetros de itens e habilidades desconhecidos, mais de um conjunto de parâmetros produz o mesmo valor no modelo. Esse problema é conhecido como a falta de identificabilidade do modelo. (ANDRADE et al., 2000)

$$L(\zeta) = \int \prod_{i=1}^{I} \prod_{j=1}^{N} P(U_{ij} = u_{ij} \mid \theta_j, \zeta_i) Normal(\theta \mid 0,1) d\theta \tag{9}$$

A maximização da equação 9 através de algum método iterativo gera a estimativa dos parâmetros dos itens. O método iterativo utilizado foi o algoritmo de expectativa-maximização (EM), que é o método padrão implementado na biblioteca MIRT do R (CHALMERS, 2012).

Em relação a estimação das habilidades dos examinados, uma vez que os parâmetros dos itens são supostos conhecidos (fixados pelos valores obtidos pela etapa anterior), existem vários métodos possíveis para tal estimação. Por exemplo, pode-se usar o método da máxima verossimilhança com os itens conhecidos. Neste estudo, utilizou-se a estimação pela média da posteriori (EAP). Em que se estima os traços latentes a partir da esperança da posteriori (ANDRADE et al., 2000).

$$\hat{\theta}_j = \frac{\int_{\mathbb{R}} \theta P(u_{.j} \mid \theta, \zeta) Normal(\theta \mid 0,1) d\theta}{\int_{\mathbb{R}} P(u_{.j} \mid \theta, \zeta) Normal(\theta \mid 0,1) d\theta} \tag{10}$$

Esse método está implementado na biblioteca MIRT (CHALMERS, 2012).

2.2.5. TAC-I

O TAC-I é um teste adaptativo computadorizado que usa um algoritmo para administrar os itens do teste (YAN et al., 2014). O indivíduo recebe o teste, e responde as questões conforme elas aparecem. A cada questão respondida, estima-se a habilidade do examinado, e então a próxima questão é escolhida. 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 questões 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 et al., 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)
- Tamanho do Teste.
- Método de Estimação das habilidades.
- Calibração dos itens.
- Critérios de seleção dos itens.
- Critério de parada do teste.

No TAC-I, as questões são escolhidas adaptadas ao traço latente do indivíduo. Entretanto, isso pode causar um problema no balanceamento do conteúdo das questões, conforme o examinado responde as questões, pode ser que ele tenha um comportamento que leve a responder questões repetidas 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 questões muito bem montado, com conteúdo variado e de dificuldades variadas, e algoritmos mais sofisticados. (SILVA et al., 2019)

No TAC-I, as questões não podem ser revisadas uma vez que respondidas. Caso houvesse a possibilidade de revisar as questões, o examinado poderia alterar o traço latente estimado em determinado ponto, o que atrapalha na ordem de aparição das questões. O examinado não tem conhecimento dos próximos itens que podem aparecer, e podem 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 trabalho, 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).

2.2.6. TAM

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 trabalho, foi feito 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 questões que envolvam o conteúdo esperado em todos os módulos, adequando a questão a dificuldade do módulo (YAN et al., 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 alunos, nele têm-se questões de dificuldades médias, em que se tenta fazer uma estimação aproximada da habilidade do aluno. Caso o aluno tenha um resultado bom, ele será apresentado a um módulo mais difícil no estágio seguinte. E mantém-se essa lógica até o último estágio.

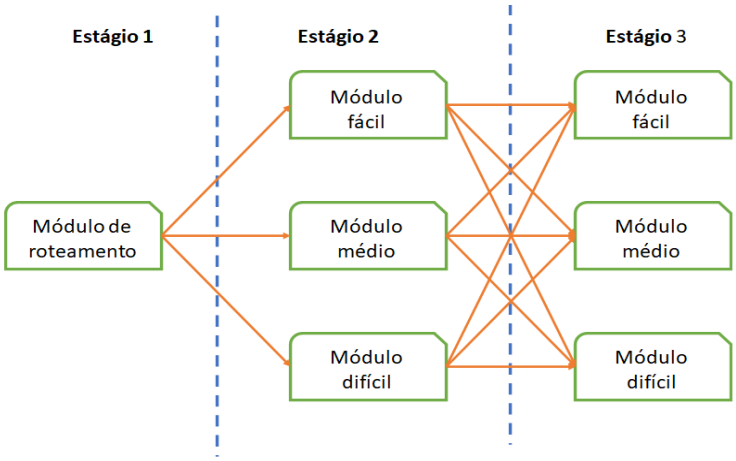
Dentro de um estágio, o examinado tem liberdade para revisar as questões e alterar as alternativas. Entretanto, quando passa para o estágio seguinte, as questões dos estágios anteriores deixam de ser acessíveis.

Quando se tem um banco de questões 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 8 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 8- Exemplo de TAM com 3 estágios.



Fonte: Elaborado pelo autor.

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.

2.3. Trabalhos Relacionados

2.3.1. Provas em larga escala com TAM

A transposição de testes lineares para testes multiestágio é um assunto extremamente debatido mundo a fora, em que já existem testes de larga escala que utilizam da metodologia TAM. Um dos principais entre esses é o *Graduate Record Examination* (GRE). O GRE é utilizado como uma das provas necessárias para o acesso a pós-graduação nos Estados Unidos.

O GRE foi primeiro introduzido em 1949, e recebeu algumas mudanças ao longo dos anos. Em 1992 deixou de ser um teste feito em papel para ser um teste feito em computador, e no ano seguinte já se introduziu o conceito do teste adaptativo. Em 2011 mudou-se o formato do teste para o TAM (WENDLER et al., 2014).

Para fazer a mudança no formato da prova, o *Educational Testing Service* (ETS - Responsáveis pela Elaboração do GRE) fez estudos de diversos fatores antes de decidirem qual formato mais se adequaria a prova.

Os principais objetivos considerados para fazer a mudança no formato foram: melhorar a estimação das habilidades para uma população tão diversa fazendo a prova, apoio a revisões na escala usada, manter o tempo de teste em até 4 horas ou menos e oferecer uma experiência mais flexível para os examinados permitindo que voltem, pulem ou refaçam questões dentro de um mesmo módulo (WENDLER et al., 2014).

Um ano após a mudança do GRE para o modelo de TAM, o ETS avaliou informações empíricas dos resultados obtidos e com isso conseguiram revisar as questões, criar e remover algumas. A qualidade do teste está sendo mantida e aprimorações contínuas do teste são feitas (WENDLER et al., 2014).

Outro teste contemporâneo que utiliza do artificio do TAM é o *Programme for International Student Assessment* (PISA). O PISA é uma prova internacional em que estudantes de diversos países são apresentados a questões sobre situações da vida real e devem respondê-las com seus conhecimentos. O PISA visa ajudar os países a

compreenderem como seus sistemas educacionais estão evoluindo, e encorajarem os países a aprenderem entre eles meios de construírem sistemas educacionais mais justos e inclusivos (PISA, 2021).

O PISA implementou o TAM em 2018 no domínio da leitura. Utilizaram a TRI para estimação dos parâmetros e conseguiram uma precisão melhor na medida das habilidades dos examinados, especialmente nas habilidades ao extremo da distribuição (YAMAMOTO et al., 2019).

2.3.2. Trabalhos Relacionados

Existe uma série de pesquisadores contemporâneos que fazem análises sobre a prova do ENEM. Os autores Piton-Gonçalves e Almeida (PITON-GONÇALVES et al., 2018) fazem uma análise clássica da prova do ENEM de 2012, procurando saber o ID e o IDS dos itens, e como eles são distribuídos na prova. O trabalho do Ricarte (RICARTE, 2016) faz uma comparação de modelos diferentes da TRI no contexto do TAM. O autor Piton-Gonçalves (PITON-GONÇALVES, 2020) analisou a viabilidade de aplicar-se um teste adaptativo para o Exame Nacional de Desempenho dos Estudantes (ENADE). O autor Souza (2019) fez um trabalho em que realizam um TAC-I para a prova do ENEM.

2.4. Considerações Finais

Neste capítulo teceram-se algumas considerações sobre as teorias importantes que fundamentaram o trabalho desenvolvido, e sobre as técnicas utilizadas para estimação de habilidade dos alunos e sobre o funcionamento de dois tipos de testes adaptativos. Ao final, discorreu-se sobre as provas em larga escala que aplicam o formato de teste multiestágio e trabalhos relacionados. A fundamentação teórica é importante para entender o desenvolvimento do trabalho. No capítulo seguinte, discutiu-se sobre a prova do ENEM de matemática de 2019, as decisões tomadas para a montagem do TAM, o processo de amostragem e as limitações.

CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO

3.1. Considerações Iniciais

A primeira parte deste estudo consistiu em procurar informações e bases de dados acerca da prova do ENEM. Felizmente, o MEC disponibiliza os microdados das provas do ENEM (INEP, 2021). Nos microdados, têm-se duas tabelas disponíveis como arquivos de valores separados por vírgulas. A primeira tabela contém os microdados dos participantes, possui informações acerca de cada participante individualmente. A segunda, se refere as provas, com informações sobre quais itens estão em cada prova e em cada tipo de prova.

A partir dos microdados, a primeira decisão tomada foi a escolha da 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).

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 da TCT. E a biblioteca ggplot2 (WICKHAM, 2016) foi a principal usada na construção dos gráficos.

3.2. Estrutura da Prova

A prova do ENEM é composta por quatro áreas Ciências Humanas, Ciências da Natureza, Linguagens e Códigos e Matemática, além da redação. Este trabalho consiste em mostrar um exemplo de como a prova do ENEM pode ser montada como um TAM, dessa forma, optou-se por usar apenas uma área da prova, matemática.

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

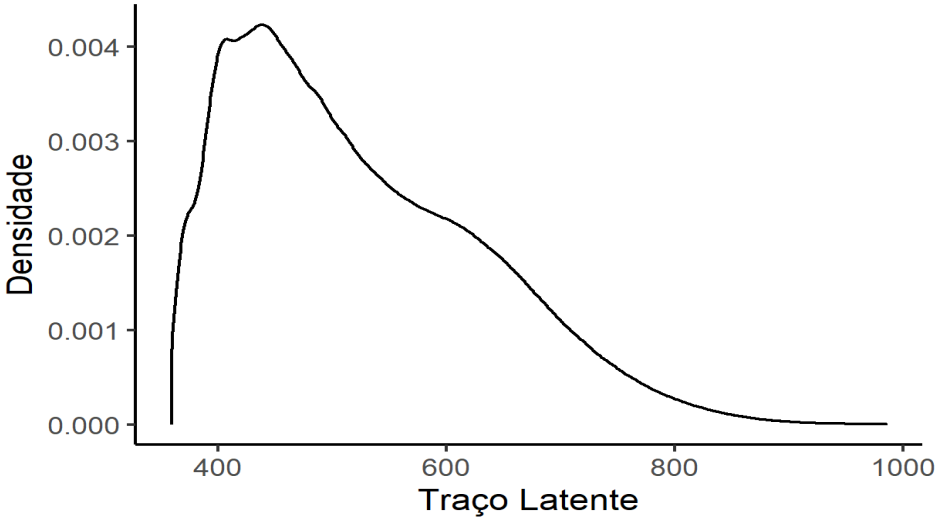
Entre as informações presentes nos microdados, tem-se a nota atribuída ao candidato em determinada área. Então, explorou-se a distribuição das notas dos examinados na prova de matemática. Nesse processo, excluíram-se os respondentes que não marcaram nenhuma questão (apesar de estarem presente no dia). Na Tabela 1 estão algumas estatísticas dessas notas, e na Figura 9 está a distribuição das notas. A métrica utilizada no ENEM possui valor de referência como 500 e valor de dispersão igual a 100 (ENEM, 2012). Pela Tabela 1 e a Figura 9, visualiza-se que existe uma concentração muito grande de examinados com notas abaixo de 600, a média das notas é próxima de 500, e o desvio-padrão próximo de 100. Outro detalhe visto na Figura 9, é o fato de a distribuição de notas ter uma cauda a direita, enquanto a esquerda tem uma concentração grande (Assimetria positiva). Isso pode ser um indicativo que os indivíduos de traços latentes maiores conseguiram ser bem diferenciados e bem estimados, enquanto os indivíduos de traços latentes menores parecem que foram estimados com valores próximos.

Tabela 1 - Estatísticas Descritivas Notas de Matemática ENEM 2019

Mínimo	Mediana	Média	Máximo	Desvio Padrão
359.0	501.2	523.3	985.5	108.8

Fonte: Microdados Enem.

Figura 9 - Distribuição das notas de Matemática



Fonte: Elaborado pelo autor.

Em seguida, as versões (tipos) da prova de matemática foram analisadas. As provas são divididas em cores (e códigos), e em cada tipo de prova os itens são dispostos em ordens diferentes. Na Tabela 2 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.

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, 2021). Optou-se por excluí-las também da análise. Existem poucos alunos que se encaixaram nessas categorias em relação ao total de alunos. Observou-se que a prova laranja possuía três itens diferentes das demais, o que, novamente, tornaria inviável a estimação desses itens.

Tabela 2 - 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

Fonte: Microdados Enem.

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.

3.3. Amostragem

Ao usar todos os participantes para a estimação dos parâmetros de itens, ou das habilidades, teria uma quantidade de processamento muito grande, consumindo muito tempo para executar. Portanto, optou-se por fazer uma amostragem dos participantes. 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 de acordo com 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 3 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. A tabela completa pode ser vista no Apêndice B.

Tabela 3 - 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

Fonte: Elaborado pelo autor.

Com a base normalizada, torna-se possível retirar uma amostra aleatória dos examinados, sem se preocupar com o tipo de prova. Escolheu-se retirar uma amostra de tamanho cinquenta mil. A probabilidade de cada um dos examinados aparecer na amostra é a mesma.

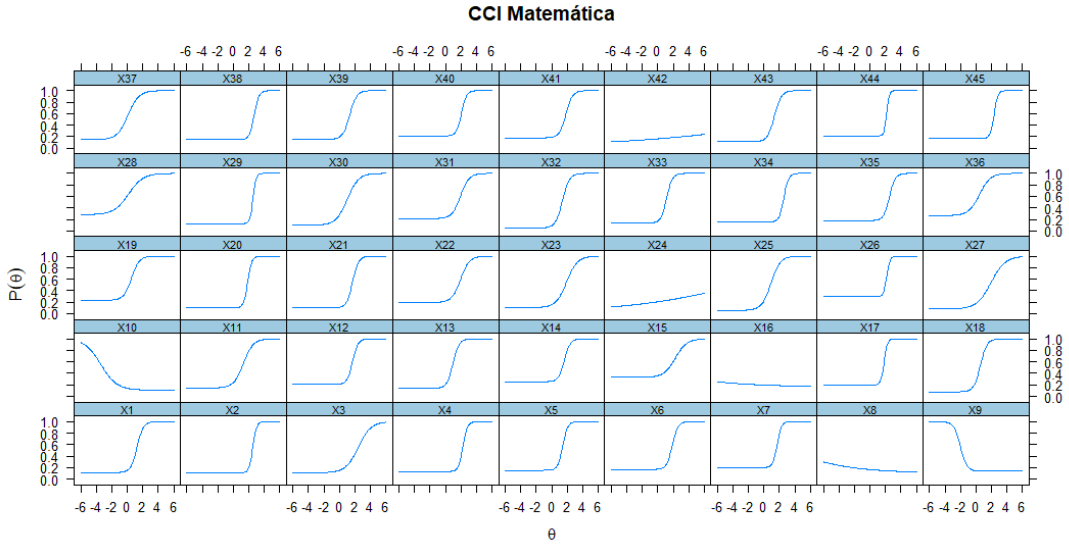
3.4. Análise dos Itens

Uma vez pronta a amostragem dos examinados, realizou-se 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). Para tornar o estudo replicável, foi utilizada a semente 1031 na geração da amostra, sendo a versão do software R a 4.0.3.

Com a finalidade de garantir que essa amostra é representativa, foram feitos testes com 20 amostras de tamanho diferentes (15 delas com 50 mil examinados, e as outras com 10 mil). Também se fez um teste considerando apenas os alunos que fizeram a prova Azul. Os testes consistiram em calibrar os itens. Desses resultados, foi possível perceber que a amostra com semente 1031 tem resultados concordantes com a população que realizou a prova.

A calibração dos itens consiste em encontrar os parâmetros dos itens, na Figura 10 é possível encontrar a CCI de cada um dos 45 itens. A calibração completa pode ser encontrada no Apêndice A.

Figura 10 – CCI prova de matemática



Fonte: Elaborado pelo autor com Mirt.

Itens com discriminação próxima de zero ou com discriminação negativa são caracterizados como itens ruins. Na prova, encontrou-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 questões considerando artifícios da TCT.

Com o pacote CTT (WILLSE et al., 2008) do R, encontrou-se o índice de dificuldade e o índice de discriminação. A Tabela 4 mostra o número do item, sua classificação, o seu ID e o seu IDS, isso para os itens classificados como ruins pela análise. Também se adicionou dois itens classificados como bons a efeito de comparação. Esses resultados são apresentados no Apêndice A, para todos os itens da prova.

Tabela 4 - Análise Clássica de alguns itens

Item	Classificação	ID	IDS
8	Ruim	0.17	-0.025
9	Ruim	0.19	-0.007
10	Ruim	0.16	-0.030
16	Ruim	0.20	-0.001
24	Ruim	0.21	0.025
42	Ruim	0.16	0.001
37	Bom	0.55	0.309
39	Bom	0.28	0.351

Fonte: Elaborado pelo autor.

Desses resultados, é possível perceber que os itens tidos como ruins, possuem um ID muito baixo, e um IDS ou negativo ou próximo de zero. Corroborando com a ideia que são itens ruins e pouco discriminativos. Esses itens podem prejudicar a montagem do TAM, entretanto foram mantidos na análise, evidenciando a bem sucedida montagem do TAM, ignorando itens ruins.

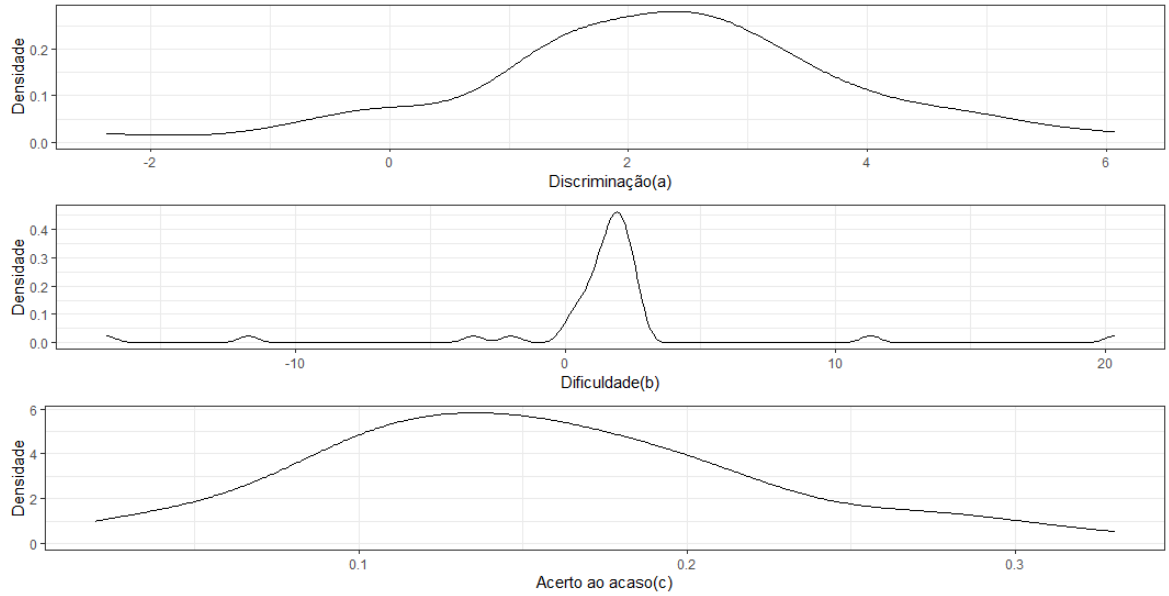
Em seguida, explorou-se a distribuição dos parâmetros de itens no contexto da TRI. É importante ter uma ideia de como a dificuldade dos itens está distribuída para verificar se o banco de itens terá condições de montar um TAM adequado a população de examinados. E verifica-se se os parâmetros estão de acordo com o esperado para um modelo da TRI.

Na Figura 11 tem-se a distribuição 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.

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 11 - Distribuição dos parâmetros dos itens.



Fonte: Elaborado pelo autor.

Neste contexto, é razoável supor que o banco de itens da prova de 2019 tem uma boa qualidade para a realização de um teste adaptativo.

3.4. Construção dos Módulos

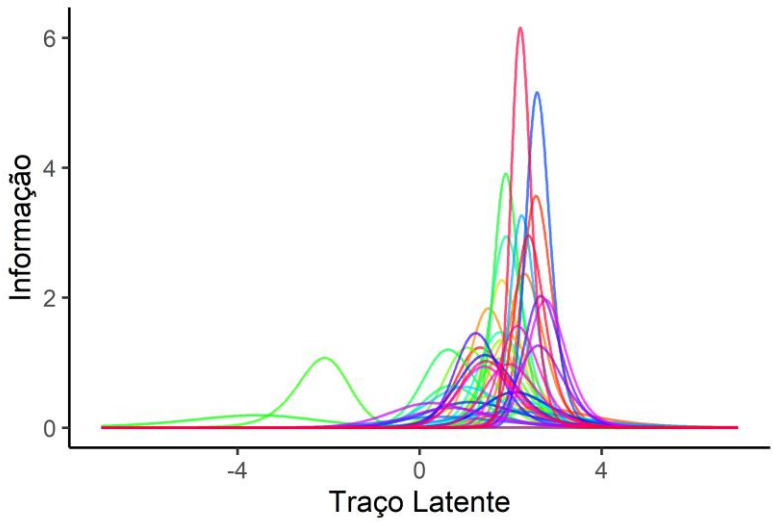
Uma vez que se tem os parâmetros dos itens, pode-se escolher alguns critérios de seleção e começar a construção de cada estágio e cada módulo do TAM.

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 de Fisher. Itens com maiores informações para um dado traço latente serão escolhidos para entrarem nos módulos. A informação de Fisher 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). Na Figura 12 tem-se a informação de todos os 45 itens.

Figura 12 - Informação dos itens



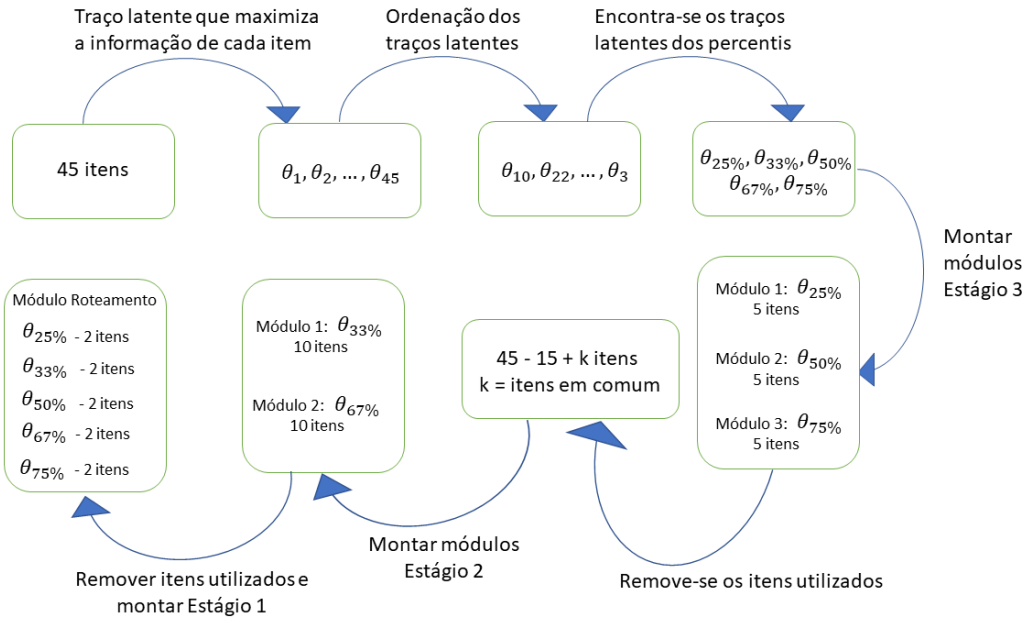
Fonte: Elaborado pelo autor.

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 foram 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 passarão 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 13 mostra um quadro esquemático das etapas de construção do módulo. Os itens escolhidos são aqueles que maximizam a informação de Fisher. Para garantir o limite de itens em comum, escolhe-se o item em comum menos informativo de um módulo e substitui ele pelo próximo item mais informativo dentro do banco de questões, e em seguida, faz-se o mesmo procedimento para o outro módulo. E esse procedimento se repete até atingir o limite de itens em comum.

Figura 13 - Fluxograma construção dos módulos



Fonte: Elaborado pelo autor.

3.5. Dificuldades e Limitações

Este estudo encontrou uma limitação no banco de itens utilizados, apenas 45 itens foram considerados no banco de questões, impedindo uma flexibilidade maior na quantidade de questões em cada módulo. As questões utilizadas limitaram a estimação ao longo da distribuição de habilidades, eram concentradas em traços latentes altos. Algumas questões eram inadequadas, e acabaram por não entrarem no TAM. Outra limitação, está no fato de não ter sido explorado outros métodos de estimação (a exemplo de métodos Bayesianos).

A maior dificuldade encontrada foi relacionada com os testes para decidir quais critérios utilizar na montagem dos módulos. Cada teste envolvia uma quantidade de processamento alta (exemplo da estimação dos parâmetros em uma amostra de 50 mil examinados), o que tornava o processo muito demorado. Outra dificuldade, foi a normalização da base, em que cada prova contém uma ordenação diferente e se não tratados, traria resultados não significativos.

3.6. Considerações Finais

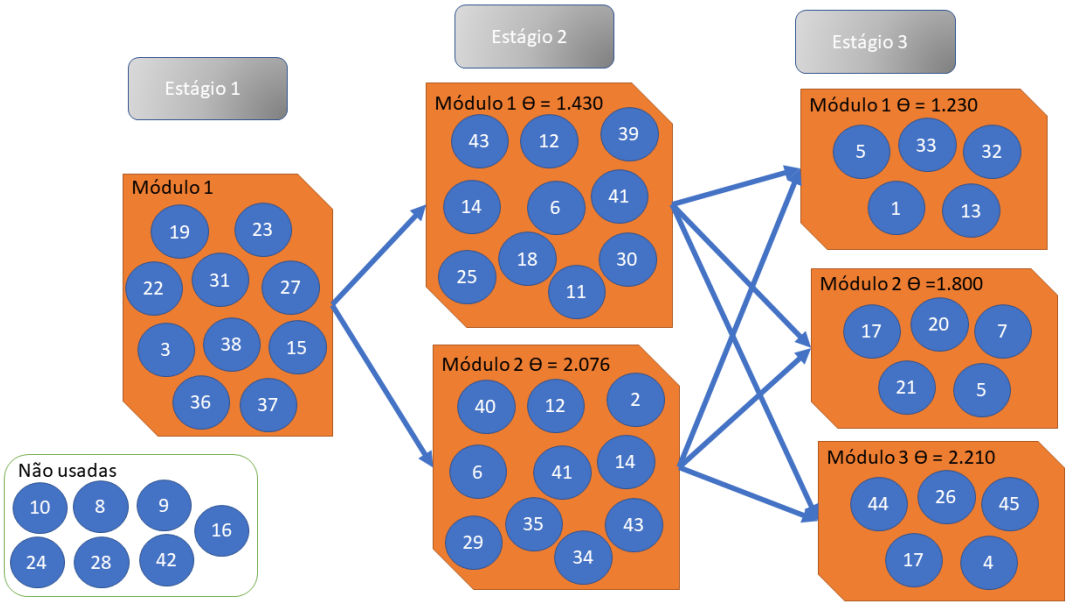
Este capítulo apresentou a estrutura da prova, identificando quais tipos de provas foram escolhidos. Em seguida, apresentou a lógica utilizada na normalização da base e da amostragem da população. Uma análise nos itens da prova de 2019 foi feita, encontrando 6 itens classificados como ruins. Por fim, mostrou-se a lógica utilizada para montagem dos estágios e dos módulos. O próximo capítulo apresenta o TAM construído e uma comparação das estimativas de habilidades dos indivíduos quando apresentados a prova completa e ao TAM.

CAPÍTULO 4: RESULTADOS

4.1. TAM Proposto

Após a montagem dos estágios, a Figura 14 ilustra o painel construído para a prova de 2019. As questões classificadas como ruins não foram selecionadas para formarem os módulos.

Figura 14 - Formato proposto para o teste

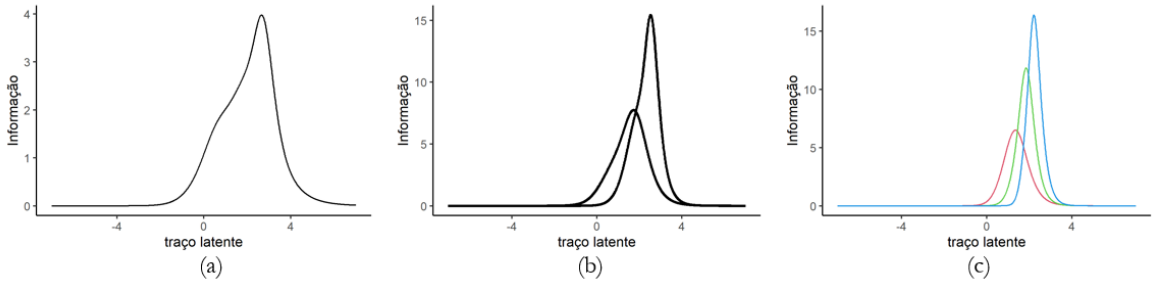


Fonte: Elaborado pelo autor

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 questões 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, 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.

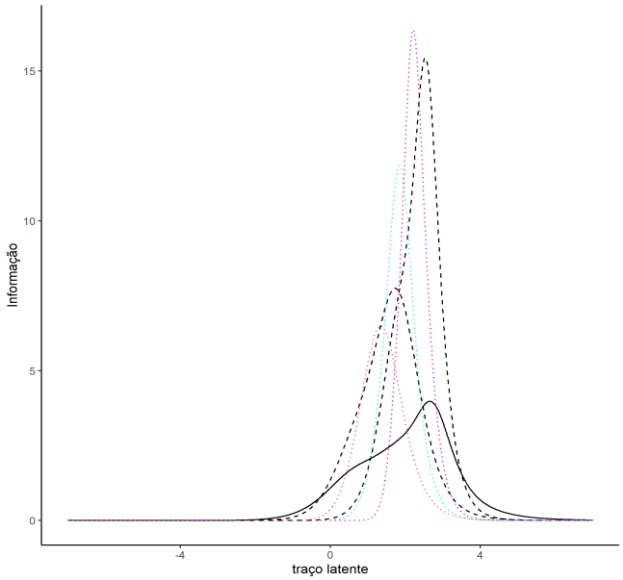
A Figura 15 mostra a informação de cada módulo, e tem-se uma ideia de quão preciso serão as habilidades estimadas em cada módulo. A Figura 16 apresenta a informação dos estágios e módulos sobrepostas.

Figura 15 - 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



Fonte: Elaborado pelo autor.

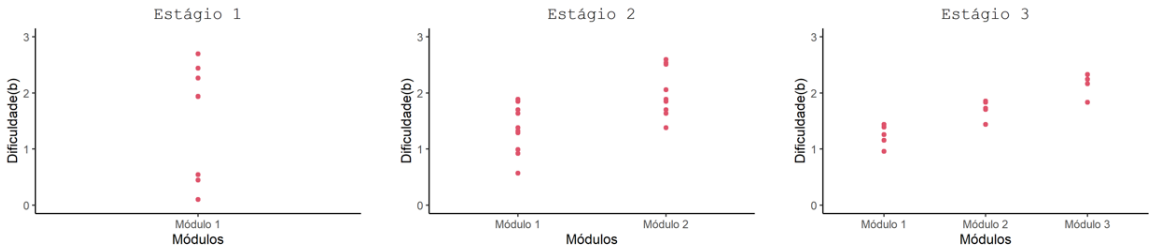
Figura 16 - Informação dos 3 Estágios sobreposta. Pontilhado colorido se refere ao estágio 3, pontilhado preto se refere ao estágio 2, linha contínua preta se refere ao estágio 3.



Fonte: Elaborado pelo autor.

Com os módulos montados, explorou-se a distribuição da dificuldade das questões (parâmetro *b*) em cada um. A Figura 17 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.

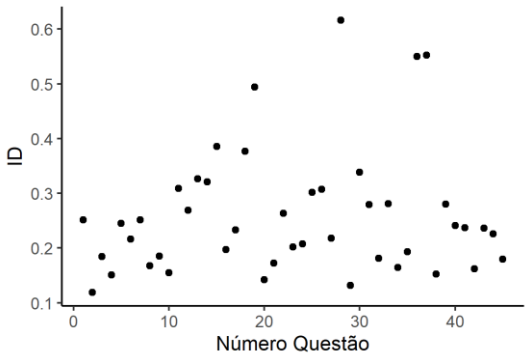
Figura 17 - Dificuldade dos módulos divididas pelos Estágios



Fonte: Elaborado pelo autor.

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 o ID dos itens, o item com maior média de acertos possui *62%* de acerto, a Figura 18 mostra como o ID está distribuído. A maior parte das questões tiveram média de acertos menor que *40%.*

Figura 18 - Média de acertos de cada item.



Fonte: Elaborado pelo autor.

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

4.2. Verificação das Estimativas da Habilidade

Apesar do banco de questões não ser o mais adequado, ainda assim foi possível montar o TAM. Na etapa final deste estudo, verifica-se se o TAM montado tem estimativas das habilidades próximas das estimativas usando o teste completo. Ou seja, se com uma redução de 45 para 25 questões é possível encontrar estimativas de habilidade parecidas com o teste com as 45 questões (indicando que muitas questões respondidas pelos examinados não agregam tanta informação para a habilidade de tal aluno).

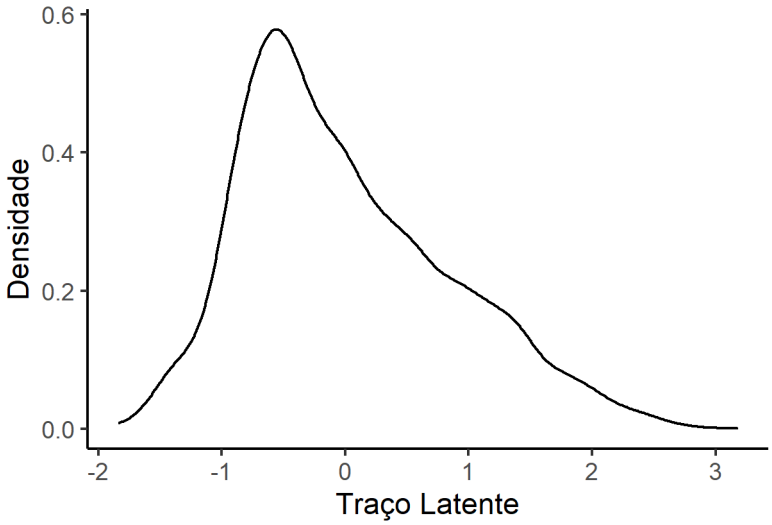
Nessa etapa, o método EAP foi utilizado nas estimativas das habilidades dos examinados. Esse método já está implementado na biblioteca MIRT (CHALMERS, 2012). A métrica utilizada na estimação das habilidades é *(0,1)*.

Uma amostra aleatória de 10 mil alunos foi utilizada. O processo de amostragem é o mesmo descrito anteriormente. Essa amostra é independente da amostra utilizada na estimação dos parâmetros dos itens.

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

Para tornar o projeto replicável, colocou-se uma semente igual a 3110 antes de gerar essa nova amostra de 10 mil. Considerando o caso em que os respondentes fizeram todas as questões, a Figura 19 apresenta a distribuição de notas desses examinados. E algumas estatísticas descritivas podem ser encontradas na Tabela 5.

Figura 19 - Distribuição dos traços latentes de 10 mil examinados



Fonte: Elaborado pelo autor.

Tabela 5 - Estatísticas Descritivas Traço Latente, prova completa

Mínimo	Mediana	Média	Máximo	Desvio Padrão	Média do Erro Padrão
-1.834	-0.158	0.011	3.178	0.852	0.50

Fonte: Elaborado pelo autor.

Para o caso do TAM, alguns problemas precisam ser ressaltados. A verdadeira prova feita como TAM não seria possível mensurar, pois tem-se apenas as respostas dos alunos em seu teste linear. Se o examinado tivesse feito o TAM verdadeiramente, ele teria um resultado diferente, como precisaria fazer menos questões, poderia dar mais atenção (e tempo) para as questões. Também poderia revisar melhor as questões dentro de um módulo. O formato da prova, por ser inédito para o examinado no ENEM, também poderia afetar seu desempenho.

Dessa forma, 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. Ou seja, considerou-se que a amostra de examinados responderam apenas as questões do TAM, com cada examinado respondendo os módulos condizentes a ele.

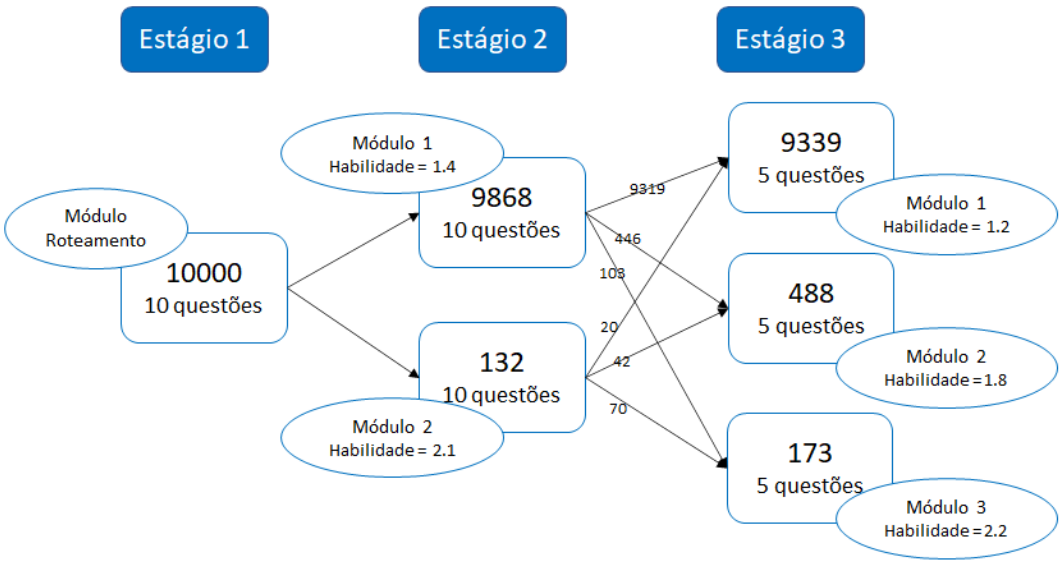
A realização de um TAM segue uma sequência de etapas. Primeiro, supôs-se que todos os examinados da amostra responderam o módulo de roteamento do Estágio 1. Em seguida, estimou-se a habilidade dos examinados, utilizando o método EAP e os parâmetros de itens já calculados. Após isso, fez-se um roteamento 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 encontrar qual módulo possui a maior informação para o traço latente estimado. Como mostrado, a Figura 15 tem a informação dos módulos de cada estágio em função do traço latente, então, após a estimação do traço latente no módulo de roteamento, encontra-se qual módulo tem maior valor para esse traço latente.

Uma vez no Estágio 2, estima-se novamente a habilidade dos examinados, considerando as questões 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 todas as questões que tal examinado teria feito nessa simulação de TAM.

A Figura 20 mostra quantos examinados dentro da amostra aleatória seguiram para cada módulo. A maioria dos examinados foram direcionados para os módulos mais fáceis.

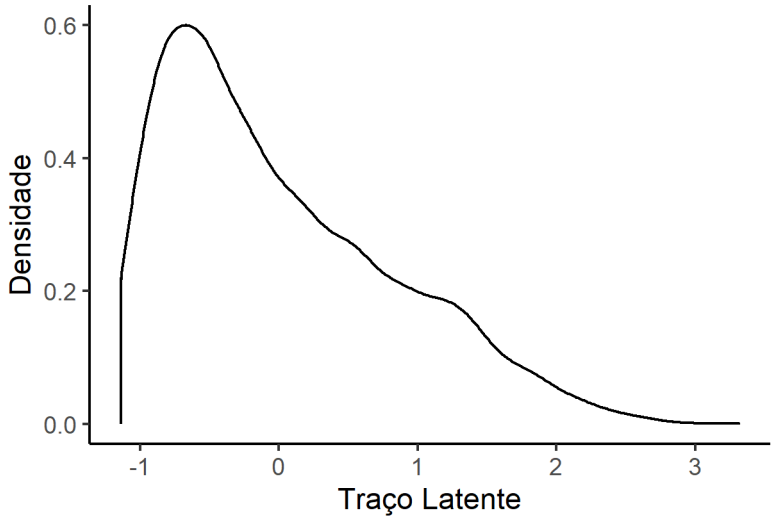
Figura 20 - Número de examinados que responderam cada módulo



Fonte: Elaborado pelo autor.

Após esse processo de roteamento e estimação, encontraram-se os traços latentes dos 10 mil examinados, simulando o TAM. A Figura 21 contém a distribuição de notas desses examinados. E algumas estatísticas descritivas estão na Tabela 6.

Figura 21 - Distribuição dos traços latentes de 10 mil examinados fazendo o TAM.



Fonte: Elaborado pelo autor.

Tabela 6 - Estatísticas Descritivas Traço Latente, prova TAM

Mínimo	Mediana	Média	Máximo	Desvio Padrão	Média do Erro Padrão
-1.14	-0.196	0.009	3.318	0.832	0.54

Fonte: Elaborado pelo autor.

Por fim, comparam-se os resultados dos examinados pelos dois métodos. Por uma análise visual dos gráficos (Figura 19 e Figura 21), percebeu-se que o lado direito de ambas as distribuições é muito semelhante, enquanto, o lado esquerdo (das pessoas com habilidades menores) tem uma diferença expressiva. Em ambas as figuras se notou que existe uma concentração de indivíduos com habilidades abaixo de zero, apesar de a prova ter informação (precisão) maior para os indivíduos com traço latente maior. A Figura 20, mostra que a maior parte dos indivíduos seguem para os módulos de menor habilidade. A prova estima melhor examinados com habilidades altas, apesar de a maioria expressiva não alcançar tais habilidades. Acredita-se que a prova deveria ter questões com dificuldades mais distribuídas.

Pelas Tabelas 5 e 6, o valor mínimo no caso do teste completo é de -1.84, enquanto no TAM é de -1.14. Essa diferença na estimação dos examinados com menor habilidade pode ser dada por alguns fatores. Por exemplo, a estimação com a prova completa considera 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. Outro exemplo, seria justificado pela falta de itens adequados a indivíduos com menor habilidade, dificultando a estimação mais precisa deles.

A média e a mediana de ambos ficaram em valores próximos. Apesar de a distribuição a esquerda estar desigual, a concentração de alunos com habilidades abaixo de zero ser bem expressiva, levou a média e a mediana serem próximas.

A média do Erro Padrão nos dois casos ficou bem próxima. Isto é, a estimação de cada traço latente dos 10 mil amostrados, 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.

Fez-se também uma análise quantitativa da diferença entre ambos os casos. Encontrou-se a correlação de Pearson entre as estimativas dos dois casos, a correlação de Spearman e a Raiz do Erro Quadrático Médio (Tabela 7). Também foi feito um gráfico de dispersão entre as estimativas (Figura 23) e um gráfico mostrando a diferença entre as duas estimativas (Figura 22). Para o Erro Quadrático Médio utilizou-se a seguinte fórmula:

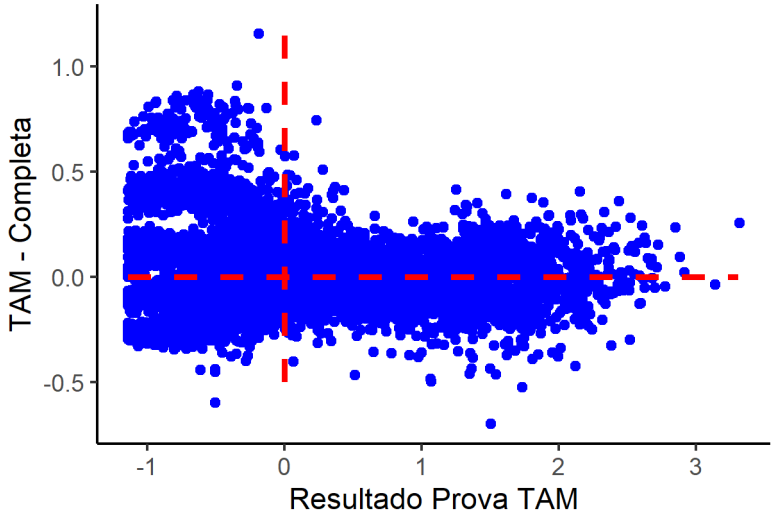
$$\frac{\sum_{i=1}^{10000}(Estimação\ TAM_i - \ Estimação\ Completa_i)^2}{10000} \quad (10)$$

Tabela 7 – Comparação TAM x Prova Completa

Correlação de Pearson	Correlação de Spearman	Raiz do Erro Quadrático Médio
0.976	0.967	0.184

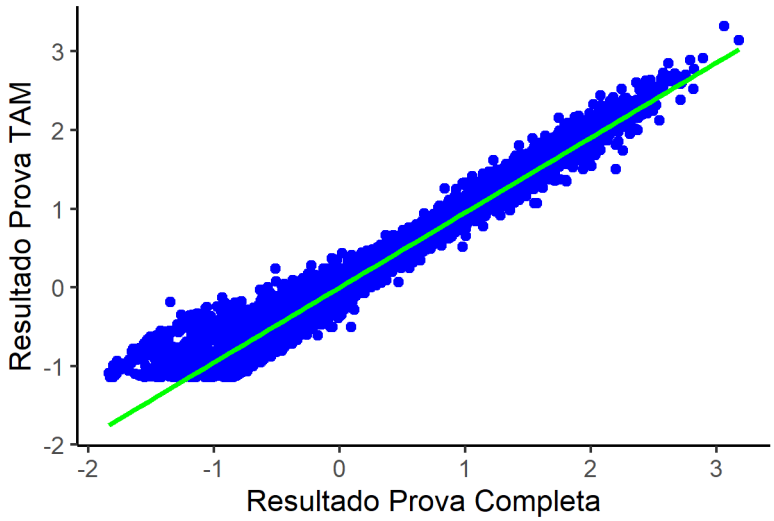
Fonte: Elaborado pelo autor.

Figura 22 - Diferença entre estimativas, prova TAM e prova completa



Fonte: Elaborado pelo autor.

Figura 23 - Gráfico de dispersão Prova Completa x Prova TAM.



Fonte: Elaborado pelo autor.

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%* das questões, conseguiram-se estimativas muito próximas. 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.

A Figura 22, explicita a ideia de que os valores maiores de estimação ficaram muito próximos, enquanto os valores menores têm uma dispersão maior nas diferenças. Na Figura 22 e na Figura 23 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.

4.3. Comparação com a Nota do ENEM

A nota do ENEM, disponibilizada pelos microdados, também é calculada utilizando o método EAP (ENEM, 2012). Essa nota, como foi citado, é calculada na métrica *(500,100)*, porém, isso não impede a comparação com a nota estimada em outra métrica.

Na Tabela 8, apresenta-se a correlação entre as notas atribuídas a amostra de 10 mil utilizando o TAM, com a nota do ENEM na prova de matemática na mesma amostra. Percebe-se que existe uma relação forte e linear entre tais notas, o que indica que a estimação de ambos está de acordo.

Tabela 8 - Comparação TAM x ENEM

Correlação de Pearson	Correlação de Spearman
0.982	0.973

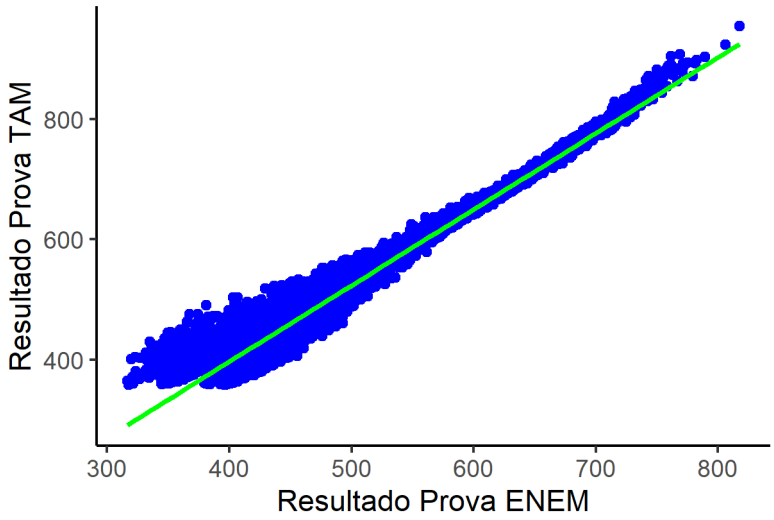
Fonte: Elaborado pelo autor.

Fez-se uma transformação linear da escala para colocar ambas as estimações na mesma métrica. Os valores estimados para o TAM foram transformados para a escala do ENEM.

$$F(\theta) = 500 + 100\theta \tag{11}$$

Após a transformação, foi possível fazer um gráfico de dispersão (Figura 24) entre as estimativas. Novamente, percebeu-se que elas seguem uma relação linear, e apenas com valores menores têm-se um espalhamento maior das notas.

Figura 24 - Prova ENEM x Prova TAM



Fonte: Elaborado pelo autor.

CAPÍTULO 5: CONCLUSÃO

5.1. Contribuições

Em 2020 ocorreu a primeira edição do Enem Digital. Essa possibilidade de implementar a prova digital criou oportunidades para aplicação de diferentes metodologias de testes. Uma dessas seriam os testes adaptativos multiestágio. Neste trabalho, avaliamos um teste adaptativo para a prova de Matemática do Enem de 2019.

Na exploração da prova de 2019, verificou-se uma falta de questões de habilidades distintas, a maioria das questões eram difíceis para a população que realizou a prova. Também foi possível encontrar questões inadequadas, as quais não agregam informação na habilidade dos indivíduos. O TAM proposto permitiu uma redução de 45 para 25 questões, sem prejudicar a estimação da nota do aluno. Pelo fato da prova ser considerada difícil, a estimativa de habilidades para examinados com habilidades menores foi prejudicada.

Sendo o TAM implementado e testado com os alunos do ensino médio, acredita-se que terá melhores possibilidades de estimar as habilidades de tais alunos. E com um banco de questões mais adequado e diversificado, as estimações se tornariam mais precisas no intervalo de habilidades dos alunos.

O TAM permite que o examinado faça menos questões. Ao manter o tempo da prova igual, o examinado então teria um melhor controle sobre o tempo, motivando-o a fazer as questões com maior foco. Com o TAM, também deixar-se-ia de ter questões inflando o teste. Questões que não sejam adequadas à habilidade do aluno não seriam apresentadas a tal aluno.

Uma vez que o ENEM se torne completamente digital, os testes adaptativos são uma alternativa a ser considerada para a realização da prova. Em que traria benefícios aos participantes, sem prejudicar as estimativas das notas dos alunos.

5.2. Trabalhos Futuros

Em trabalhos futuros, espera-se adicionar um estudo sobre as provas adaptadas, indivíduos que necessitem fazer a prova adaptada devem ser incluídos no mesmo formato do TAM, com as questões sendo condizentes (adequadas à realidade dos examinados).

Outro trabalho seria aplicar o TAM para todas as áreas do conhecimento, verificando a adequação das questões e, se é possível, realizar tal formato para as demais áreas. Por exemplo a prova de Linguagens, é uma prova mais específica, em que os alunos podem responder questões de inglês ou espanhol. E essas questões devem ser consideradas em um possível TAM.

Um outro estudo futuro, seria sobre o funcionamento da aplicação (aplicabilidade) do TAM, e consequentemente, sobre os resultados dos alunos quando aplicados a tal. Uma vez que o TAM realizado, como seria a resposta dos alunos a essa prova e a adequação dos indivíduos a esse modelo de prova.

Por último, outra possibilidade seria a exploração de um TAM quando o banco de questões é bem diverso, contendo questões de variadas dificuldades. Procurando descobrir se o TAM se adaptaria a essas questões, se os módulos ficariam muito distintos (com menos questões em comum), e se as estimações para indivíduos de habilidades menores seriam mais precisas.

REFERÊNCIAS

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. **Teoria da resposta ao item: conceitos e aplicações**. ABE, São Paulo, 2000.

ARÊA, L. Enem. **Divulgados os resultados finais do exame**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. 2021. Disponível em: < https://www.gov.br/inep/pt-br/assuntos/noticias/enem/divulgados-os-resultados-finais-do-exame>. Acesso em: 01 jun. 2021.

BAKER, F. **The Basics of Item Response Theory**. 2001. ERIC. Capítulo 6.

CHALMERS, R. P. **mirt: A multidimensional item response theory package for the R environment**. Journal of Statistical Software, v. 48, n. 6, p. 1–29, 2012.

CHALMERS, R. P. **Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications**. 2016.

ENEM. **Guia do participante**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Brasília, 2012.

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

GRÉGOIRE, J., Laveault, D. **Introdução às Teorias dos Testes em Ciências Humanas**. Porto, Portugal: Porto. 2002.

INEP. **Edital Enem 2019**. Diário Oficial da União – Seção 3. 2019. Disponível em: <https://download.inep.gov.br/educacao_basica/enem/edital/2019/edital_enem_2019.pdf> Acesso em: 01 jun. 2021.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Exame Nacional do Ensino Médio (Enem).** Disponível em: <https://www.gov.br/inep/pt-br >. Acesso em: 01 jun. 2021.

JATOBÁ, V. Delgado, K. V. Farias, J. Freire, V. **Comparação de regras de seleção de itens em testes adaptativos computadorizados: um estudo de caso no Enem**. Simpósio Brasileiro de Informática na Educação-SBIE. 2018. v. 29, n. 1, p. 1453.

LAROS, J. A. **O uso da Teoria de Resposta ao Item em Avaliações Educacionais: Diretrizes para pesquisadores**. 2021. Avaliação Psicológica, páginas 421-435.

LE, L. **Item point-biserial discrimination**. 2012. Disponível em: <https://www.acer.org/files/Conquest-Notes-5-ItemPointBiserialDiscrimination.pdf>. Acesso em: 01 jun. 2021.

LOPES, A. **Enem Digital**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. 2019. Disponível em: <http://portal.mec.gov.br/images/stories/noticias/2019/junho/03.07.2019_Coletiva-lanamento-Enem-Digital.pdf>. Acesso em: 01 jun. 2021.

LY, A., Marsman, M., Verhagen, J., Grasman, R., Wagenmakers, E. **A Tutorial on Fisher Information**. University of Amsterdam. Department of Psychological Methods. Vol. X. 2017. Disponível em: <https://arxiv.org/pdf/1705.01064.pdf> Acesso em: 01 jun. 2021.

MEC. Ministério da Educação. **Atendimento especializado deve ser requerido até 17 de maio**. Brasília. Disponível em: <http://portal.mec.gov.br/component/tags/tag/atendimento-especializado>. Acesso em: 01 jun. 2021.

PISA. **OECD's Programme for International Student Assessment**. 2021. Disponível em: <https://www.oecd.org/pisa/aboutpisa/>. Acessado em: 01 jun 2021

PITON-GONÇALVES, J., Almeida, A. **Análise da dificuldade e da discriminação de itens de Matemática do ENEM**. REMAT, Bento Gonçalves, RS, Brasil, v. 4, n. 2, p. 38-53. 2018.

PITON-GONÇALVES, J. **Testes Adaptativos para o Enade: uma aplicação metodológica**. 2020. Revista Meta: Avaliação.

RICARTE, T. A. M. **Multistage adaptative Testing based on logistic positive**. Universidade de São Paulo. 2016.

RICARTE, T. A. M., CURI, M., DAVIER, A. V. **Modeling Accidental Mistakes in Multistage Testing: A Simulation Study**. Springer Proceedings in Mathematics & Statistics, v. 233, p. 55-65, 2018.

SARTES, L. M. A., Souza-Formigoni, M. L. O. **Avanços na psicometria: da Teoria Clássica dos Testes à Teoria de Resposta ao Item**. 2013.

SILVA, V. R., CURI, M. **Academic English proficiency assessment using a computerized adaptive test**. TEMA**:** Tendências em Matemática Aplicada e Computacional, v. online, p. 000, 2019.

SOUZA, W. F**. Testes Adaptativos Computadorizados Aplicados às Provas do ENEM**. 2019. Universidade Federal de Ouro Preto. Disponível em: <https://www.monografias.ufop.br/bitstream/35400000/1804/1/MONOGRAFIA_TestesAdaptativosComputadorizados.pdf> Acessado em: 01 jun. 2021.

UFC. Universidade Federal do Ceará. **Notas de Corte – Processo Seletivo Sisu 1º/2019**. 2019. Disponível em: <https://sisu.ufc.br/wp-content/uploads/2019/10/sisu-2019-notas-de-corte.pdf>. Acesso em: 01 jun. 2021.

VAN DER LINDEN, W. J., Glas, C. A. W. **Elements of Adaptative Testing**. Springer. 2010. Capítulo 18.

VAN DER LINDEN, W. J. **Handbook of Item Response Theory, volume 1**. CRC Press. 2015 Capítulo 2.

WENDLER, C., Bridgeman, B. **The Research Foundation for the GRE revised General Teste: A Compendium of Studies**. ETS. 2014 Capítulo 1 e Capítulo 3.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York. 2016.

WILLSE, J. T. Shu, Z. **Package CTT**. 2008.

YAMAMOTO, K., Shin, H. J., Khorramdel, L. **Introduction of multistage adaptive testing design in PISA 2018**. 2019. OECD Education Working Paper Número 209.

YAN, D., Lewis, C., von Davier, A. **Computerized multistage testing: Theory and applications**. Capítulo 1. CRC. 2014.

APÊNDICE A – Tabela Complementar - itens

Item	Discriminação	Dificuldade	Acerto ao acaso	ID	IDS
1	2.47	1.26	0.11	0.25	0.4
2	4.2	2.51	0.11	0.12	0.05
3	1.22	2.44	0.11	0.18	0.19
4	3.49	2.24	0.13	0.15	0.14
5	3.13	1.44	0.15	0.24	0.36
6	2.8	1.89	0.17	0.22	0.23
7	3.65	1.73	0.2	0.25	0.24
8	-0.23	-11.74	0.11	0.17	-0.02
9	-2.37	-2	0.14	0.19	-0.01
10	-0.97	-3.41	0.11	0.16	-0.03
11	1.59	1.29	0.15	0.31	0.3
12	2.84	1.7	0.21	0.27	0.26
13	2.54	0.96	0.14	0.33	0.41
14	2.65	1.63	0.26	0.32	0.24
15	1.37	2.26	0.33	0.39	0.12
16	-0.19	-17	0.17	0.2	0
17	4.79	1.83	0.2	0.23	0.19
18	2.37	0.57	0.08	0.38	0.45
19	2	0.54	0.23	0.49	0.34
20	3.78	1.85	0.1	0.14	0.28
21	2.67	1.7	0.1	0.17	0.33

22	1.77	1.93	0.19	0.26	0.21
23	1.44	1.94	0.1	0.2	0.26
24	0.13	11.31	0.02	0.21	0.03
25	1.68	0.92	0.06	0.3	0.41
26	4.78	2.17	0.29	0.31	0.08
27	1.08	1.94	0.08	0.22	0.25
28	1.07	0.16	0.28	0.62	0.22
29	5.15	2.54	0.13	0.13	0.05
30	1.38	0.99	0.1	0.34	0.33
31	1.81	1.93	0.21	0.28	0.2
32	2.22	1.39	0.05	0.18	0.43
33	2.76	1.16	0.14	0.28	0.41
34	3.32	2.59	0.16	0.16	0.08
35	2.67	2.51	0.18	0.19	0.1
36	1.43	0.45	0.27	0.55	0.28
37	1.42	0.1	0.15	0.55	0.31
38	3.24	2.7	0.15	0.15	0.06
39	2.24	1.32	0.15	0.28	0.35
40	3.06	2.06	0.21	0.24	0.15
41	2.35	1.85	0.18	0.24	0.24
42	0.09	20.33	0.03	0.16	0
43	2.25	1.38	0.11	0.24	0.37
44	6.07	2.16	0.21	0.23	0.11
45	4.05	2.33	0.17	0.18	0.09

APÊNDICE B – Normalização da base

Item	Prova Azul	Prova Amarela	Prova Rosa	Prova Cinza	Código Item
1	160	165	171	155	8368
2	168	173	179	163	8401
3	172	138	147	167	8442
4	137	140	143	152	9779
5	166	171	177	161	10360
6	161	166	172	156	13303
7	162	167	173	157	14324
8	175	178	138	170	17642
9	163	168	174	158	18002
10	145	149	153	141	23065
11	151	155	159	147	28434
12	148	152	156	144	30436
13	153	157	161	149	30570
14	147	151	155	143	30740
15	167	172	178	162	31184
16	156	160	164	174	39708
17	164	169	175	159	42803
18	143	147	151	139	54448
19	140	144	148	136	63187
20	142	146	150	138	63808
21	174	177	169	179	78445

22	155	159	163	173	83754
23	171	137	146	166	83792
24	177	180	140	172	83955
25	180	143	137	169	83994
26	154	158	162	150	84374
27	150	154	158	146	84409
28	144	148	152	140	86432
29	173	176	168	178	89634
30	179	142	136	168	111518
31	178	175	141	180	111593
32	138	141	144	153	111608
33	139	163	167	177	111696
34	152	156	160	148	117607
35	141	145	149	137	117612
36	136	139	142	151	117635
37	149	153	157	145	117725
38	159	164	170	154	117726
39	165	170	176	160	117743
40	176	179	139	171	117767
41	170	136	145	165	117783
42	146	150	154	142	117798
43	169	174	180	164	117818
44	158	162	166	176	117845
45	157	161	165	175	117950