

FUNDAÇÃO GETULIO VARGAS  
ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp  
CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA

SIMULAÇÕES EM TEORIA DE RESPOSTA AO ITEM E TESTES  
ADAPTATIVOS COMPUTADORIZADOS

FELLIPE LOPES LIMA LEITE

Rio de Janeiro  
2018

**FUNDAÇÃO GETULIO VARGAS**  
**ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp**  
**CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA**

**SIMULAÇÕES EM TEORIA DE RESPOSTA AO ITEM E TESTES  
ADAPTATIVOS COMPUTADORIZADOS**

”Declaro ser o único autor da presente monografia, requisito parcial para a obtenção do grau de Bacharel em Matemática Aplicada, e ressalto que não recorri a qualquer forma de colaboração ou auxílio de terceiros para realizá-lo a não ser nos casos e para os fins autorizados pelo professor orientador”.

---

**Fellipe Lopes Lima Leite**

**Orientador: Paulo Cezar P. Carvalho**

**Rio de Janeiro**  
**2018**

**FELLIPE LOPES LIMA LEITE**

**SIMULAÇÕES EM TEORIA DE RESPOSTA AO ITEM E TESTES  
ADAPTATIVOS COMPUTADORIZADOS**

“Projeto de Monografia apresentado à Escola de Matemática Aplicada – FGV/EMAp  
como requisito parcial para continuidade ao trabalho de monografia.”

Aprovado em \_\_\_\_ de \_\_\_\_ de \_\_\_\_.

Grau atribuído ao Projeto de Monografia: \_\_\_\_.

---

**Professor Orientador: Paulo Cezar P. Carvalho**  
**Escola de Matemática Aplicada - FGV/EMAp**  
**Fundação Getulio Vargas**

# Sumário

<b>1</b>	<b>Resumo</b>	<b>4</b>
<b>2</b>	<b>Introdução Teórica</b>	<b>5</b>
2.1	Teoria Clássica da Medida . . . . .	5
2.2	Teoria de Resposta ao Item . . . . .	5
2.2.1	Unidimensionalidade . . . . .	6
2.2.2	Independência Local . . . . .	6
2.3	Curva Característica do Item . . . . .	6
2.4	Estimação da Habilidade . . . . .	7
2.4.1	Máxima Verossimilhança . . . . .	8
2.4.2	Método de Newton-Raphson . . . . .	9
2.5	Informação do Item . . . . .	10
2.6	Informação e Erro Padrão do Teste . . . . .	12
<b>3</b>	<b>Teste Adaptativo Computadorizado</b>	<b>14</b>
<b>4</b>	<b>Desenvolvimento</b>	<b>15</b>
4.1	Montagem da Base de Estudantes . . . . .	15
4.2	Montagem da Base de Itens . . . . .	15
4.3	Estimação das Habilidades . . . . .	15
4.4	Erro entre Habilidade Real e Estimada . . . . .	16
4.5	Comparação com Notas Brutas . . . . .	17
4.6	Erro entre Habilidade Real e Estimada . . . . .	18
<b>5</b>	<b>Conclusão</b>	<b>19</b>

# 1 Resumo

Estimar a proficiência de um aluno baseado em uma prova não é um problema simples. O Exame Nacional do Ensino Médio (ENEM), por exemplo, utiliza a Teoria de Resposta ao Item (TRI) para estimar os níveis de habilidade dos alunos que realizam o exame para ingressar na faculdade. Uma das grandes vantagens de usar a TRI é a possibilidade de comparar graus obtidos em testes devidamente calibrados. O presente trabalho objetiva simular uma avaliação, donde são conhecidos os parâmetros dos itens. Os acertos e erros de cada aluno que fez o teste são computados probabilisticamente e guardados. Os itens foram modelados com o modelo logístico de três parâmetros e a proficiência de cada aluno será estimada pelo estimador de máxima verossimilhança.

## 2 Introdução Teórica

### 2.1 Teoria Clássica da Medida

A teoria clássica possui alguns problemas na hora de avaliar a habilidade de um indivíduo. O principal deles é a interdependência entre teste e aluno que faz o teste.

A dependência do teste aparece quando tentamos estimar a habilidade de um indivíduo. Por exemplo, coloque este indivíduo para fazer um teste difícil e outro fácil. No primeiro a nota tenderá a ser baixa, levando a concluir que o aluno tem a habilidade em questão pouco desenvolvida. Já no segundo, a nota tenderá a ser alta, causando a impressão de que o aluno tem a habilidade melhor desenvolvida. Mas a habilidade do aluno em si é uma só, não deveria haver duas medidas distintas. Por isso, na teoria clássica, comparar testes diferentes é uma questão complexa.

A dependência do grupo a ser avaliado surge quando queremos calibrar o teste. Se um dado teste, cujos itens ainda não foram calibrados, for feito por alunos de maior habilidade, a tendência é o índice de acertos de cada item ser alta, o que geraria um conjunto de parâmetros para aqueles itens. Se o mesmo teste não calibrado for feito por alunos menos habilidosos, o índice de acertos de cada item tende a ser menor, o que geraria outro conjunto de parâmetros para os mesmos itens.

A TRI surge como um conjunto de modelos utilizados para estimar habilidades e calibrar itens sem o problema da interdependência entre teste e aluno.

### 2.2 Teoria de Resposta ao Item

Segundo [1], a teoria é montada sobre dois princípios básicos: a performance de uma pessoa no item pode ser prevista por um conjunto de fatores - chamados de habilidade, e a relação entre a performance no item e o nível de habilidade pode ser descrito como uma função monótona crescente chamada função característica do item.

Em contraposição à teoria clássica, os modelos utilizados na TRI fazem com que os parâmetros dos itens sejam invariantes ao grupo de indivíduos realizando o teste assim como a habilidade estimada seja invariante ao teste aplicado, a menos de algum erro de medida. Isso dá um grande poder de comparabilidade aos testes montados utilizando-se desta teoria. Basta que os testes estejam calibrados na mesma régua

para viabilizar a comparação entre as estimativas das habilidades.

Para que os modelos funcionem e possam ser desenvolvidos, dois pressupostos são tomados: unidimensionalidade e independência local.

### 2.2.1 Unidimensionalidade

Apenas uma habilidade está sendo medida pelo conjunto de itens do teste. Sabemos que isso é infactível em termos de mundo real, mas basta que haja uma habilidade dominante. Numa prova de matemática, por exemplo, além da habilidade matemática, há a habilidade de interpretação da questão também. O teste é montado de modo que a habilidade matemática seja dominante.

Há também modelos que consideram multidimensionalidade, mas não serão abordados neste trabalho.

### 2.2.2 Independência Local

Uma vez fixada a habilidade de um indivíduo, acertar ou não cada item do teste são eventos independentes. Isto é, a habilidade fixada - já que tratamos de unidimensionalidade - é o único fator que influencia a probabilidade de acertar ou não uma questão. O conjunto de questões em si torna-se independente. Desta forma, poderemos usar que:

$$P(U_1, U_2, \dots, U_n | \theta) = P(U_1 | \theta) * P(U_2 | \theta) * \dots * P(U_n | \theta) \quad (1)$$

## 2.3 Curva Característica do Item

A função característica do item, por ser monotônica crescente, diz que quanto maior a habilidade de um indivíduo, maior a probabilidade do mesmo acertar o item. Para a simulação feita aqui, nos basearemos no modelo logístico de três parâmetros conforme apresentado em [2]. Assim, a probabilidade de um item ser acertado dada um certo nível de habilidade é dada pela expressão a seguir:

$$P_i(\theta) = c_i + (1 - c_i) * \frac{1}{1 + e^{-Da_i(\theta - b_i)}} \quad (2)$$

Neste modelo,  $a_i$  é o parâmetro discriminante do item  $i$ . Isto é, o quanto este

item consegue distinguir um aluno com maior nível de habilidade de um com menor nível de habilidade. Pode ser entendido também como a tangente da curva no ponto de inflexão, onde a probabilidade de acerto é  $\frac{1+c_i}{2}$ .  $b_i$  é o parâmetro de dificuldade do item  $i$ . Quanto maior  $b_i$ , maior o nível de habilidade necessário para se ter  $\frac{1+c_i}{2}$  de chance de acerto.  $c_i$  é o parâmetro de acerto ao acaso. Não necessariamente ele valerá 0.2 para um item com 5 alternativas ou 0.25 para um item com 4 alternativas, isso depende de como as opções foram montadas, podendo haver um viés de acerto ou de erro.

A figura 1 mostra como fica a curva característica do item, considerando aplicado o modelo logístico. Para entender como os parâmetros influenciam no formato da curva, observe que a curva verde tem o maior valor de  $b$ , pois está mais deslocada para a direita, possui  $c = 0$ , que é a probabilidade de acerto supondo um nível de habilidade extremamente baixo ( $-\infty$ ) e possui o menor valor de  $a$ , o que faz com que ela fique menos inclinada. A curva azul tem  $b$  intermediário,  $c = 0.25$  e o maior valor de  $a$ , o que faz com que ela fique mais inclinada. Por fim, a curva laranja possui o menor valor de  $b$ , valor intermediário pro  $a$  e  $c = 0.25$ .

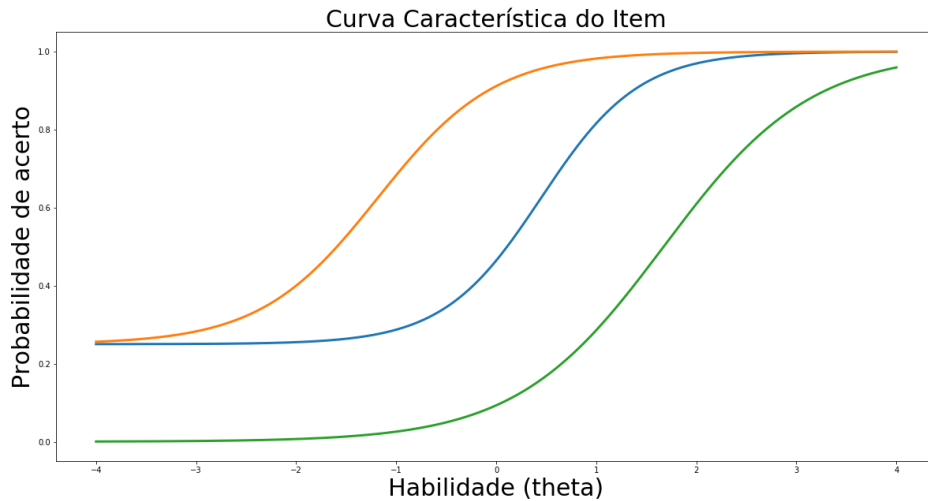


Figure 1: Curva característica do item

## 2.4 Estimação da Habilidade

Como pressuposto da simulação, consideramos os parâmetros dos itens conhecidos. Além disso, as respostas de um certo aluno aos itens do exame são variáveis



aleatórias binárias (1 sendo acerto e 0 erro), dadas pelo vetor  $U = (U_1, U_2, \dots, U_n)$ . Pela suposição de independência local, podemos considerar que a probabilidade conjunta do dado conjunto de respostas ocorrer é dada por:

$$P(U_1, U_2, \dots, U_n | \theta) = P(U_1 | \theta) * P(U_2 | \theta) * \dots * P(U_n | \theta) \quad (3)$$

Podendo ser reescrito como:

$$P(U_1, U_2, \dots, U_n | \theta) = \prod_{j=1}^n P(U_j | \theta) \quad (4)$$

Como  $U_j$  assume valores binários, considerando também  $P_j = P(U_j | \theta)$  e  $Q_j = 1 - P(U_j | \theta)$ , obtemos a expressão:

$$P(U_1, U_2, \dots, U_n | \theta) = \prod_{j=1}^n P_j^{U_j} * Q_j^{1-U_j} \quad (5)$$

#### 2.4.1 Máxima Verossimilhança

Conforme apresentado em [3], o método de estimação por máxima verossimilhança assume que o valor mais razoável para um estimador  $\theta$  é aquele que maximiza a probabilidade da amostra observada ocorrer. A função de verossimilhança é, portanto, a probabilidade de um evento ocorrer dado um valor do parâmetro  $\theta$ . Muitas vezes é utilizado o log da função de verossimilhança por simplificar as contas. Neste método, casos extremos como errar todos os itens ou acertar todos os itens causam problema no estimador, pois o mesmo iria para  $-\infty$  ou  $+\infty$ , respectivamente.

Uma vez que conhecemos o padrão de respostas  $U_j = u_j$  para cada questão  $j$ , nossa expressão de probabilidade passa a representar a função de verossimilhança, denotada por:

$$L(u_1, u_2, \dots, u_n | \theta) = \prod_{j=1}^n P_j^{u_j} * Q_j^{1-u_j} \quad (6)$$

Por se tratar de um produtório, aplicamos log em ambos os lados da igualdade para obter a expressão final a ser maximizada:

$$\ln L(u | \theta) = \sum_{j=1}^n [u_j * \ln P_j + (1 - u_j) * \ln(1 - P_j)] \quad (7)$$

Alguns exemplos de gráfico para a função log-verossimilhança são apresentados na figura 2. Perceba que o máximo é bem caracterizado, mas pode ser difícil de encontrar computacionalmente dependendo dos parâmetros do item, como é o caso do gráfico verde.

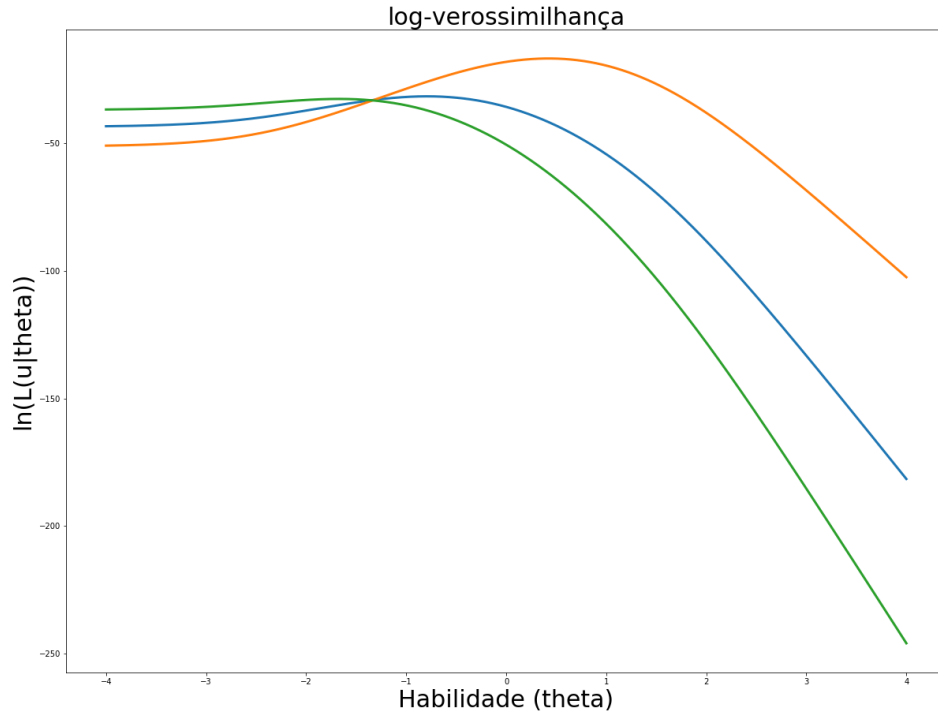


Figure 2: Função Log-verossimilhança

Para maximizar a função da equação 7, buscaremos a raiz da primeira derivada, cuja expressão é:

$$\frac{d \ln L}{d\theta} = \sum_{j=1}^n \left[ u_j * \frac{P'_j}{P_j} - (1 - u_j) * \frac{P'_j}{1 - P_j} \right] \quad (8)$$

Esta função chamaremos de  $f(x)$  para o desenvolvimento seguinte.

#### 2.4.2 Método de Newton-Raphson

O método é utilizado para encontrar raízes de uma função  $f(x)$  derivável, podendo ser estendido para sistemas de equações. Neste trabalho, foi utilizada sua versão mais simples, segundo a qual é encontrada uma linearização  $l(x)$  da função  $f$

num ponto inicial  $x_0$ . A raiz dessa linearização é utilizada como novo ponto inicial, e o processo se repete até que a alteração sofrida de uma aproximação para uma nova aproximação esteja abaixo de uma certa tolerância.

Equacionando o método. A primeira linearização é dada por:

$$l(x) = f'(x) * (x - x_0) + f(x_0) \quad (9)$$

Donde podemos encontrar sua raiz  $x_1$ :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \quad (10)$$

Generalizando, obtemos a seguinte relação de recorrência:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (11)$$

A qual é repetida até que  $|x_{n+1} - x_n|$  seja menor que uma tolerância  $\epsilon$ .

Considerando a função de verossimilhança em XXX como  $f(x)$ , temos que sua derivada é dada por:

$$f'(x) = \sum_{j=1}^n [u_j * \frac{P_j'' * P_j - P_j'^2}{P_j^2} - (1 - u_j) * \frac{P_j'' * (1 - P_j) + P_j'^2}{(1 - P_j)^2}] \quad (12)$$

O pseudocódigo utilizado para implementação do algoritmo foi obtido de [4], conforme apresentado a seguir:

## 2.5 Informação do Item

A função de informação do item é utilizada para descrever itens, fazer seleção de itens a serem colocados em um teste e é utilizada também para comparar testes. A informação que um item gera no nível de habilidade  $\theta$  é dado pela expressão:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta) * Q_i(\theta)} \quad (13)$$

A informação gerada num nível  $\theta$  de habilidade depende unicamente dos parâmetros dos itens. Ela será maior se  $b$  estiver mais próximo de  $\theta$ , ou seja, se a dificuldade do item estiver próxima do nível de habilidade do aluno; quanto maior o valor de

---

**Algorithm 1** Método de Newton-Raphson

---

```
1: procedure NEWTON( $f, f', x, nmax, \epsilon, \delta$ )
2:   integer  $n, nmax$ ; real  $x, fx, fp, \epsilon, \delta$ 
3:   external function  $f, f'$ 
4:    $fx \leftarrow f(x)$ 
5:   output  $0, x, fx$ 
6:   for  $n = 1$  to  $nmax$  do
7:      $fp \leftarrow f'(x)$ 
8:     if  $\|fp\| < \delta$  then
9:       output "small derivative"
10:    return
11:     $d \leftarrow fx/fp$ 
12:     $x \leftarrow x - d$ 
13:     $fx \leftarrow f(x)$ 
14:    output  $n, x, fx$ 
15:    if  $\|d\| < \epsilon$  then
16:      output "convergence"
17:    return
18:
```

---

a - isto é, quanto mais discriminante for o item; e quanto mais  $c$  tende a zero - pois reduziria o acerto ao acaso. A figura 3 dá uma dimensão visual de como os parâmetros afetam a informação do item.

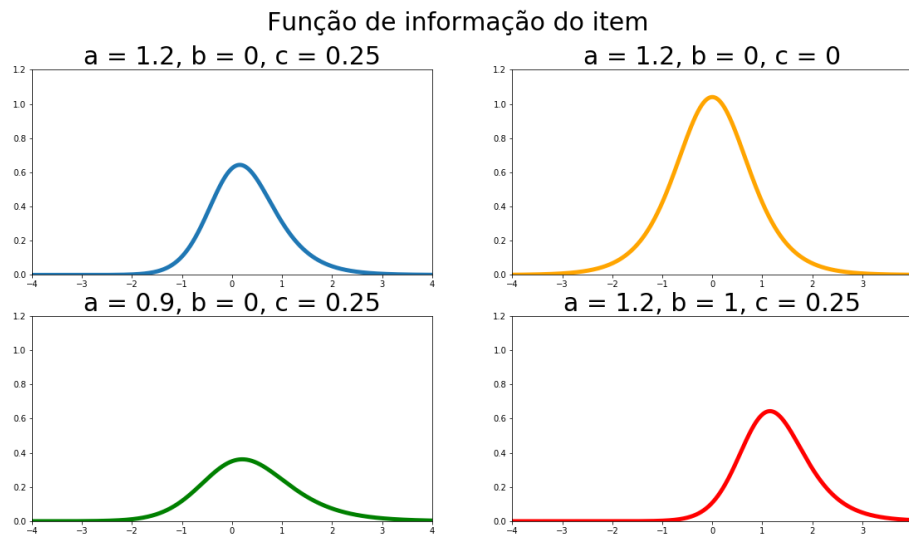


Figure 3: Exemplos de curva de informação do item

## 2.6 Informação e Erro Padrão do Teste

A informação do teste em si é a soma da informação de cada item que compõe esse teste. Ou seja:

$$I(\theta) = \sum_{i=1}^n I_i(\theta) \quad (14)$$

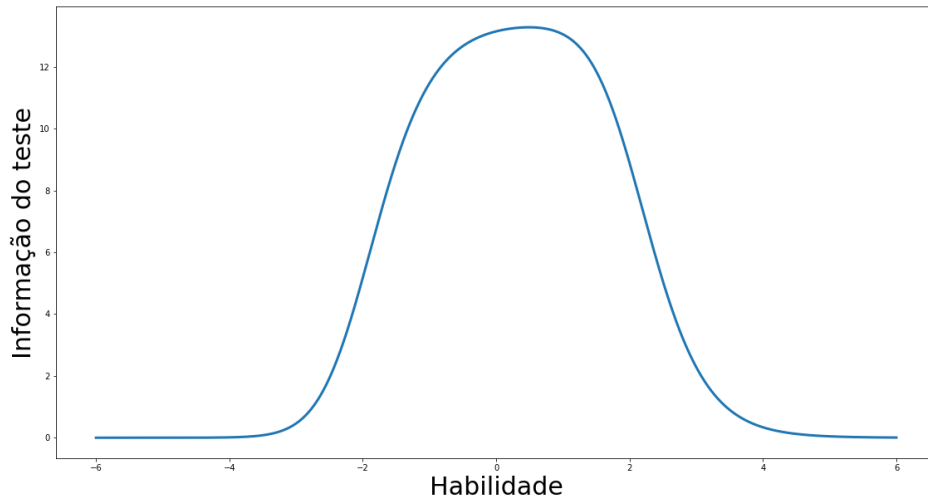


Figure 4: Curva de informação do teste simulado

Observe na figura 4 que o teste gera uma boa informação para valores de  $\theta$  entre -2 e 2, e passa a gerar informação menor para valores fora desse intervalo.

O erro padrão do teste, como apresentado por [1], é uma medida da precisão com a qual o nível de habilidade pode ser estimado naquele ponto.

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \quad (15)$$

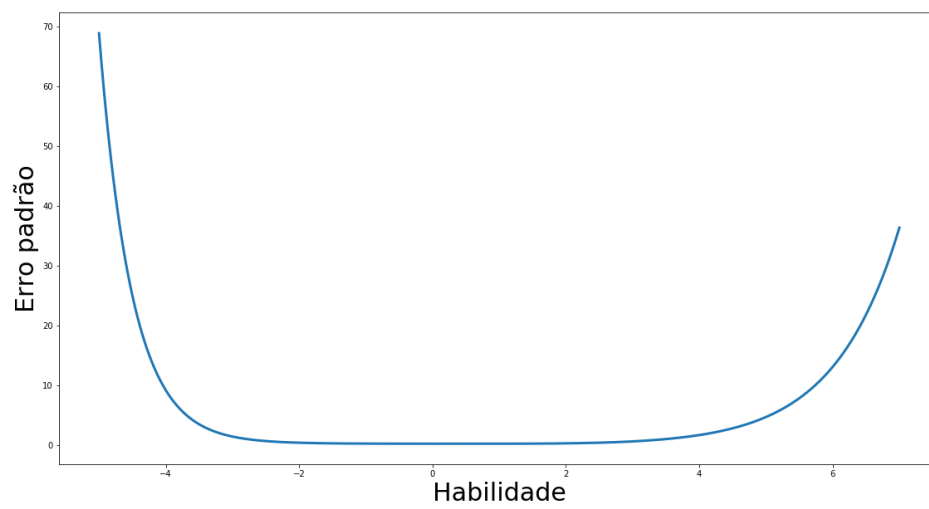


Figure 5: Desvio padrão do teste simulado

### 3 Teste Adaptativo Computadorizado

O Teste Adaptativo Computadorizado (TAC) surge como consequência da TRI. A ideia central é que um teste não consegue fornecer o mesmo grau de informação para todos os valores possíveis de habilidade. Deste modo, seria interessante que testes fossem personalizados de acordo com a capacidade inata do aluno. O algoritmo se baseia nos seguintes passos:

- O indivíduo responde a um conjunto pré selecionado de itens
- Geralmente são apenas dois ou três itens. Mas no caso de estimação por máxima verossimilhança, as questões iniciais são feitas até que haja um erro e um acerto pelo menos, para evitar os casos extremos.
- Baseado nas respostas até então, é estimada a habilidade do indivíduo.
- Calcula-se a informação gerada por cada item restante para a habilidade estimada.
- O próximo item da prova é o item que fornece maior informação.
- O processo de estimar habilidade e calcular a informação dos itens restantes é repetido até que haja um número máximo de questões no teste ou até que uma tolerância seja alcançada para o erro padrão do teste.

## 4 Desenvolvimento

### 4.1 Montagem da Base de Estudantes

Foram simulados 10.000 estudantes cujas proficiências foram amostradas a partir de uma distribuição normal padrão. A figura 6 apresenta o histograma das proficiências simuladas.

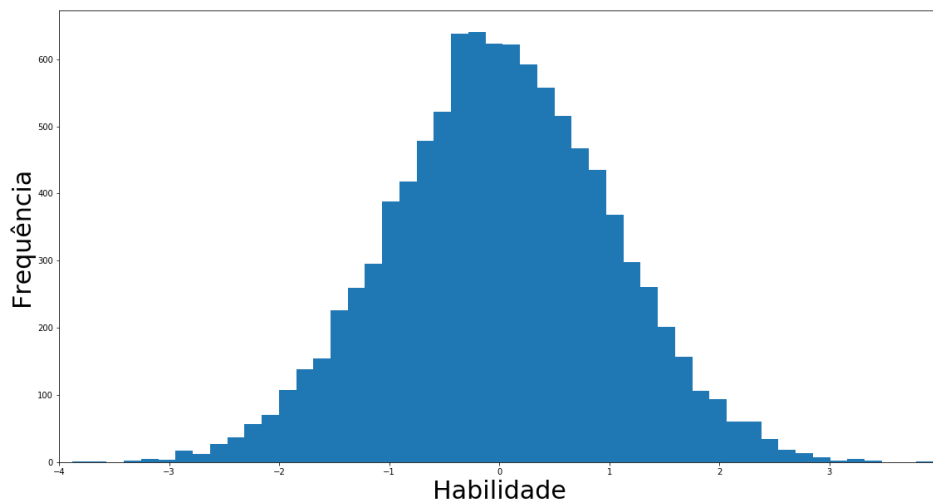


Figure 6: Distribuição simulada da proficiência dos 10.000 alunos

### 4.2 Montagem da Base de Itens

Para montagem da base de itens, considerando o modelo de três parâmetros, foram tomados  $a$  e  $c$  como constantes, respectivamente 1.2 e 0.25. Segundo [2], o parâmetro  $b$  usualmente assume valores entre  $-2$  e  $+2$ . Assim sendo, para formar os 45 itens do teste simulado, os valores de  $b$  foram tomados uniformemente no intervalo de  $-2$  a  $+2$  uniformemente.

### 4.3 Estimação das Habilidades

Os alunos foram colocados para fazer o teste, de modo que foi computado o acerto ou erro de cada aluno em cada questão. A probabilidade de acertar uma questão qualquer é dada pela expressão de  $P_i(\theta)$  na equação 2, utilizando o  $\theta$  real do aluno.



Uma vez obtido o vetor de respostas de cada aluno, foi aplicado o método de máxima verossimilhança, obtendo a distribuição da figura 7 para os níveis estimados de proficiência. Houve o caso de 109 alunos cujo cálculo da raiz da derivada deu erro, estes alunos foram desconsiderados na montagem do histograma, pois não influenciariam significativamente o resultado da simulação.

Entende-se que o erro dos 109 alunos foi causado por erro computacional. Ao buscar o máximo da função log-verossimilhança, alguns alunos apresentaram gráficos próximos ao da linha verde na figura 2 ou mais achatados. Isto fez com que a derivada tivesse valores muito próximos de zero que acabaram não passando no algoritmo de Newton-Raphson aplicado. Como o objetivo do trabalho não é a otimização dos algoritmos em si, foi decidido excluir estes 109 alunos da análise, já que correspondem a apenas 1% do espaço amostral.

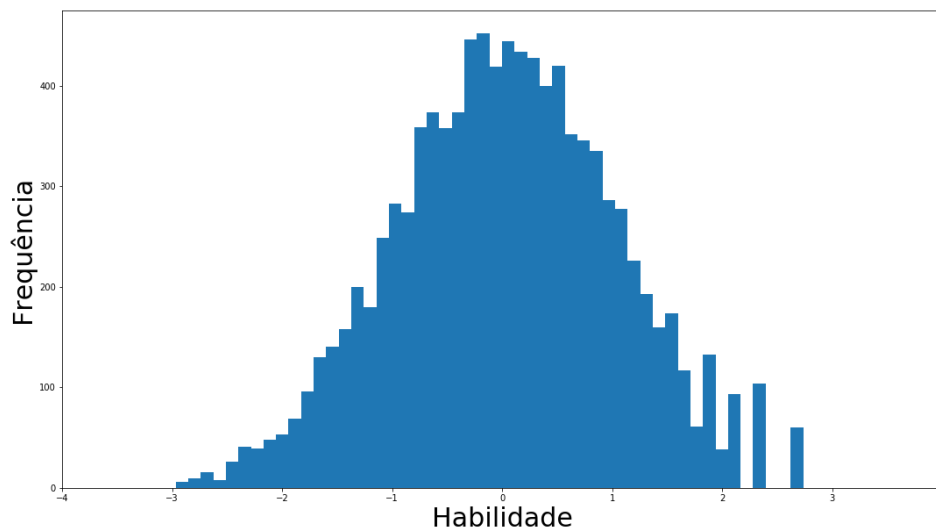


Figure 7: Distribuição estimada da proficiência dos alunos

#### 4.4 Erro entre Habilidade Real e Estimada

O erro padrão do teste foi calculado através da equação 15 para cada um dos alunos baseado na lista simulada de suas proficiências. A média encontrada para tais erros foi 0.3020. O erro em si é alto, em se tratando de uma distribuição normal padrão. Em grande parte, ele advém de erros de aproximação feitos durante a programação.

## 4.5 Comparação com Notas Brutas

Se calcularmos a nota bruta - obtida pelo número de acertos ao longo das 50 questões dividido por 5 - de cada aluno que realizou o teste, obteremos o gráfico apresentado na figura 8. É esperado que o gráfico não tenha o mesmo formato que o das habilidades simuladas, haja vista que o resultado que um aluno terá dentro de uma prova é uma variável aleatória composta pelo acerto ou erro de cada item. A probabilidade de acertar cada item, conforme visto na expressão 2, é função da habilidade. Para encontrar as notas estimadas, faremos a conversão para a escala de 0 a 10 segundo a equação 16 a seguir:

$$y = 10 * \frac{x - a}{b - a} \quad (16)$$

Onde  $a$  é o valor mínimo na lista de habilidades estimadas,  $b$  é o valor máximo na lista de habilidades estimadas,  $x$  é o valor de cada habilidade estimada e  $y$  é o valor correspondente na escala de 0 a 10 conforme desejamos. O gráfico correspondente será, claro, o mesmo que o da figura 7, salvo a mudança de escala no eixo x.

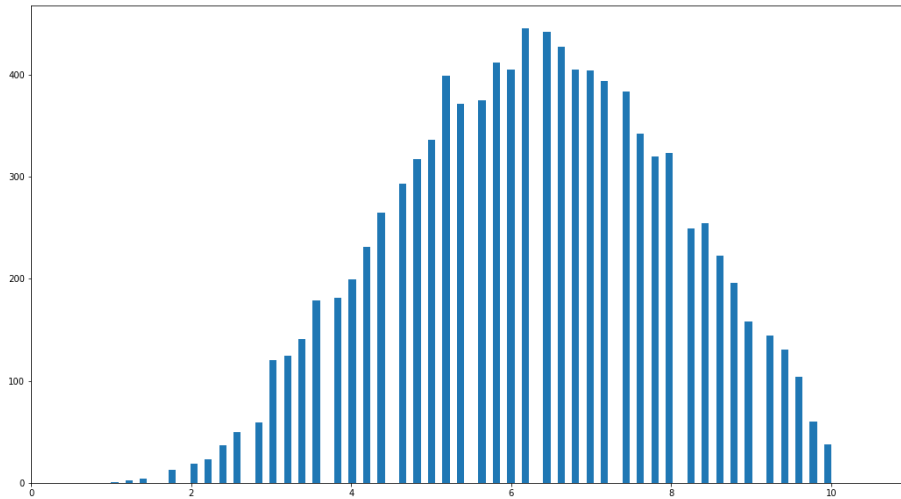


Figure 8: Distribuição das notas brutas obtidas pelos alunos

## 4.6 Erro entre Habilidade Real e Estimada

Para calcular o erro entre as notas estimadas e simuladas, faremos a variância entre a nota simulada e a nota estimada de cada aluno. Como os alunos são considerados independentes entre si, a variância total será a soma das variâncias de cada aluno. O resultado pro erro quadrático obtido seguindo este caminho foi de 1,14.

## 5 Conclusão

A simulação de um teste utilizando a Teoria de Resposta ao Item cumpriu seu objetivo: criar experiência prática de como métodos estatísticos podem ser utilizados para melhorar a informação obtida num exame.

A nível de trabalhos futuros, utilizar o método Bayesiano para estimação possivelmente resultará num erro menor, pois este método consegue lidar com os casos extremos de zerar ou gabaritar o exame. Além disso, seria interessante usar um teste real, com alunos reais fazendo, mas neste caso seria necessário estimar também os parâmetros dos itens, e não supô-los conhecidos. Espera-se, neste caso, que a distribuição das habilidades seja uma normal assimétrica, como é no ENEM, a exemplo da figura 9.

Além disso, implementar um modelo de teste adaptativo computadorizado e simular quantas questões um aluno precisaria fazer para ter o mesmo nível de erro que usando o modelo simples de TRI.

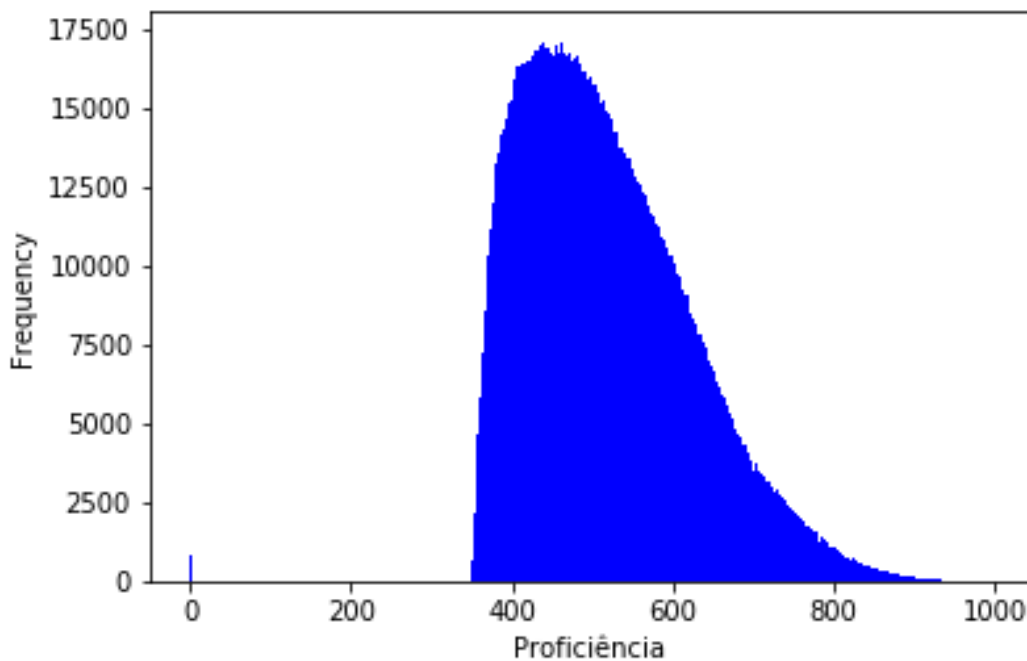


Figure 9: Distribuição de notas do ENEM 2017

## References

- [1] R. K. Hambleton, *Fundamentals of Item Response Theory (Measurements Methods for the Social Science)*. Sage Publications, 1991.
- [2] D. F. de Andrade, H. R. Tavares, and R. da Cunha Valle, *Teoria da Resposta ao Item: Conceitos e Aplicações*. SINAPE, 2000.
- [3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference and Prediction*. Springer, 2009.
- [4] W. Cheney and D. Kincaid, *Numerical Mathematics and Computing*. Thomson, 6 ed.