UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

VICTOR MIRANDA GONÇALVES JATOBÁ

# Uma abordagem personalizada no processo de seleção de itens em Testes Adaptativos Computadorizados

São Paulo
2019

VICTOR MIRANDA GONÇALVES JATOBÁ

# Uma abordagem personalizada no processo de seleção de itens em Testes Adaptativos Computadorizados

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 08 de outubro de 2018. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018,' de 13 de outubro de 2011.

Orientador: Profa. Dra. Karina Valdivia Delgado

São Paulo

2019

CATALOGAÇÃO-NA-PUBLICAÇÃO
(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB-8 4936

Jatobá, Victor Miranda Gonçalves
Uma abordagem personalizada no processo de seleção de itens em Testes Adaptativos Computadorizados / Victor Miranda Gonçalves Jatobá ; orientadora, Karina Valdivia Delgado. – 2019.
59 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, em 2018.
Versão corrigida

1. Metodologia e técnicas de computação. 2. Teoria de resposta ao item. 3. Exame Nacional do Ensino Médio. I. Delgado, Karina Valdivia, orient. II. Tìtulo.
CDD 22.ed.– 005.1

Dissertação de autoria de Victor Miranda Gonçalves Jatobá, sob o título **"Uma abordagem personalizada no processo de seleção de itens em Testes Adaptativos Computadorizados"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 08 de outubro de 2018 pela comissão julgadora constituída pelos doutores:

---

**Profa. Dra. Karina Valdivia Delgado**

Universidade de São Paulo

Presidente

---

**Profa. Dra. Ana Amélia Benedito Silva**

Universidade de São Paulo

---

**Prof. Dr. Francisco de Assis Zampirolli**

Universidade Federal do ABC

---

**Prof. Dr. Esteban Fernandez Tuesta**

Universidade de São Paulo

*Ao meu filho Miguel. Que o sirva de inspiração para o eterno processo de construção do saber.*

## Agradecimentos

São muitas, as pessoas da qual devo imenso agradecimento. Porém, certamente devo começar pelos meus pais. Eles me ensinaram todo o valor do que eu entendo por vida, me incentivando e apoiando em todas as minhas decisões. Em seguida, mas não menos importante, devo agradecer a minha esposa Priscila. Ela me apoiou incondicionalmente em todo o processo de mestrado e me deu força, todos os dias, para seguir em frente. Além disso segurou a maior barra dentro de casa, para que eu tivesse um pouco mais de tempo para concluir este trabalho. Também agradeço minha orientadora Prof.ª Karina, que com bastante atenção e dedicação, sempre atendeu aos meus chamados com sua disponibilidade e paciência ímpar.

Devo meus agradecimentos sinceros à minha sogra Denise, que cuidou do meu filho nas duras horas em que tive que abrir mão, para a concepção deste trabalho. Agradeço também ao Professor Jorge Farias que dedicou parte do seu tempo refletindo e sugerindo melhorias nos textos produzidos aqui e nos artigos publicados.

Agradeço a toda a minha família, sem exceções. Irmãos, tios, avós, cunhados, todos. Vocês também fizeram parte dessa trajetória. Muitas outras pessoas fizeram parte deste resultado de forma direta e indireta. Agradeço a todos as pessoas que tive o prazer de compartilhar alegrias e angustias durante o mestrado, criando vínculos de bastante amizade. O Rodrigo Siqueira, que me acolheu no CCSL-USP e me ajudou em vários aspectos diretos, como as escritas em inglês. Ao meu amigo William, que sempre esteve muito presente e solícito. A toda turma da EACH e do IME, pessoas marcantes que levarei pro resto da vida. Agradeço também a Leila e a Marcio por terem dado espaço e tempo no trabalho, para que eu me dedicasse ao depósito dessa dissertação.

Certamente vou pecar e esquecer de citar diretamente muitas pessoas importantes. De antemão, peço desculpas. Mas quero deixar registrado que essa pequena página não é suficiente para todo o meu sentimento de gratidão.

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

*"Poderoso para mim não é aquele que descobre ouro. Para mim poderoso é aquele que descobre as insignificâncias (do mundo e as nossas). Por essa pequena sentença me elogiaram de imbecil. Fiquei emocionado. Sou fraco para elogios."*

*(Manoel de Barros)*

## Resumo

JATOBÁ, Victor Miranda Gonçalves. **Uma abordagem personalizada no processo de seleção de itens em Testes Adaptativos Computadorizados**. 2019. 59 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

Testes Adaptativos Computadorizados (CAT) baseados na Teoria de Resposta ao Item permitem fazer testes mais precisos com um menor número de questões em relação à prova clássica feita a papel . Porém a construção de CAT envolve alguns questionamentos-chave, que quando feitos de forma adequada, podem melhorar ainda mais a precisão e a eficiência na estimativa das habilidades dos respondentes. Um dos principais questionamentos está na escolha da Regra de Seleção de Itens (ISR). O CAT clássico, faz uso, exclusivamente, de uma ISR. Entretanto, essas regras possuem vantagens, entre elas, a depender do nível de habilidade e do estágio em que o teste se encontra. Assim, o objetivo deste trabalho é reduzir o comprimento de provas dicotômicas – que consideram apenas se a resposta foi correta ou incorreta – que estão inseridas no ambiente de um CAT que faz uso, exclusivo, de apenas uma ISR sem perda significativa de precisão da estimativa das habilidades. Para tal, cria-se a abordagem denominada ALICAT que personaliza o processo de seleção de itens em CAT, considerando o uso de mais de uma ISR. Para aplicar essa abordagem é necessário primeiro analisar o desempenho de diferentes ISRs. Um estudo de caso na prova de *Matemática e suas tecnologias* do ENEM de 2012, indica que a regra de seleção de Kullback-Leibler com distribuição a *posteriori* ($KLP$) possui melhor desempenho na estimativa das habilidades dos respondentes em relação as regras: Informação de Fisher ($F$); Kullback-Leibler ($KL$); Informação Ponderada pela Máxima Verossimilhança ($MLWI$); e Informação ponderada a *posteriori* ($MPWI$). Resultados prévios da literatura mostram que CAT utilizando a regra $KLP$ conseguiu reduzir a prova do estudo de caso em $46,6\%$ em relação ao tamanho completo de 45 itens sem perda significativa na estimativa das habilidades. Neste trabalho, foi observado que as regras $F$ e a $MLWI$ tiveram melhor desempenho nos estágios inicias do CAT, para estimar respondentes com níveis de habilidades extremos negativos e positivos, respectivamente. Com a utilização dessas regras de seleção em conjunto, a abordagem ALICAT reduziu a mesma prova em $53,3\%$.

Palavras-chave: Testes Adaptativos Computadorizados, Teoria de Resposta ao Item, Exame Nacional do Ensino Médio, ENEM, Regra de Seleção de Item, Critério de Seleção de Item.

**Abstract**

JATOBÁ, Victor Miranda Gonçalves. **A personalized approach to the item selection process in Computerized Adaptive Testing**. 2019. 59 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

Computerized Adaptive Testing (CAT) based on Item Response Theory allows more accurate assessments with fewer questions than the classic paper test. Nonetheless, the CAT building involves some key questions that, when done properly, can further improve the accuracy and efficiency in estimating examinees' abilities. One of the main questions is in regard to choosing the Item Selection Rule (ISR). The classic CAT makes exclusive use of one ISR. However, these rules have differences depending on the examinees' ability level and on the CAT stage. Thus, the objective of this work is to reduce the dichotomous - which considers only correct and incorrect answers - test size which is inserted on a classic CAT without significant loss of accuracy in the estimation of the examinee's ability level. For this purpose, we create the ALICAT approach that personalizes the item selection process in a CAT considering the use of more than one ISR. To apply this approach, we first analyze the performance of different ISRs. The case study in *mathematics and its technologies* test of the ENEM 2012 shows that the Kullback-Leibler Information with a Posterior Distribution ($KLP$) has better performance in the examinees' ability estimation when compared with: Fisher Information ($F$); Kullback-Leibler Information ($KL$); Maximum Likelihood Weighted Information($MLWI$); and Maximum Posterior Weighted Information ($MPWI$) rules. Previous results in the literature show that CAT using $KLP$ was able to reduce this test size by 46.6% from the full size of 45 items with no significant loss of accuracy in estimating the examinees' ability level. In this work, we observe that the $F$ and the $MLWI$ rules performed better on early CAT stages to estimate examinees' proficiency level with extreme negative and positive values, respectively. With this information, we were able to reduce the same test by 53.3% using an approach that uses the best rules together.

Keywords: Computerized Adaptive Testing. CAT. Item Response Theory. IRT. Fisher Information. Item Selection Rule. Item Selection Method.

## Lista de figuras

## Lista de tabelas

**Lista de abreviaturas e siglas**

| | |
|---|---|
| BI | Banco de itens |
| BIAS | Viés |
| CAT | Teste adaptativo computadorizado |
| CCI | Curva característica do item |
| ENEM | Exame nacional do ensino médio |
| F | Informação de Fisher |
| IRT | Teoria de resposta ao item |
| KL | Informação de Kullback-Leibler |
| KLP | Informação de Kullback-Leibler com distribuição a |
| MIRT | Teoria de resposta ao item multidimensional |
| MLWI | Informação ponderada pela máxima verossimilhança |
| ML1 | Modelo logístico de 1 parâmetro |
| ML2 | Modelo logístico de 2 parâmetros |
| ML3 | Modelo logístico de 3 parâmetros |
| MPWI | Máxima informação ponderada a *posteriori* |
| P&P | Prova clássica, realizada por papel e lápis *posteriori* |
| RMSE | Raiz do erro quadrático médio |
| SE | Erro padrão |

## Lista de símbolos

| | |
|---|---|
| $\theta$ | Escore verdadeiro. Estimativa do nível de habilidade dos avaliados considerando a prova completa. |
| $\hat{\theta}$ | Escore parcial. Nível de habilidade dos avaliados estimados durante a execução do CAT. |
| $a$ | Taxa de discriminação da questão considerada no cálculo da IRT. |
| $b$ | Parâmetro de dificuldade da questão considerado no cálculo da IRT. |
| $c$ | Parâmetro de acerto casual considerado no cálculo da IRT. |
| $D$ | Constante numérica para que a função logística forneça resultados semelhantes aos da função normal. |
| $P(\theta)$ | Probabilidade do respondente, com habilidade $\theta$, acertar um dado item. |

**Sumário**

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

# 1 Introdução

A etapa de avaliação de estudantes sempre foi muito importante no processo de aprendizagem. Na computação, os sistemas de aprendizado geradores de testes ajudam os estudantes a identificarem se atingiram o nível adequado de conhecimento aprendido (GUZMÁN; CONEJO, 2004). Um exemplo desse tipo de sistema é o Teste Adaptativo Computadorizado (do inglês, *Computerized Adaptive Testing* – CAT).

CATs são testes administrados por computadores que, de forma eficiente, reduzem o número de itens (questões) mantendo um melhor diagnóstico do desempenho do respondente (KOVATCHEVA; NIKOLOV, 2009). No CAT clássico, inicialmente é selecionada uma questão e, a cada nova, é estimado o nível de habilidade do estudante. Caso o critério de parada não seja atendido, outra questão é selecionada.

Existem diversas formas de avaliar o nível de habilidade do respondente. Um modelo bastante utilizado recentemente é o da Teoria de Resposta ao Item (do inglês *Item Response Theory* – IRT) (LÓPEZ-CUADRADO et al., 2010; GUZMÁN; CONEJO, 2004; WONG et al., 2010). Esta teoria inclui um conjunto de modelos matemáticos que procuram estabelecer a probabilidade de um respondente qualquer acertar uma determinada questão, dadas as características do item e as habilidades do avaliado (ANDRADE; TAVARES; VALLE, 2000). Esse é o modelo adotado, por exemplo, pela prova objetiva do Exame Nacional do Ensino Médio (ENEM) para calcular o desempenho dos estudantes (BRASIL, 2013).

CATs baseados na IRT permitem fazer testes mais precisos (KOVATCHEVA; NIKOLOV, 2009) pois é possível identificar as áreas de carência do estudante e assim selecionar uma sequência de itens adaptada ao conhecimento do usuário (CHEN; LEE; CHEN, 2005).

## *1.1 Apresentação do problema*

CATs possuem diversas vantagens em relação à prova no formato papel e lápis (do inglês, *Paper and Pencil* – P&P) (SEGALL, 2004). Uma delas é a possibilidade de estimar com maior precisão as habilidades latentes dos respondentes considerando um menor número de itens (WAINER et al., 2000; BERNES-LEE, 2006).

Porém a construção de CATs envolve alguns questionamentos-chave, como a escolha dos critérios de inicialização e finalização do teste e da regra de seleção de itens (do inglês, *Item Selection Rule* – ISR) (WAINER et al., 2000). Além disso, a depender do cenário, CATs podem usar, junto com as ISRs, algum mecanismo para controlar a exposição de itens, ou algum tipo de balanceamento de conteúdo (CHEN; ANKENMAN, 2004).

A escolha mais adequada dos questionamentos-chave, pode melhorar a precisão e a eficiência na estimativa das habilidades dos respondentes, principalmente em relação às ISRs (CHEN; ANKENMANN; CHANG, 2000; CHEN; ANKENMAN, 2004; BARRADA et al., 2008; WANG; CHANG; HUEBNER, 2011).

A técnica de Informação de Fisher (do inglês, *Fisher Information* – F) é um exemplo de uma ISR. Segundo Chen, Ankenmann e Chang (2000), esta técnica é mais precisa para estimar o nível de habilidade de usuários medianos em estágios iniciais de CATs se comparada as técnicas de Máxima Informação ponderada a *posteriori* (do inglês, *Maximum Posterior Weighted Information* – $MPWI$) e a Informação de Kullback-Leibler com distribuição a *posteriori* (do inglês, *Kullback-Leibler Information with a Posterior Distribution* – $KLP$). Em contra partida, segundo Chen e Ankenman (2004) as estratégias $MPWI$ e $KLP$ possuem melhores resultados em relação à regra $F$ para níveis de habilidades extremos.

Também há trabalhos que indicam que as regras de Informação Ponderada pela Máxima Verossimilhança (do inglês, *Maximum Likelihood Weighted Information* – $MLWI$) e $MPWI$ apresentam um melhor desempenho geral que a regra $F$ (VEERKAMP; BERGER, 1997; LINDEN, 1998; LINDEN; PASHLEY, 2009). Esses resultados foram obtidos em aplicações que estão inseridas no ambiente de provas dicotômicas (que considera apenas se a resposta foi correta ou incorreta). Ademais, essas aplicações não fazem uso de controles de exposição de itens e nem de balanceamento de conteúdo.

O processo de seleção de itens considerado na construção do CAT clássico, envolve a escolha de apenas uma regra de seleção de itens (OLEA; PONSODA, 1996). Entretanto essas regras possuem vantagens e desvantagens a depender do nível de habilidade do respondente e do estágio - quantidade de itens que já foram selecionados - em que o teste se encontra (CHEN; ANKENMANN; CHANG, 2000; CHEN; ANKENMAN, 2004; VEERKAMP; BERGER, 1997; LINDEN, 1998; LINDEN; PASHLEY, 2009).

Assim, essa pesquisa se baseia na seguinte questão: é possível reduzir o comprimento de CATs que fazem uso exclusivo de apenas uma ISR sem perda significativa de precisão da estimativa das habilidades dos respondentes?

## *1.2 Objetivos*

O objetivo geral é reduzir o comprimento de provas dicotômicas, que estão inseridas no ambiente de um CAT que faz uso, exclusivo, de apenas uma ISR sem perda significativa de precisão da estimativa das habilidades dos respondentes. A proposta é analisar o desempenho de diferentes ISR e, com isso, personalizar o processo de seleção de itens em CAT, considerando o uso de mais de uma ISR (esta abordagem recebeu o nome de ALICAT, *personALIzed* CAT).

## *1.3 Organização*

Este texto é composto por mais cinco capítulos e as Referências. O Capítulo 2 contém a fundamentação teórica. Em seguida, é apresentado no Capítulo 3, a abordagem de personalização do processo de seleção de itens em CATs, ALICAT. Posteriormente, no Capítulo 4, é descrito o método de pesquisa. Por fim, são apresentados os resultados dos experimentos no Capítulo 5, e as considerações finais no Capítulo 6.

# 2 Fundamentos teóricos

A seguir são apresentados um levantamento do estado da arte e o histórico que deu base ao surgimento dos Testes Adaptativos Computadorizados. Além disso são expostos os principais conceitos que envolvem a configuração e validação de um CAT. Por fim são elucidados alguns detalhes do funcionamento do ENEM dando base e justificando o seu uso no estudo de caso deste trabalho.

## *2.1 Avaliação clássica do conhecimento*

Um dos componentes mais importantes em sistemas educacionais é a avaliação, a qual fornece as estruturas necessárias para determinar o nível de aprendizado do estudante (ÖZYURT et al., 2012). O uso de testes geralmente é uma forma comum de se avaliar o conhecimento e as habilidades dos aprendizes (RAJAMANI; KATHIRAVAN, 2013).

Os testes convencionais, suportados pela Teoria Clássica dos Testes (do inglês, *Classical Test Theory* – CTT) (NOVICK, 1966), avaliam o desempenho dos alunos por meio da quantidade de questões respondidas corretamente dentre um conjunto total de questões, obtendo o que se conhece por escore do teste (QUARESMA, 2014).

A CTT possui algumas limitações, principalmente pelo fato de ser uma avaliação orientada a teste (prova) (AL-A'ALI, 2007). Essa característica a torna incapaz de predizer o comportamento da resposta de um avaliado, ou de um grupo, a uma questão de forma isolada (JACOBSON, 1997).

Outro aspecto importante diz respeito a generalização da formação dos testes, que podem não considerar as individualidades dos avaliados (AL-A'ALI, 2007). Por exemplo, a relação entre acertar 75% de questões fáceis é diferente de acertar 75% de questões difíceis (SEGALL, 2009; PASQUALI; PRIMI, 2003). Além disso, as provas são elaboradas para avaliar, maximamente, os sujeitos de habilidades medianas, sendo, portanto, menos apropriadas e válidas para avaliar indivíduos com níveis mais extremos (PASQUALI; PRIMI, 2003). Um indivíduo com alto grau de habilidade pode considerar a prova entediante e consequentemente não ter um bom desempenho por considerar que algumas questões, ou a maioria delas, não são desafiadoras. No outro extremo, indivíduos com habilidades reduzidas podem sofrer frustrações e terem os mesmos resultados negativos na avaliação por acharem, algumas, ou a maioria das questões muito difíceis (SIE, 2014).

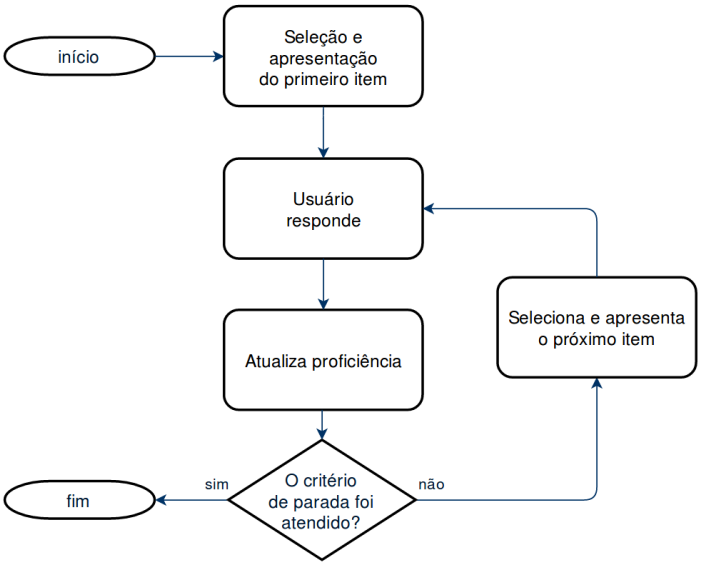
## *2.2 Testes adaptativos computadorizados*

Testes adaptativos computadorizados também conhecidos como sistemas geradores de testes, são testes administrados por computador, que de forma eficiente, reduzem o número de itens mantendo um melhor diagnóstico do desempenho do respondente (KOVATCHEVA; NIKOLOV, 2009). Essa ferramenta surgiu como uma alternativa aos testes convencionais CTT aplicados via formato papel e lápis (SPENASSATO; BORNIA; TEZZA, 2015). Trabalhos pioneiros datam da década de 70 e 80 (LORD, 1977; WEISS, 1982).

As principais vantagens do CAT em relação aos testes convencionais são: (i) testes mais rápidos e menores; (ii) aplicação em horários flexíveis; (iii) maior segurança, pois é muito mais difícil o respondente enganar e/ou trapacear; (iv) melhor controle da exposição das questões; (v) maior aderência entre os conteúdos das áreas avaliadas e o nível de conhecimento dos estudantes; (vi) maior agilidade na atualização das questões dos testes; (vii) podem fornecer feedbacks imediatos; (viii) melhor experiência na resolução dos testes; e (ix) maior precisão nas estimativas das habilidades dos respondentes considerando um menor número de itens (AL-A'ALI, 2007; WAINER et al., 2000; GUZMÁN; CONEJO, 2004).

O fluxograma de um CAT clássico, pode ser visto na Figura 1. Inicialmente é selecionada uma questão para o usuário, e após a resposta dele é estimado o nível de proeficiência. Caso o critério de parada não seja atendido, outra questão é selecionada e o fluxo se repete.

Figura 1 – Fluxograma clássico de funcionamento de um CAT




Fonte: (OLEA; PONSODA, 1996)

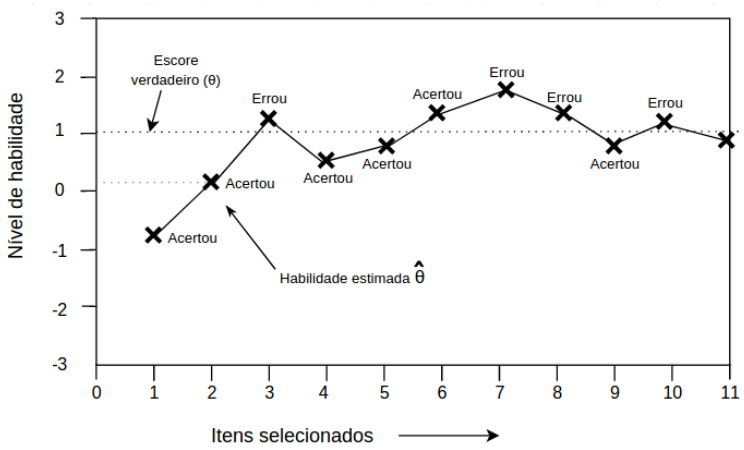
Vale destacar que os questionamentos chave encontrados em CAT são (WAINER et al., 2000):

- Como escolher a primeira questão quando não se sabe nada a respeito do respondente?
- Como escolher o próximo item após a resposta do item atual?
- Como saber o momento ideal de parar?

Esses questionamentos justificam a existência de diferentes modelos que suportam a criação de CAT, os quais, podem atender a cada problema de forma particular. Alguns exemplos desses modelos são o teste de razão de probabilidade sequencial (RECKASE; WEISS, 1983), a combinação de hierarquias de granularidade e redes bayesianas (COLLINS; GREER; HUANG, 1996), a teoria da decisão de medição (RUDNER, 2002) e a IRT. Entretanto a IRT é o modelo mais utilizado em sistemas CAT (LÓPEZ-CUADRADO et al., 2010; GUZMÁN; CONEJO, 2004; WONG et al., 2010).

A Figura 2 exemplifica um teste realizado por CAT. O termo *escore verdadeiro*, simbolizado pela letra grega $\theta$, representa o valor do nível de habilidade conhecido. Assim, esse CAT foi hipoteticamente executado considerando um usuário com *escore verdadeiro* de 1,1. Nos momentos iniciais, a medida que o respondente acerta, o seu *nível de habilidade estimado* $\hat{\theta}$ cresce. Mas percebe-se que logo após errar a terceira questão a estimativa do nível de habilidade decresce. À medida que são selecionadas mais questões, a estimativa do nível de habilidade vai ficando cada vez mais precisa.

Figura 2 – Exemplo do funcionamento de um teste realizado por CAT




Fonte: Victor Miranda Gonçalves Jatobá, 2017

Sistemas CAT são compostos por cinco componentes principais: (1) banco de itens (BI), (2) critérios iniciais (3) algoritmo ou regra para a seleção de itens, (4) método de estimação da habilidade e (5) regra de parada do teste (THOMPSON; WEISS, 2011).

## *2.3 Teoria de resposta ao item*

A Teoria de resposta ao item é composta por modelos matemáticos que procuram estabelecer a probabilidade de um respondente qualquer acertar uma determinada questão, dadas as características do item e as habilidades do avaliado (ANDRADE; TAVARES; VALLE, 2000). Apesar de ter suas origens na década de 1930, ela só foi transformada em axioma na década de 1960, tomando conta de grande parte da psicometria nos anos 1980 (PASQUALI; PRIMI, 2003).

Uma das grandes vantagens da IRT sobre a Teoria Clássica é que o cálculo do nível de aptidão do sujeito e dos parâmetros dos itens (dificuldade e discriminação) independem da amostra de itens utilizados (HAMBLETON; SWAMINATHAN; ROGERS, 1991). Isso permite a comparação entre populações, desde que submetidas a provas que tenham alguns itens comuns, ou ainda, a comparação entre indivíduos da mesma população que tenham sido submetidos a provas totalmente diferentes (QUARESMA, 2014).

Três técnicas são muito utilizadas para o cálculo da probabilidade de acerto e erro a itens dicotômicos. Essas técnicas diferem entre si no número de parâmetros que utilizam para descrever o item. São elas: (i) o modelo Rasch de uma dimensão que considera apenas

o parâmetro de dificuldade do item (ML1) (RASCH, 1960); (ii) o modelo matemático logístico unidimensional com os parâmetros de dificuldade e discriminação (ML2), originado na década de 1950 (BAKER; KIM, 2004); e por fim (iii) o modelo logístico com três parâmetros (ML3), que acrescenta o parâmetro de probabilidade de acerto ao acaso (BIRNBAUM, 1968). Considerando o modelo ML3, a probabilidade do usuário $j$, com habilidade $\theta_j$ responder corretamente o item $i$ pode ser calculada por:

$$P(U_{ji} = 1|\theta_j) = c + (1 - c)\frac{1}{1 + \exp^{-D.a_i.(\theta_j - b_i)}}, \quad (1)$$

em que:

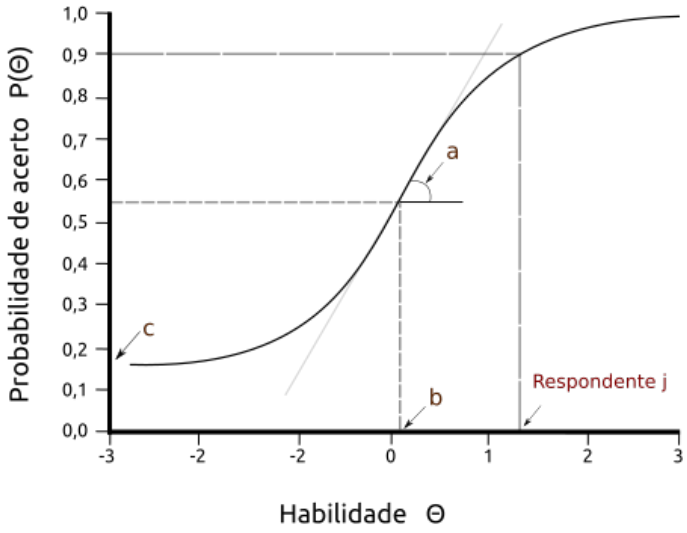
- $U_{ji}$ tém o valor 0 ou 1, representando, respectivamente, se o respondente $j$ errou ou acertou a questão $i$.
- $b_i$ é o parâmetro de dificuldade da questão;
- $a_i$ é o poder de discriminação que cada questão possui para diferenciar os participantes que dominam, dos participantes que não dominam a habilidade avaliada na questão $i$ (BRASIL, 2013). O valor de $a_i$ é proporcional à derivada da tangente da curva no ponto de inflexão no ponto $b_i$. Valores baixos de $a_i$ indicam que o item tem pouco poder de discriminação, isto é, alunos com habilidades bastante diferentes tem a mesma probabilidade de acertar a questão. Já para valores muito altos, os alunos são discriminados em dois grupos: os que possuem habilidades abaixo do valor de $b_i$ e os que possuem acima;
- $c$ é o parâmetro de acerto ao acaso. Ou seja, é a probabilidade de um participante acertar a questão não dominando a habilidade exigida (BRASIL, 2013);
- $\theta_j$ é a habilidade latente do usuário $j$;
- D é um fator de escala constante. Geralmente é utilizado o valor 1,7 para que a função logística forneça resultados semelhantes aos da função normal (GALVAO; NETO; BORGES, 2013).

Para o ML2, basta fixar $c = 0$ para todos os itens na Equação 1 e para o modelo Rasch, basta adicionalmente igualar $a_i = 1$ para todos os itens $i$.

A IRT consegue antever o desempenho do aluno, ou grupo, a uma determinada questão através da curva característica do item (CCI). Essa curva é representada através

de um gráfico que fornece informações sobre a probabilidade de cada respondente acertar um determinado item (BAKER, 2001).

Figura 3 – Exemplo de uma curva característica do item para o IRT ML3




Fonte: Victor Miranda Gonçalves Jatobá, 2017

A Figura 3 apresenta um exemplo de uma CCI para o ML3 do IRT. É possível visualizar a relação entre a probabilidade de acerto, os parâmetros do item e o nível de habilidade de respondentes. Quanto maior for o nível de habilidade, maior a probabilidade de acerto ao item ($P(\theta)$). No gráfico, o respondente $j$, com habilidade em torno de 1,4, tem, aproximadamente, probabilidade de 0,9 de acertar a questão representada. Em outras palavras, é esperado que 90% dos respondentes, com esse nível de habilidade, acertem essa questão.

Uma variação da IRT é a IRT Multidimensional (do inglês, *Multidimensional Item Response Theory* – MIRT). A principal diferença é que a MIRT analisa várias habilidades latentes do respondente e há vários parâmetros de discriminação do item (RECKASE, 2009).

## *2.4 Testes adaptativos computadorizados baseados na IRT*

CAT baseados na IRT permitem fazer testes mais precisos sem a necessidade de um número fixo de questões em um menor tempo em relação aos testes clássicos

(KOVATCHEVA; NIKOLOV, 2009), pois é possível identificar as áreas de carência do estudante e assim selecionar uma sequência de itens adaptadas aos conhecimentos do usuário (CHEN; LEE; CHEN, 2005).

A ampla adoção de IRT em CAT pode ser justificada por ela ser uma teoria moderna, que aborda as deficiências inerentes aos métodos clássicos de teste para projetar, construir e avaliar testes educacionais e psicológicos (HAMBLETON; SWAMINATHAN; ROGERS, 1991). Isso torna a IRT mais aderente em testes orientados a item (JACOBSON, 1997).

## *2.5 Banco de itens*

A montagem de um banco de itens (BI) de qualidade é importante em um sistema CAT, porque o sucesso do teste depende de dois fatores básicos: (i) o BI deve ter questões suficientes para fornecer as informações necessárias sobre o teste; e (ii) deve fornecer itens com diferentes níveis de dificuldade que consigam cobrir todos os tópicos a serem avaliados (WISE, 1997).

Em uma situação real, um BI pequeno poderia gerar imprecisão do escore do respondente, pois ao selecionar itens para aplicação, o algoritmo pode não encontrar itens disponíveis para o nível de habilidade do respondente (SPENASSATO; BORNIA; TEZZA, 2015). Alguns trabalhos sugerem que, o número de questões no BI deve ser de 5 a 10 vezes o tamanho da prova que se pretente aplicar (ÖZYURT et al., 2012). Por exemplo, se uma prova contém 30 questões é aconselhável que o BI tenha entre 150 e 300 questões.

Alguns BIs podem passar por um processo inicial de eliminação de questões muito fáceis e / ou muito difíceis (LENDYUK; RIPPA; SACHENKO, 2013). Para CAT baseados em IRT podem sem excluídas também, questões que não estão de acordo com o modelo logístico adotado (BIRNBAUM, 1968).

## *2.6 Seleção de itens*

A escolha da ISR influencia diretamente na maior eficiência e precisão na estimativa da habilidade dos respondentes de CAT em comparação a testes P&P (EGGEN; STRAETMANS, 2000). No entanto, CATs ainda possuem problemas de falta de acurácia do $\hat{\theta}$ nos estágios inicias do teste (CHEN; ANKENMANN; CHANG, 2000). Logo, a maior

acurácia do $\hat{\theta}$ depende da escolha do critério de seleção de itens mais apropriado (CHEN; ANKENMANN; CHANG, 2000; CHEN; ANKENMAN, 2004; BARRADA et al., 2008; WANG; CHANG; HUEBNER, 2011).

Apesar dos métodos para seleção de itens em CAT não estarem tão refinados como os utilizados em testes lineares (LINDEN; PASHLEY, 2009), existem duas abordagens gerais bastante consolidadas: (i) baseada em informação e (ii) bayesiana (RIJN et al., 2002). A primeira seleciona o item que fornece maior informação a respeito da estimativa de um $\theta$ específico. Já a abordagem bayesiana, seleciona itens baseados em distribuições anteriores e posteriores das estimativas de um $\theta$ específico (BUTTERFIELD, 2016).

### 2.6.1 Informação de Fisher

Em CAT, uma regra comumente utilizada para selecionar o próximo item baseado em informação é o critério de informação de Fisher (F) (LORD, 1980). A informação de Fisher do item $i$ é dada pela Equação 2.

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \tag{2}$$

em que $P_i(\theta)$ é a probabilidade do usuário acertar o item $i$ dado o nível de habilidade $\theta$. A Equação 1 traz o cálculo de $P_i(\theta)$ para o ML3. Já $P_i'(\theta)$ é a primeira derivada de $P_i(\theta)$, isto é:

$$P_i'(\theta) = \frac{\partial}{\partial\theta} P_i(\theta). \tag{3}$$

Usando a informação de Fisher, o item a ser selecionado é aquele que fornece a maior informação para um dado nível de habilidade. Quanto mais itens forem selecionados, maior será a precisão do $\theta$ e a informação do teste (HAMBLETON; SWAMINATHAN; ROGERS, 1991).

A regra de seleção F possui maior eficiência e precisão na estimativa dos escores, quando $\hat{\theta}$ está próximo do *escore verdadeiro*. Em estágios iniciais do teste, caso o $\hat{\theta}$ esteja distante do $\theta$ original, o item a ser selecionado não será o que realmente irá fornecer a maior informação do $\theta$. Isso resulta em redução de eficiência e precisão na estimativa(CHEN; ANKENMANN; CHANG, 2000).

### 2.6.2 Informação Intervalar de Fisher

O problema da regra $F$, exposto na subseção 2.6.1, pode ser menos grave se for usado um intervalo do nível de habilidade ao invés da estimativa pontual (CHEN; ANKENMANN; CHANG, 2000). Essa foi a estratégia utilizada na construção da função geral de informação ponderada (do inglês, *General Weighted Information Function* – GWI) (VEERKAMP; BERGER, 1997) vista na Equação 4.

$$GWI_i(\theta) = \int_{-\infty}^{+\infty} W(\theta)\, I_i(\theta)\, d\theta, \tag{4}$$

em que $W(\theta)$ é a função ponderada e $I_i(\theta)$ a função de informação de Fisher (Equação 2). A $GWI$ usa valores de F em uma faixa de níveis de $\theta$ ao invés de um único valor devolvido pela F. Esses valores são agregados em um único valor através de uma média ponderada. Então, ao usar o $GWI$ como critério de seleção, o item $i$ a ser escolhido é aquele que possuir a máxima média ponderada.

A depender dos requisitos de um CAT, o critério de seleção do próximo item pode ser adaptado aplicando ajustes no $W(\theta)$. Por exemplo, quando o $\hat{\theta}$ estiver próximo ao $\theta$ é mais apropriado o uso da F para selecionar o próximo item (CHEN; ANKENMAN, 2004). Nesse caso, $W(\theta)$ poderia ser expressa da seguinte forma:

$$W(\theta) = \begin{cases} 1 & Se \quad \theta = \hat{\theta} \\ 0 & \text{Caso contrário.} \end{cases} \tag{5}$$

Quando se considera a incerteza da estimativa do $\theta$ nos estágios iniciais do CAT, pode ser razoável considerar o peso de cada nível de $\theta$ quando este estiver em um intervalo de confiança do $\theta$ verdadeiro, caso contrário, deve-se igualar o peso a zero. Neste caso, a função ficaria:

$$W(\theta) = \begin{cases} 1 & Se \quad \theta \in \left(\hat{\theta}_l, \hat{\theta}_u\right) \\ 0 & \text{Caso contrário.} \end{cases} \tag{6}$$

Na Equação 6, $(\hat{\theta}_l, \hat{\theta}_u)$ é o intervalo de confiança para a habilidade atual estimada. Considerando esse intervalo de confiança e substituindo a função W($\theta$) da Equação 6 na Equação 4, obtemos a função de informação intervalar de Fisher (FII) (VEERKAMP; BERGER, 1997):

$$FII_i(\theta) = \int_{\hat{\theta}_l}^{\hat{\theta}_u} I_i(\theta)\, d\theta. \tag{7}$$

Esta função, representa uma área em $F$ que está sobre o intervalo de confiança $\left(\hat{\theta}_l, \hat{\theta}_u\right)$. Ao escolher a $FII$ como critério de seleção em CAT, o próximo item será aquele que fornece a área máxima (CHEN; ANKENMANN; CHANG, 2000).

A diferença entre a regra $F$ e a $FII$ é o uso da função ponderada $W(\theta)$. Essa diferença fornece a $FII$ uma gama de níveis de $\theta$ e, portanto, deve (teoricamente) funcionar de forma diferente de $F$, que considera apenas um único nível de $\theta$ (CHEN; ANKENMANN; CHANG, 2000).

Outra característica da $FII$ é que a medida que itens sucessivos são selecionados em CAT, o intervalo de confiança se estreita e converge para uma estimativa pontual do verdadeiro $\theta$. Assim, após a seleção de um grande número de itens, a função ponderada se torna uma função pontual. Ou seja, a $FII$ em um determinado ponto se torna a própria função $F$. Essa é a principal vantagem do uso da $FII$, já que a função $F$ é mais adequada em estágios mais avançados de CAT (CHEN; ANKENMANN; CHANG, 2000).

A delimitação da distância entre os pontos do intervalo de confiança da $FII$ é muito importante no processo de seleção do item. Por conta desse fator, não é possível demonstrar as vantagens desta sobre a $F$ (VEERKAMP; BERGER, 1997).

### 2.6.3 Informação Ponderada pela Máxima Verossimilhança

A regra de Informação Ponderada pela Máxima Verossimilhança ($MLWI$) (VEERKAMP; BERGER, 1997) utiliza a função de verossimilhança como a função de peso e é definida como a área sob uma função, que é um produto da função de verossimilhança e da função de informação:

$$MLWI_i(\theta) = max_{i \in I_n} \int_{-\infty}^{+\infty} I_i(\theta)\, L(\theta | x_1, ..., x_k)\, d\theta, \tag{8}$$

na qual $k$ é o numero de itens administrados, $x_1, ..., x_k$ do padrão de respostas provisórias e $L(\theta|x_1, ..., x_k)$ a função de verossimilhança avaliada em $\theta$, dado o padrão de respostas provisórias. O problema de encontrar múltiplos máximos locais na função de verossimilhança é contornado usando este critério (VEERKAMP; BERGER, 1997).

### 2.6.4 Informação Ponderada a *Posteriori*

Outra alternativa é aplicar a distribuição a *posteriori* como uma função de peso. Esta abordagem deu origem a regra de Máxima Informação ponderada a *Posteriori* (do inglês, *Maximum Posterior Weighted Information* – MPWI) (LINDEN, 1998). Considerada uma abordagem bayesiana, esta função de peso considera toda a gama de níveis de $\theta$. Com isso, a regra $MLWI$, prevista na Equação 8, adiciona a distribuição a *posteriori* de $\theta$.

$$MPWI_i(\theta) = max_{i \in I_n} \int_{-\infty}^{+\infty} I_i(\theta)\, \pi(\theta \mid X_n)\, L(\theta|x_1, ..., x_k)\, d\theta, \tag{9}$$

A função $\pi(\theta \mid X_n)$ representa a distribuição a *posteriori* de $\theta$ após a seleção de $n$ itens. Nesta abordagem todos os níveis de $\theta$ são considerados para a seleção do próximo item. A distribuição a *posteriori* é uma função de uma distribuição a priori e a função de verossimilhança. À medida que questões são selecionadas, a função de verossimilhança engloba a distribuição a priori, tornando-se assim muito parecida com a $F$.

A principal limitação dessa ISR é a grande influência que a distribuição a priori possui no desempenho das estimativas (CHEN; ANKENMANN; CHANG, 2000). Esse problema se torna ainda mais grave em estágios iniciais de um CAT, já que poucas questões foram respondidas.

### 2.6.5 Informação de Kullback-Leibler

Outra alternativa às regras anteriores é a Informação de Kullback-Leibler ($KL$) (CHANG; YING, 1996). Essa estratégia, que é baseada em informação, é uma medida geral para a *distância* entre duas distribuições (LINDEN; PASHLEY, 2009).

A Equação 10 mostra o funcionamento da equação geral da função de informação de $KL$.

$$KL_i\left(\hat{\theta}, \theta_0\right) = E\left[\log \frac{L(\theta_0)}{L(\hat{\theta})}\right], \tag{10}$$

em que $\theta_0$ representa o valor real do nível de habilidade do respondente. A partir da resolução da Equação 10, é obtida a Equação 11, na qual é definida uma informação global entre duas distribuições de probabilidade.

$$KL_i\left(\hat{\theta}, \theta_0\right) = P_i(\theta_0) \log\left[\frac{P_i(\theta_0)}{P_i(\hat{\theta})}\right] + [1 - P_i(\theta_0)] \log\left[\frac{1 - P_i(\theta_0)}{1 - P_i(\hat{\theta})}\right]. \tag{11}$$

Quando $\hat{\theta} = \theta_0$ o valor de $KL$ é zero. Isso significa que o item não consegue distinguir entre respondentes de mesmo nível de habilidade. Na perspectiva contrária, quando $\theta$ e $\theta_0$ são muito diferentes, o valor de $KL$ é muito grande.

Para tratar essa limitação, pode ser usado, assim como na $FII$, um intervalo de confiança para uma função de peso. Essa alternativa à KL, pode ser vista na Equação 12.

$$KL_i(\hat{\theta}) = \int_{\hat{\theta}_l}^{\hat{\theta}_u} KL_i(\hat{\theta}, \theta)\, d\theta. \tag{12}$$

Dessa forma, a $KL$ conforme a Equação 12, herda algumas característica da $FII$. Uma delas é a de não convergir para o verdadeiro escore quando os $\hat{\theta}$s estão distantes dos $\theta$s originais. Essa característica geralmente está presente nos estágios inicias de CATs.

Ademais, impor um intervalo de confiança finito na função assimétrica $KL_i(\hat{\theta}, \theta_0)$ pode não ser apropriado, sugerindo assim a implementação de uma abordagem bayesiana (CHEN; ANKENMANN; CHANG, 2000).

### 2.6.6 Informação de Kullback-Leibler com Distribuição a Posteriori

De forma similar ao $MPWI$, a $KL$ também pode incorporar uma distribuição a *posteriori* (do inglês, *Kullback-Leibler Information with a Posterior Distribution* – KLP) (CHANG; YING, 1996) como forma de melhoria na seleção do próximo item. A abordagem bayesiana $KLP_i(\hat{\theta})$ (representada na Equação 13), troca a função de peso pela distribuição a *posteriori* e, ao invés do intervalo de confiança finito, torna o $\theta \in (-\infty, \infty)$.

$$KLP_i(\hat{\theta}) = \int_{-\infty}^{+\infty} p(\theta \mid X_n)\, KL_i(\theta, \hat{\theta})\, d\theta, \tag{13}$$

Com o $KLP$, os altos valores de $KL$ para níveis extremos de $\theta$ são equilibrados pela distribuição de densidade posterior, o qual, tipicamente, possui valores de probabilidade menores para níveis de $\theta$ extremos (CHEN; ANKENMANN; CHANG, 2000). Portanto é mais apropriado o uso de uma distribuição a *posteriori* do que um intervalo de confiança usado pela $KL$. Porém, a influência da distribuição a priori, especialmente nos estágios iniciais do CAT, continua sendo um ponto importante.

### 2.6.7 Desempenho das regras de seleção de itens

Comumente CATs podem usar junto com as ISRs algum mecanismo para controlar a exposição de itens ou aplicar algum tipo de balanceamento de conteúdo (CHEN; ANKENMAN, 2004). O uso destas vai depender do nível de segurança e sigilo que é exigido quanto ao acesso as questões (SPENASSATO et al., 2016). No entanto a inserção de restrições deve ser feita com cautela, pois pode resultar em diminuição da acurácia da estimativa das habilidades se comparado ao CAT sem restrições (CHEN; ANKENMAN, 2004; LINDEN, 1999)

Considerando CATs sem restrições, as técnicas $MPWI$ e $KLP$ tendem a melhorar a seleção de itens em estágios iniciais de CAT, especialmente para níveis de habilidade com valores negativos extremos (para $\theta = -3, -2$) se comparados a técnica F (CHEN; ANKENMANN; CHANG, 2000; CHEN; ANKENMAN, 2004).

Já a técnica F, sob o mesmo cenário, é mais precisa para estimar o nível de habilidade de usuários com $\theta$s próximos a zero em estágios iniciais de CATs se comparada as técnicas $MPWI$ e $KLP$ (CHEN; ANKENMANN; CHANG, 2000). Entretanto estudos também indicam que as regras $MLWI$ e $MPWI$ tem melhor desempenho que a $F$ considerando todos os níveis de $\theta$ (VEERKAMP; BERGER, 1997; LINDEN, 1998; LINDEN; PASHLEY, 2009).

## *2.7 Estimação*

Existem vários métodos para estimar os parâmetros dos itens (processo também conhecido por calibração) e as habilidades latentes (também conhecido por nível de habilidade e representado pela letra grega $\theta$) de respondentes. Um deles é o método de máxima verossimilhança (do inglês, *Maximum Likelihood Estimator* – MLE).

Para estimar a habilidade latente de um usuário em provas dicotômicas, o MLE examina um conjunto de avaliações $E$, em que cada avaliação é representada por uma tripla $(u, q, r)$, em que $r$ é a resposta do usuário $u$ à questão $q$, que pode ser 1 (um) para acertou ou 0 (zero) para errou. Considerando $P$ como a probabilidade do usuário acertar a questão (Equação 1) e Q a probabilidade do o usuário errar ($Q = 1 - P$), o cálculo para determinar a máxima verossimilhança pode ser visto na Equação 14.

$$L_E(\theta) = \prod_{e=1}^{E} P(\theta)^{r_e} Q(\theta)^{1-r_e}. \tag{14}$$

$L_E$ é calculado para obter a estimativa atual da habilidade de um respondente ($\hat{\theta}$) considerando um conjunto de respostas $E$. No MLE, o $\hat{\theta}$ é o valor de $\theta$ que maximiza a função $L_E$ para um determinado padrão de escore do item. No CAT, é selecionado o próximo item em relação a esse $\hat{\theta}$ (MEIJER; NERING, 1999).

Porém o MLE possui alguns problemas. Um deles é a inexistência de uma estimativa perfeita para padrões de escores com todas as respostas corretas ou incorretas. Além disso, o $\hat{\theta}$ é superestimado para valores positivos de $\theta$, e subestimado para valores negativos de $\theta$ (LORD, 1983).

Uma alternativa ao MLE é a estratégia de máxima verossimilhança ponderada (do inglês, *Weighted Maximum Likelihood* – WML) (WARM, 1989). Para o ML1 e o ML2 o peso presente na WML equivale a raiz quadrada da função de informação do teste (MEIJER; NERING, 1999). A WML é menos tendenciosa (*unbiased*) do que MLE quando a mesma variância assintótica é utilizada em uma distribuição normal (WARM, 1989).

Outras duas alternativas são a, Esperança a Posteriori (do inglês, *Expected a Posteriori* – EAP ) (BOCK; MISLEVY, 1982) e a Máximo a Posteriori (do inglês, *Maximum a Posteriori* – MAP). Uma limitação presente nessas abordagens é que quando

a diferença entre a verossimilhança estimada e a média da distribuição anterior é muito grande, o resultado de $\hat{\theta}$ regressará a média do anterior (MEIJER; NERING, 1999).

A abordagem bayesiana é outra alternativa. A escolha entre essa abordagem e um método de máxima verossimilhança dependerá do papel que a precisão do erro padrão e BIAS exercem na tomada de decisão a partir dos resultados do CAT (WANG; VISPOEL, 1998).

Apesar das distinções entre os métodos de estimação, só é perceptível a diferença na eficiência das diferentes técnicas apenas em estágios iniciais de um teste (WANG; VISPOEL, 1998; DOEBLER, 2012).

## *2.8 Critérios iniciais*

Um dos primeiros dilemas em testes adaptativos é como iniciar um teste quando não se sabe nada a respeito do respondente. Uma alternativa é considerar um valor médio para o seu $\theta$ inicial. Na IRT o valor 0.0 representa o valor médio na escala de habilidade que vai de -4 a 4 (THOMPSON; WEISS, 2011). Outra alternativa é iniciar a habilidade pela média da distribuição de habilidade da população ou estabelecer um valor aleatório(LINDEN, 1999).

Quando é possível, algumas aplicações submetem o usuário a um conjunto de questões fixas, e a partir das respostas, as habilidades são estimadas para que os próximos itens sejam adaptados (RAJAMANI; KATHIRAVAN, 2013). Além disso, o teste pode ser inicializado, também, considerando um conjunto de informações a *priori*. Essas informações podem ser escores de outros testes, sejam eles tradicionais ou adaptativos.

Spenassato, Bornia e Tezza (2015) fizeram um levantamento sobre formas de inicialização em diferentes testes e perceberam que testes na área de saúde geralmente usam a primeira questão fixa, para todos os respondentes. Em contra partida, avaliações educacionais costumam variar o primeiro item, para evitar compartilhamento de respostas, entre os participantes.

## *2.9 Critérios de parada*

A escolha de uma regra de finalização do teste pode variar bastante a depender do cenário de teste a ser aplicado. Alguns CATs podem adotar até mais de um critério de parada. Apesar da diversidade, existem alguns critérios de parada geralmente adotados, como:

- Atingiu o limite máximo de questões do BI (Heitink; Veldkamp, 2015; JEONG; HONG, 2013; RAJAMANI; KATHIRAVAN, 2013).
- Atingiu um número fixo de questões a serem aplicadas (VISPOEL, 1998; DOEBLER, 2012).
- Atingiu um limite de variação do Erro Padrão da habilidade do respondente entre um item e outro (ZITNỲ et al., 2012; VELDKAMP; MATTEUCCI, 2013; MEIJER; NERING, 1999; SPENASSATO et al., 2016).
- Atingiu o tempo limite. Entretanto não é necessário em CAT predeterminar um tempo limite, pois estudantes podem gerar escores imprecisos, por se sentirem cansados em provas com longas durações (AL-A'ALI, 2007).
- O BI não possui mais questões relevantes. Geralmente acontece com um BI pequeno ou com pouca diversidade em níveis de dificuldade.
- O usuário escolhe quando parar (Heitink; Veldkamp, 2015; LENDYUK; RIPPA; SACHENKO, 2013; JEONG; HONG, 2013).

## *2.10 Exame nacional do ensino médio*

O Exame Nacional do Ensino Médio é uma prova de formato P&P promovido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (GONÇALVES; ALUÍSIO, 2015). Foi criado em 1998, mas recebeu uma reformulação a partir de 2009, em que passou a ser usado como meio de seleção para o ingresso nas universidades (COSTA, 2015).

O ENEM parte do conceito de competências, que se traduz em habilidades, conhecimentos e atitudes para resolver cada situação-problema. O exame é estruturada em cinco provas. Uma das provas é a Redação e as outras estão estruturadas em quatro macroáreas, são elas: Ciências da Natureza e suas Tecnologias; Ciências Humanas e suas Tecnologias;

Linguagens, Códigos e suas Tecnologias; e Matemática e suas Tecnologias (ANDRADE, 2013).

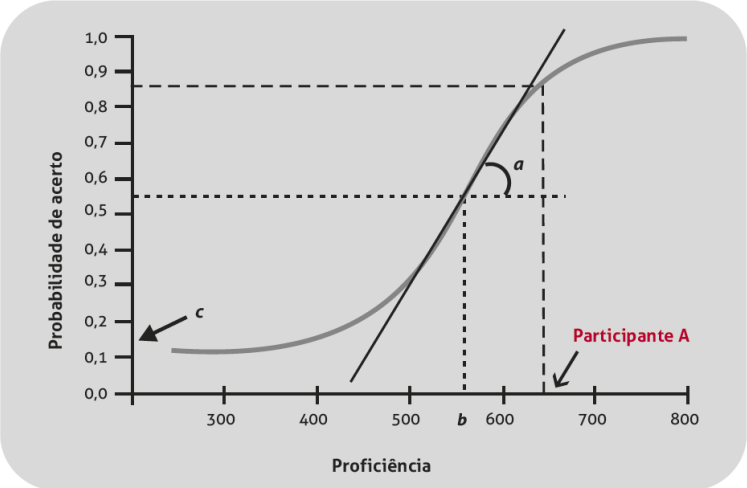
Os resultados do ENEM podem ser utilizados para: (i) compor a avaliação de medição da qualidade do Ensino Médio no país; (ii) a implementação de políticas públicas; (iii) a criação de um teste de referência nacional para o aperfeiçoamento dos currículos do Ensino Médio e (iv) o desenvolvimento de estudos e indicadores sobre a educação brasileira (KARINO; BARBOSA, 2011).

As provas são dicotômicas com possibilidade de múltipla escolha e para calcular a nota dos participantes, o exame, a partir 2009, passou a adotar a IRT. O Modelo Logístico utilizado pela IRT é o de três parâmetros (ML3). Cada prova tem 45 itens e o cálculo do nível de habilidade dos respondentes é estimado pelo método Esperança a Posteriori (EAP).

Para medir o nível de conhecimento (proficiência) do participante nas quatro macroáreas, o ENEM utiliza uma métrica (escala) criada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Cada macroárea possui uma escala que depende de dois valores: (a) a posição de referência que configura o desempenho médio e é representado pelo valor 500; e (b) o valor de dispersão representado pelo desvio padrão das notas, com valor 100 atribuído (BRASIL, 2013). Por exemplo, um participante com nota 400 apresenta proficiência com uma unidade de desvio padrão abaixo do nível médio dos participantes.

A Figura 4 apresenta um exemplo hipotético de uma CCI na prova do ENEM. A única diferença em relação à Figura 3, está no eixo $x$. Nesse eixo, os valores de proficiência são dados pela escala adotada pelo ENEM. No exemplo da Figura 4 o *Participante A*, com proficiência de 650, possui 85% de chance de acertar a questão.

Figura 4 – Curva característica do item para o IRT ML3 da prova do ENEM




Fonte: (BRASIL, 2013)

# 3 Abordagem personalizada para o processo de seleção de itens em CATs

No CAT clássico, apenas uma ISR é considerada para a seleção da sequência das questões. Entretanto as regras de seleção de itens possuem variações de desempenho na estimativa das habilidades a depender da proficiência do respondente e do estágio em que o teste se encontra (CHEN; ANKENMANN; CHANG, 2000; CHEN; ANKENMAN, 2004; VEERKAMP; BERGER, 1997; LINDEN, 1998; LINDEN; PASHLEY, 2009).

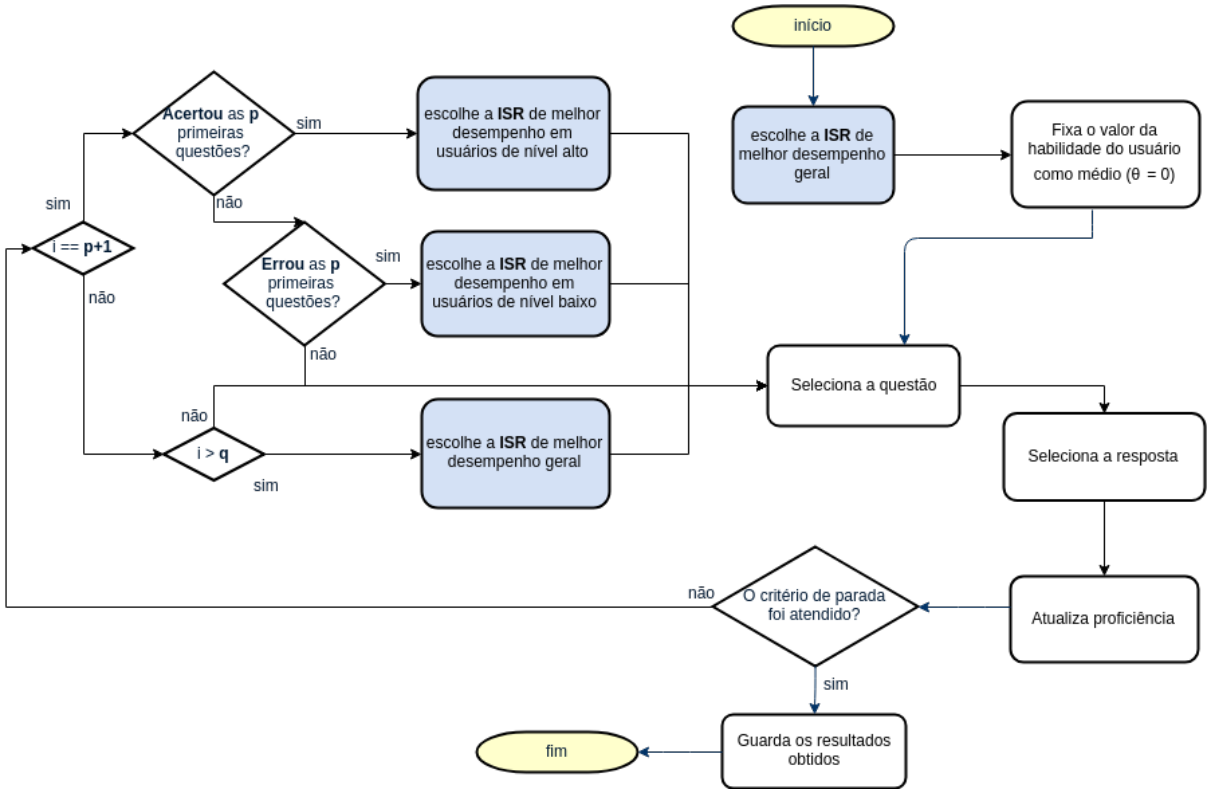
Neste capítulo é apresentada a principal contribuição deste trabalho, uma estratégia para personalizar o processo de seleção de itens que escolhe a ISR a depender do desempenho do aluno no cumprimento da prova e do momento em que o teste se encontra, isto é, são consideradas mais de uma ISR. Essa estratégia é chamada de ALICAT (*personALIzed* CAT, em português, CAT personalizado), a qual escolhe a regra de seleção de itens a depender do padrão de respostas do usuário. Se este respondeu corretamente as **p** primeiras perguntas, então a ISR selecionada para os próximos **q** itens, será a que obteve o melhor desempenho na estimativa de escores de usuários de nível alto. Errando as **p** primeiras, será o inverso. Nesse caso a ISR selecionada será a que obteve melhor desempenho para usuários de nível baixo. Em qualquer outro caso, será utilizada a regra que obteve melhor desempenho geral (ver Figura 5).

Na Figura 5, a variável $i$ representa a quantidade de itens selecionados. Já as variáveis $p$ e $q$ (com $p < q$) representam o momento do teste em que pode ser selecionada outra ISR diferente da que começou. As ações em azul, contém o momento do teste em que são selecionadas diferentes regras.

A estratégia de voltar a utilizar a melhor regra geral após a aplicação de **q** itens, se deve ao fato de que, a medida que o teste cresce, o desempenho de todas as regras tende a se manter similar (CHEN; ANKENMANN; CHANG, 2000).

Dessarte, para o funcionamento da abordagem ALICAT é preciso identificar, no cenário do teste a ser aplicado, qual a regra que apresenta o melhor desempenho geral e quais as regras que possuem melhor desempenho na estimativa dos escores de usuários com níveis altos e baixos.

Figura 5 – Fluxograma de funcionamento da abordagem ALICAT




Fonte: Victor Miranda Gonçalves Jatobá, 2017

# 4 Método de pesquisa

A seguir são descritas as abordagens para a captação dos dados e a montagem do banco de itens. Outrossim, são elucidados os processos de condução e avaliação do estudo comparativo entre as regras de seleção de itens. Por fim, são apresentadas a configuração e a forma de avaliação da abordagem ALICAT.

## *4.1 Captação dos dados e montagem do banco de itens*

Foram utilizados os dados da prova do ENEM de 2012, os quais são públicos e foram retirados do portal da transparência (TEIXEIRA, 2016). O modelo de dados é composto por um conjunto de respostas dicotômicas de avaliados selecionados de forma aleatória. Cada resposta $u$, pode conter apenas dois possíveis valores, que são: 1 para "acertou" e 0 para "errou", isto é, $u = \{0, 1\}$, caracterizando a nossa distribuição como uma Bernoulli.

Como suposição a ser utilizada por todo o texto, temos que nossa amostra é independente e igualmente distribuída, pois as respostas oriundas de indivíduos diferentes são independentes.

A amostra inicial foi composta por 1000.000 respondentes da prova Rosa de Matemática e suas tecnologias. Os valores dos parâmetros $a$, $b$ e $c$ dos 45 itens foram retirados do trabalho de SPENASSATO et al. (2016). Esses parâmetros serviram de base para estimar o nível de habilidade dos respondentes.

## *4.2 Estimação das habilidades*

Após a montagem do banco de itens, a próxima etapa é estimar o nível de habilidade dos respondentes. Estes níveis foram estimados com o uso do *software* ICL (HANSON, 2002).

O método para a estimação das habilidades foi a Esperança a Posteriori (do inglês, *Expected a Posteriori* – EAP ), o mesmo utilizado pelo trabalho de SPENASSATO et al. (2016). Essas estimativas são aqui consideradas como os *escores verdadeiros* e

representados pela letra grega $\theta$. São denominadas assim, pois foram estimados em relação à prova completa de 45 itens.

Após estimar os *escores verdadeiros*, estes foram classificados em 10 grupos entre -2 e 3.5 (ver Tabela 1). A primeira intenção foi entender o comportamento da etapa de estimação. O menor e maior $\theta$ foram, respectivamente, $-1.716215$ e 3.083216. Nota-se que existe uma menor quantidade de respondentes com $\theta$s altos (maiores que 2) e $\theta$s baixos (menores que -1.5).

Tabela 1 – Classificação, por intervalo, dos $\theta$s verdadeiros

| Intervalo dos $\theta$s verdadeiros | Tamanho da amostra | Menor $\theta$ | Maior $\theta$ | Média dos $\theta$s |
|---|---|---|---|---|
| [ -2 ; -1,5 ] | 12193 | -1,716215 | -1,500013 | -1,58 |
| ] -1,5 ; -1 ] | 121331 | -1,499992 | -1,000002 | -1,19 |
| ] -1 ; -0,5 ] | 211557 | -0,999995 | -0,500002 | -0,75 |
| ] -0,5 ; 0 ] | 193117 | -0,499994 | -1e-06 | -0,25 |
| ] 0 ; 0,5 ] | 163131 | 1,3e-05 | 0,499995 | 0,24 |
| ] 0,5 ; 1 ] | 139804 | 0,500006 | 0,999997 | 0,74 |
| ] 1 ; 1,5 ] | 95364 | 1,000001 | 1,499987 | 1,23 |
| ] 1,5 ; 2 ] | 53275 | 1,500001 | 1,999989 | 1,71 |
| ] 2 ; 2,5 ] | 9435 | 2,000034 | 2,499647 | 2,16 |
| ] 2,5 ; 3,5 ] | 793 | 2,500007 | 3,083216 | 2,66 |

De cada grupo, foram retirados, de forma aleatória, 500 respondentes que totalizaram outra amostra de 5000. Isso foi importante para garantir que todos os grupos de níveis de usuários façam parte da etapa de simulação do CAT. Desse grupo de 5000 respondentes, foram considerados apenas aqueles que responderam no mínimo a 40 itens, totalizando 4979 respondentes.

### *4.3 Configuração do CAT*

Para a montagem dos CATs, utilizamos o pacote catR (MAGIS; RAÎCHE et al., 2012) do *software* R. Não foi preciso implementar nenhum critério de exposição de itens, pois todos os avaliados foram submetidos aos mesmos 45 itens. Também nenhuma restrição para balanceamento de conteúdo foi desenvolvida, pois não é divulgado a qual conteúdo cada questão pertence.

Foi adotada a medida utilizada por Spenassato et al. (2016) para identificar o comprimento dos CATs. Esta medida permite encontrar a quantidade máxima de questões

necessária para que um CAT consiga estimar, com um certo grau de precisão, o escore do avaliado. Para tal, é preciso verificar em qual ponto do teste a precisão da estimativa do $\hat{\theta}$ tende a se manter estável.

Para verificar o ponto de estabilidade considera-se o cálculo do Erro Padrão (do inglês, *Standard Error* – $SE$) utilizando o método $EAP$. O $SE$ necessita de três parâmetros: (a) o $\hat{\theta}$ do respondente (b) o conjunto de itens, contendo seus respectivos parâmetros $a$, $b$ e $c$; e (c) as respostas do usuário a estes itens. O $SE$ foi estimado pelo método *semTheta* disponível no pacote catR do *software* R.

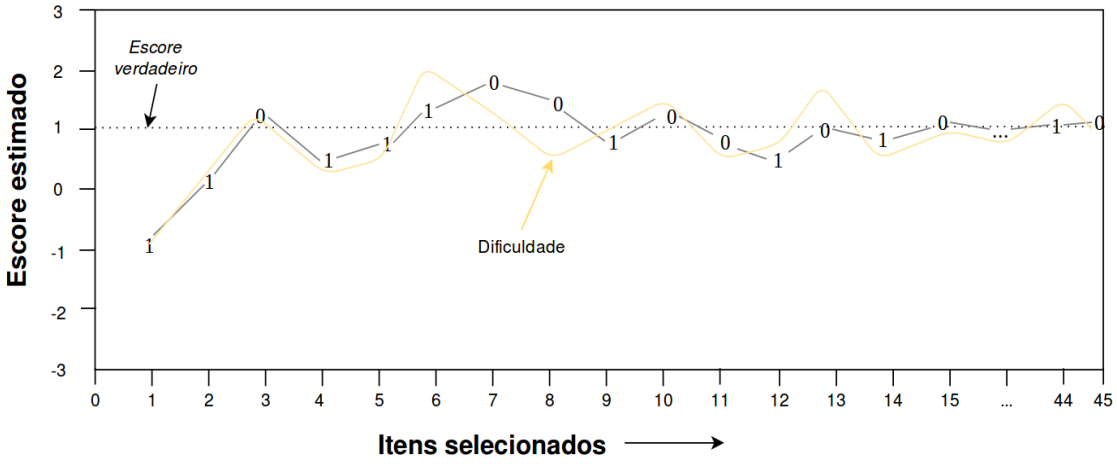
Temos então que o ponto de estabilidade é o momento do teste, no qual, a diferença do erro padrão entre o item atual aplicado ($SE_{i,j}$) e o item anterior ($SE_{i-1,j}$ ) é inferior a 1% do erro padrão do item anterior. O cálculo é expresso na seguinte equação:

$$|SE_{i,j} - SE_{i-1,j}| < |0,01 \times SE_{i-1,j}|, \tag{15}$$

em que, a variável $j$ representa o avaliado, que vai de 1 até 4979. Já a variável $i$ se refere ao item, que vai de 1 até, no mínimo 40, e no máximo 45. O $SE_{i,j}$ corresponde ao erro padrão estimado do escore do respondente $j$, após a aplicação de $i$ itens. O valor da Equação 15 informa que, mesmo após aplicação de mais itens, a precisão tende a se manter estável. Isso significa que o teste pode ser finalizado sem perdas ou prejuízos na precisão das estimativa dos escores (SPENASSATO et al., 2016).

A Figura 6 ilustra um exemplo hipotético de uma aplicação de um CAT a um respondente com *escore verdadeiro* igual a $1,1$ para uma prova de 45 itens. O valor 1 (um) representa que o respondente acertou a questão e 0 (zero) que errou. A linha amarela representa o valor do parâmetro de dificuldade de cada item selecionado de forma sequencial.

Figura 6 – Exemplo da aplicação de um CAT com 45 itens a um respondente com $\theta = 1,1$




Fonte: Victor Miranda Gonçalves Jatobá, 2017

É possível observar que a partir do item 13, a estimativa do escore atual ($\hat{\theta}$) já está bem próximo do verdadeiro escore. Isso significa que, caso o comportamento de todos os avaliados fosse similar, o valor 13 poderia ser um bom comprimento para o CAT. No exemplo, isso representaria uma redução de 71,1% da prova original de 45 itens.

Isto posto, foram aplicados, para cada CAT, uma das regras de seleção de itens $F$, $KL$, $KLP$, $MLWI$ e $MPWI$ com $\hat{\theta}$ inicial fixado em 0 (zero). Na execução dos CATs foram identificados os pontos de estabilidade de cada respondente. Todos os pontos foram classificados em 10 grupos, que vão de $-2$ a $3,5$. Para cada grupo, foi retirada a média dos pontos pertencentes. Assim o comprimento máximo do CAT foi definido pela maior média dentre as 10.

## *4.4 Avaliação das ISRs*

Após calcular o comprimento máximo $n$ do teste para cada ISR, estes são novamente executados, agora considerando o novo comprimento fixo.

Para avaliar o desempenho das ISRs na estimação das habilidades, foram calculados o BIAS (viés) médio (Equação 16) e a raiz do erro quadrático médio (do inglês, *Root Mean Squared Error of Estimation* – RMSE) definida na Equação 17.

$$BIAS(n) = \frac{1}{R} \sum_{k=1}^{R} \left( \hat{\theta}_{n,k} - \theta_k \right), \tag{16}$$

$$RMSE(n) = \sqrt{\frac{1}{R}\sum_{k=1}^{R}\left(\hat{\theta}_{n,k} - \theta_k\right)^2}. \quad (17)$$

Nessas equações $\theta_k$ é o *escore verdadeiro* do $k$-ésimo respondente; $R$ é o número total de respondentes e $\hat{\theta}_{n,k}$ é o valor estimado da habilidade do $k$-ésimo respondente após aplicar $n$ itens. Essas estatísticas foram capturadas na seleção dos primeiros 30 itens e no final da execução de cada CAT.

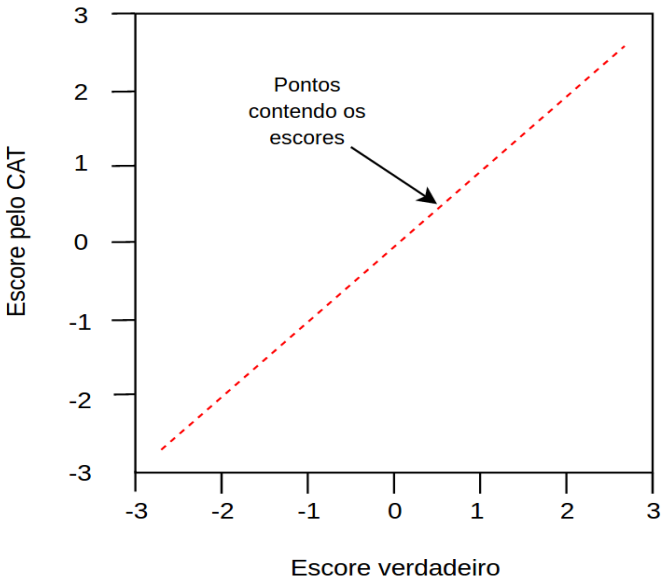
## *4.5 Configuração e avaliação da abordagem ALICAT*

Após os resultados da avaliação das ISRs, são conhecidas as regras de melhor desempenho geral, e as de melhor desempenho na estimativa dos escores de usuários com níveis altos e baixos. São considerados de nível alto, os valores acima ou iguais a $2,5$, e de nível baixo, os valores abaixo ou iguais a $-1,5$. Somente após a definição das regras supracitadas é que será possível construir o ALICAT (abordagem proposta na Seção 3).

Tendo definida a configuração do ALICAT, a próxima etapa é avaliar o desempenho deste e o comparar com os outros CATs que utilizam exclusivamente apenas uma ISR. Essa etapa segue o mesmo processo da Seção 4.4, pelos quais são avaliadas as ISRs. Após identificar o comprimento do ALICAT, os testes serão processados novamente para todos os respondentes da amostra, porém com o número fixo de questões. Durante a execução do teste, serão calculados o $BIAS$ e o $RMSE$.

Por fim, como parte da comparação do desempenho entre o ALICAT e os CATs que utilizam exclusivamente uma regra de seleção de itens, serão apresentados gráficos como os da Figura 7. Essa figura contém um exemplo de comparação entre os *escores verdadeiros* e os escores obtidos via CAT. Percebe-se nesse exemplo que os escores obtidos pelo CAT tiveram os mesmos valores que os *escores verdadeiros*, que seria o caso ideal.

Figura 7 – Gráfico exemplificando o que seria um contexto ideal de estimativas das habilidades dos respondentes realizadas por um CAT. A linha tracejada vermelha, representa possíveis valores dos escores obtidos pelo CAT em relação ao *escore verdadeiro*.




Fonte: Victor Miranda Gonçalves Jatobá, 2017

# 5 Resultados

Os resultados foram extraídos da execução de cada CAT, considerando sua respectiva ISR. Foram contemplados o desempenho geral destas, além de seus respectivos pontos de estabilidade. Também, foram calculados o BIAS e o RMSE nos estágios iniciais e no final de cada CAT. Por fim, são apresentados os resultados da etapa de configuração e execução do ALICAT.

## *5.1 Desempenho das ISRs no teste completo*

A Tabela 2 contém a quantidade de respondentes ($\sigma$) e a média de itens selecionados ($\bar{x}$) para cada intervalo de $\hat{\theta}$. Os valores em negrito foram os comprimentos máximos definidos para cada ISR (denotado por $n$). Com isso, todas as ISRs foram novamente executadas considerando a regra de parada fixada em $n$.

Tabela 2 – Resultados da identificação do ponto de estabilidade (média de itens selecionados - $\bar{x}$) e do tamanho da amostra ($\sigma$) para cada ISR em diferentes intervalos de $\hat{\theta}$.

| Intervalo dos $\hat{\theta}$ estimados | F | | KL | | KLP | | MLWI | | MPWI | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma$ | $\bar{x}$ | $\sigma$ | $\bar{x}$ | $\sigma$ | $\bar{x}$ | $\sigma$ | $\bar{x}$ | $\sigma$ | $\bar{x}$ |
| ] -2 ; -1.5 ] | 502 | 7 | 48 | 14 | 301 | 8 | 0 | - | 0 | - |
| ] -1.5 ; -1 ] | 439 | 8 | 548 | 8 | 552 | 8 | 3 | **23** | 797 | 8 |
| ] -1 ; -0.5 ] | 611 | 10 | 477 | 8 | 740 | 9 | 7 | 17 | 723 | 8 |
| ] -0.5 ; 0 ] | 490 | 14 | 1047 | 7 | 385 | 13 | 2544 | 6 | 517 | 9 |
| ] 0 ; 0.5 ] | 666 | 21 | 629 | 8 | 506 | 14 | 771 | 6 | 585 | 15 |
| ] 0.5 ; 1 ] | 281 | 19 | 369 | 10 | 708 | 9 | 377 | 7 | 395 | 13 |
| ] 1 ; 1.5 ] | 665 | **35** | 406 | **21** | 242 | **24** | 166 | 10 | 454 | **18** |
| ] 1.5 ; 2 ] | 152 | 25 | 554 | 12 | 1512 | 7 | 35 | 19 | 1506 | 9 |
| ] 2 ; 2.5 ] | 1142 | 19 | 901 | 15 | 3 | 22 | 769 | 14 | 2 | 2 |
| [ 2.5 ; 3 ] | 31 | 27 | 0 | - | 0 | - | 307 | 9 | 0 | - |

Em linhas gerais, as regras $KLP$ e $MPWI$ praticamente não conseguiram estimar usuários de nível alto ($\theta \geq 2$). Com isso, colocaram todos os respondentes com $\theta$s verdadeiros maiores que 1.5 no intervalo ] 1.5 ; 2 ]. No outro extremo, a regra $MLWI$ teve pouco êxito em estimar usuários com níveis baixos ($\theta \leq -0.5$).

A regra de seleção $F$ foi a que selecionou, em média, um maior número de itens para um determinado grupo de $\hat{\theta}$ e a $MPWI$ foi a que selecionou menos, $n = 35$ e $n = 18$ itens,

respectivamente. Praticamente todas as médias máximas de número de itens selecionados foram encontradas no intervalo de 1 a 1.5, exceto para a regra de seleção $MLWI$. O resultado obtido para a regra F é similar ao encontrado no trabalho de SPENASSATO et al. (2016), no qual a média máxima de itens foi 33.

A Figura 8 traz o desempenho de cada regra em relação ao número de questões selecionadas para cada grupo de $\hat{\theta}$ estimado. Na maioria dos casos, a regra F tem maior média de itens selecionados para $\hat{\theta}$ maiores que $-0.5$. Em contrapartida, possui melhor desempenho em usuários de $\theta \leq -1.5$. Já a regra MLWI foi melhor para $\hat{\theta}$s entre $-0.5$ e 1.5 e maiores que 2.5.

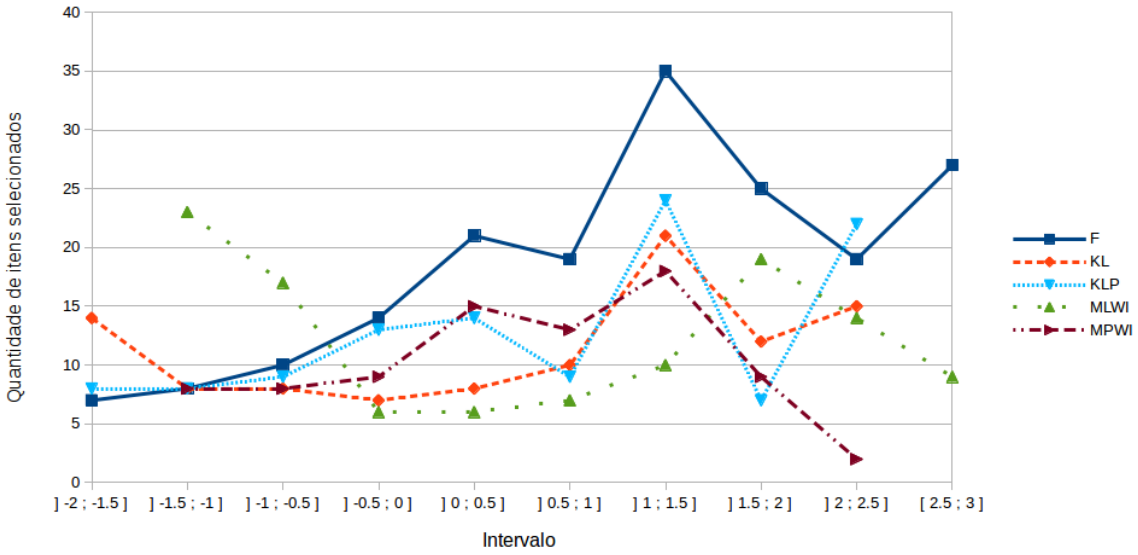



Figura 8 – Média de itens selecionados pelas ISRs para cada intervalo dos escores estimados.

Fonte: Victor Miranda Gonçalves Jatobá, 2017

Os resultados do BIAS e do RMSE podem ser vistos na Tabela 3. Há evidências de que as regras $KL$ e $MPWI$ estão subestimando a habilidade dos respondentes, pois estas possuem BIAS negativo. As regras com a menor raiz do erro quadrático médio, ou seja, as que possuem melhores estimadores são a $F$ e a $KLP$.

Tabela 3 – BIAS e RMSE para cada regra de seleção de itens

| | F | KL | KLP | MLWI | MPWI |
|---|---|---|---|---|---|
| BIAS | 0,030 | -0,002 | 0,001 | 0,067 | -0,028 |
| RMSE | 0,174 | 0,273 | 0,193 | 0,400 | 0,294 |

De forma geral, a regra de seleção com maior destaque é a $KLP$, pois tem o menor BIAS, o segundo menor RMSE e permite reduzir em 46,6% o tamanho da prova, sem

perda significativa da estimativa do escore do respondente, se comparado ao teste completo com 45 questões.

## *5.2 Desempenho das ISRs nos estágios iniciais dos CATs*

A seguir são apresentados o valor do BIAS e RMSE na seleção dos 30 primeiros itens da execução dos CATs.

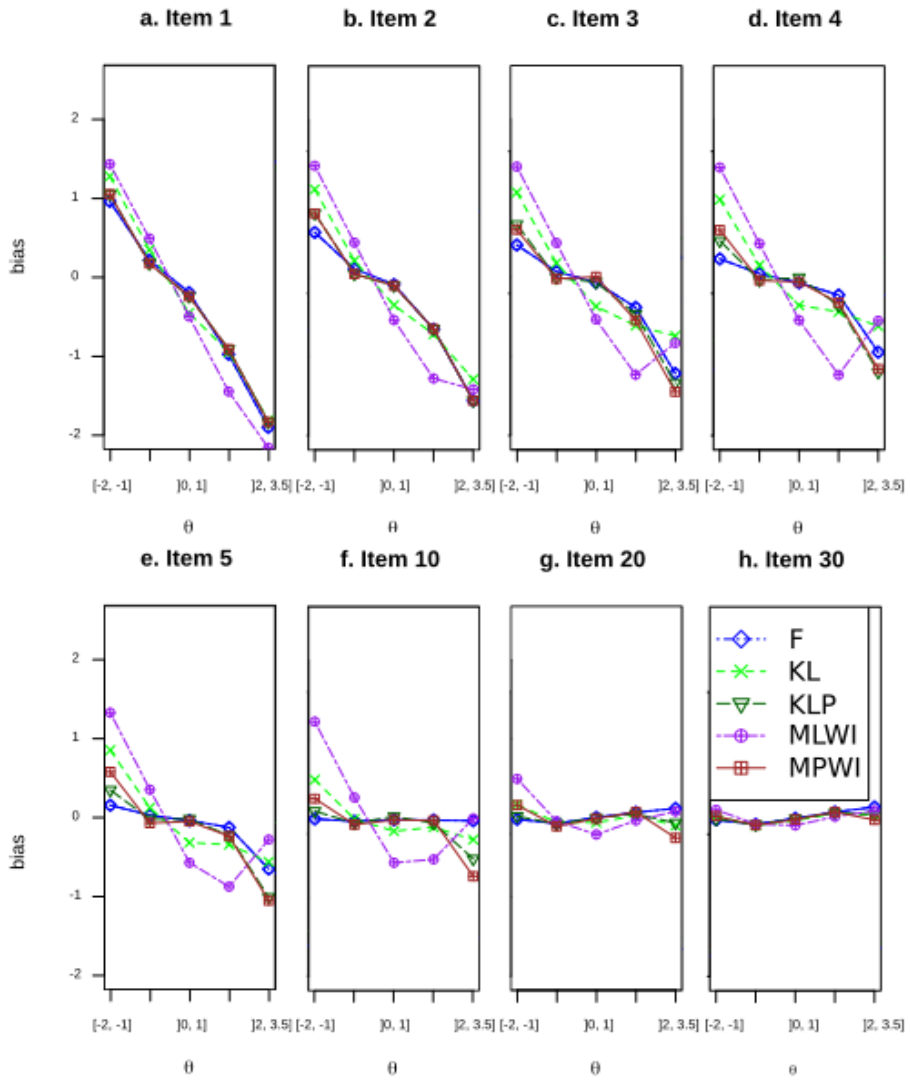
### 5.2.1 BIAS

A Figura 9 resume os resultados do BIAS (Equação 16) realizados nos momentos iniciais do teste para cada uma das regras de seleção de itens. De forma geral, a diferença do desempenho das regras foi ficando cada vez menor, a medida que o teste aumentava em número de questões. A partir do item 30, por exemplo, todas as regras possuem desempenho muito similar.

A maior diferença dos valores do BIAS está contida nos grupos mais extremos dos $\theta$s. Para o extremo negativo ($-2 \leq \theta < -1$) a regra $F$ teve um menor BIAS nos 10 primeiros itens. Já no extremo positivo ($2 < \theta \leq 3.5$), houve bastante variação. Apesar da regra $KL$ ter tido um bom desempenho nos itens 2 e 3, foi a regra $MLWI$ que obteve destaque nesse grupo. Esta regra, a partir do item 4 até o item 10, foi a que esteve mais próxima de zero.

De forma geral, o padrão de convergência do BIAS ao longo da aplicação de mais itens se manteve similar ao trabalho de CHEN et al. (2000). Mas os resultados, principalmente da regra $F$ para extremos negativos, foram divergentes. Isso pode ser explicado pela diferença da natureza dos testes, pela forma como o CAT foi configurado e também pela escolha e configuração dos método de estimação das habilidades.

Figura 9 – Resultados do BIAS advindos da execução dos CATs de cada ISR. Os BIAS foram catalogados por grupos de $\theta$ para os primeiros 30 itens. Os grupos vão de -2 a 3,5 (eixo x).

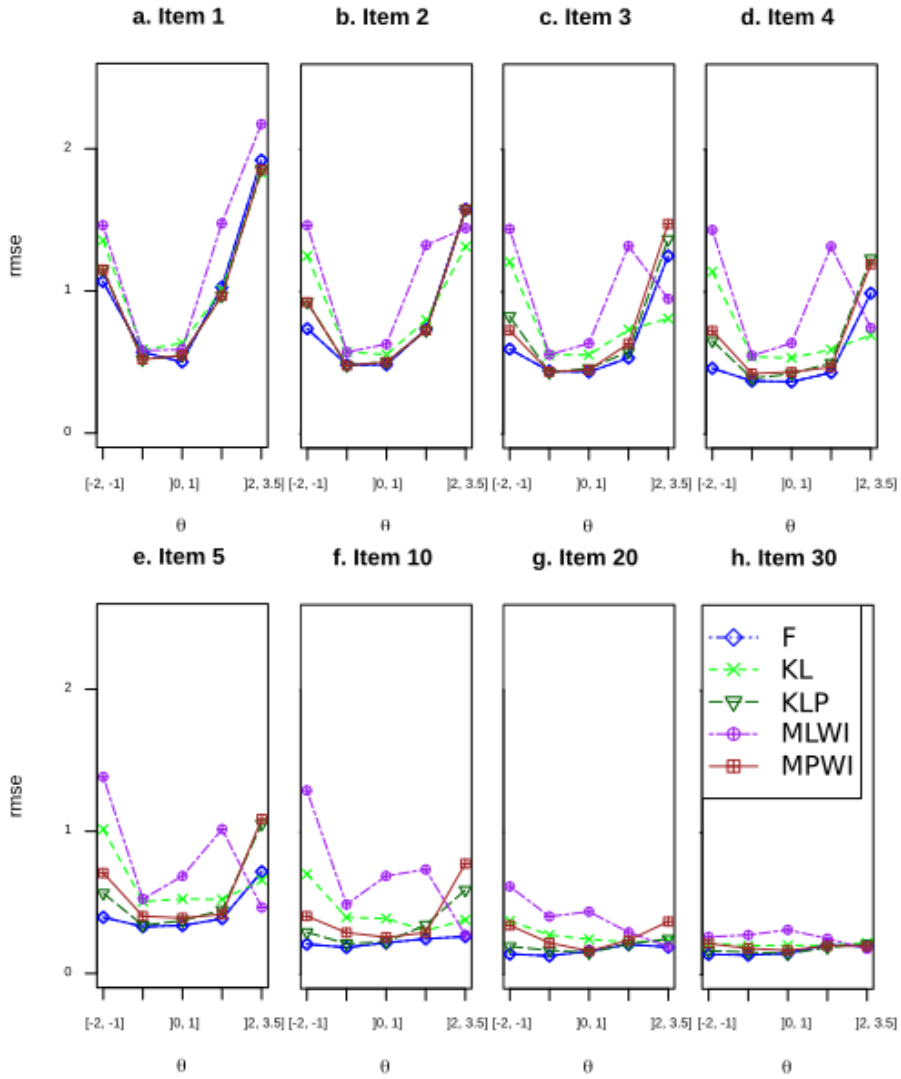



Fonte: Victor Miranda Gonçalves Jatobá, 2017

### 5.2.2 RMSE

A Figura 10 mostra que o comportamento do RMSE das diferentes ISRs, para os $\theta$s extremos, foi similar ao observado para o BIAS. As regras $F$ e $MLWI$ obtiveram melhor RMSE para $\theta$s extremos negativos e positivos, respectivamente. Essas observações seguem as mesmas ressalvas da Subseção 5.2.1.

Figura 10 – Resultados do RMSE advindos da execução dos CATs de cada ISR. Os RMSE foram catalogados por grupos de $\theta$ para os primeiros 30 itens. Os grupos vão de -2 a 3,5 (eixo x).

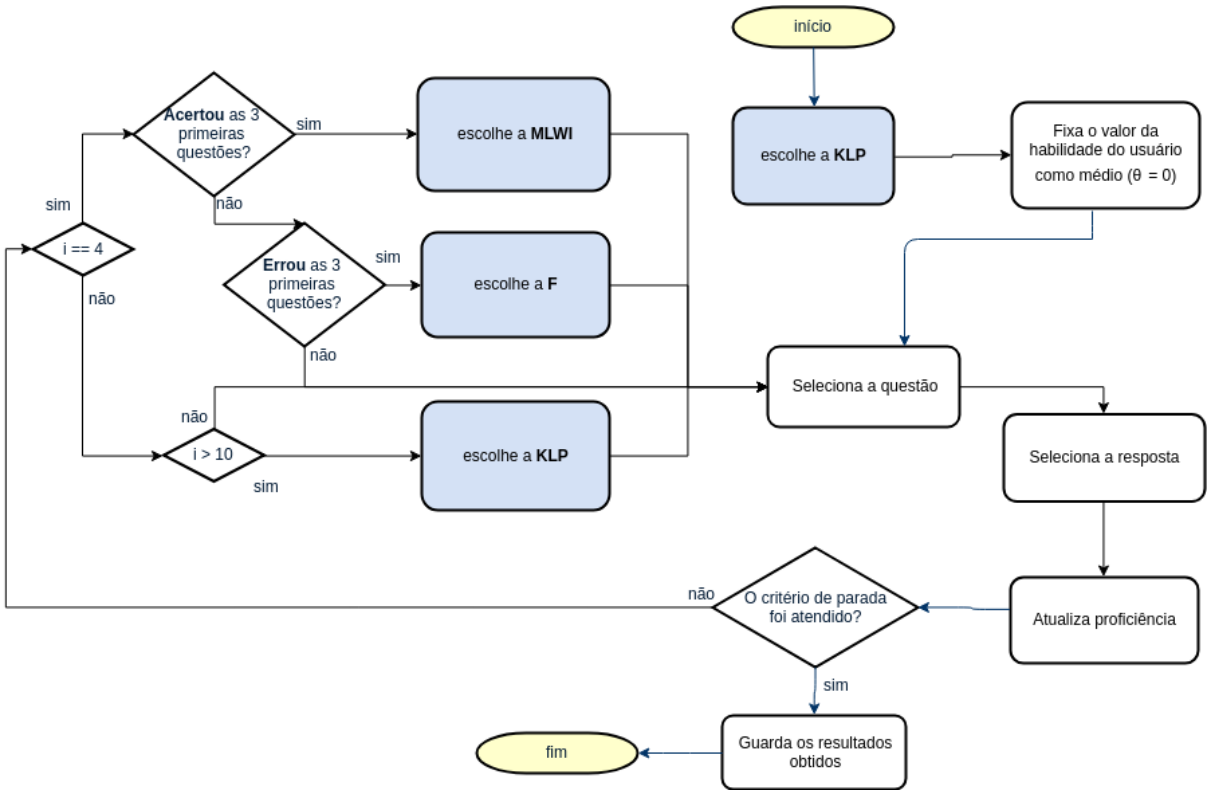



Fonte: Victor Miranda Gonçalves Jatobá, 2017

## *5.3 Configuração e análise do desempenho do ALICAT*

Com os resultados das Subseções 5.1 e 5.2, foi possível construir o ALICAT. Essa abordagem ficou sob a seguinte definição: A ISR de melhor desempenho geral é a $KLP$; a de melhor desempenho em usuários de nível alto ($\theta \geq 2.5$) é a $MLWI$; e a de melhor desempenho para usuários de nível baixo ($\theta \leq -1.5$) é a $F$. Assim, a configuração do ALICAT modelada na Figura 5, pode ser vista na Figura 11.

Figura 11 – Fluxograma contendo a configuração final do ALICAT após a análise das ISRs.




Fonte: Victor Miranda Gonçalves Jatobá, 2017

Com a concepção do ALICAT, este foi executado considerando a prova completa e a amostra dos 4979 respondentes. O ponto de estabilidade (Seção 4.3) identificado, foi de **21** questões. Os resultados da Tabela 4 mostram que todos os intervalos de $\hat{\theta}$ foram contemplados. Além disso, a quantidade de respondentes por grupo, ficou muito próximo dos 500 originalmente retirados dos $\theta$s *verdadeiros*.

Com o ponto de estabilidade definido, o CAT foi então executado novamente, agora considerando o valor fixo de 21 questões para o critério de parada. O valor do BIAS foi de **0,004** e o do RMSE foi **0,190** . Ou seja, muito próximo dos resultados da $KLP$, que obteve melhor desempenho geral.

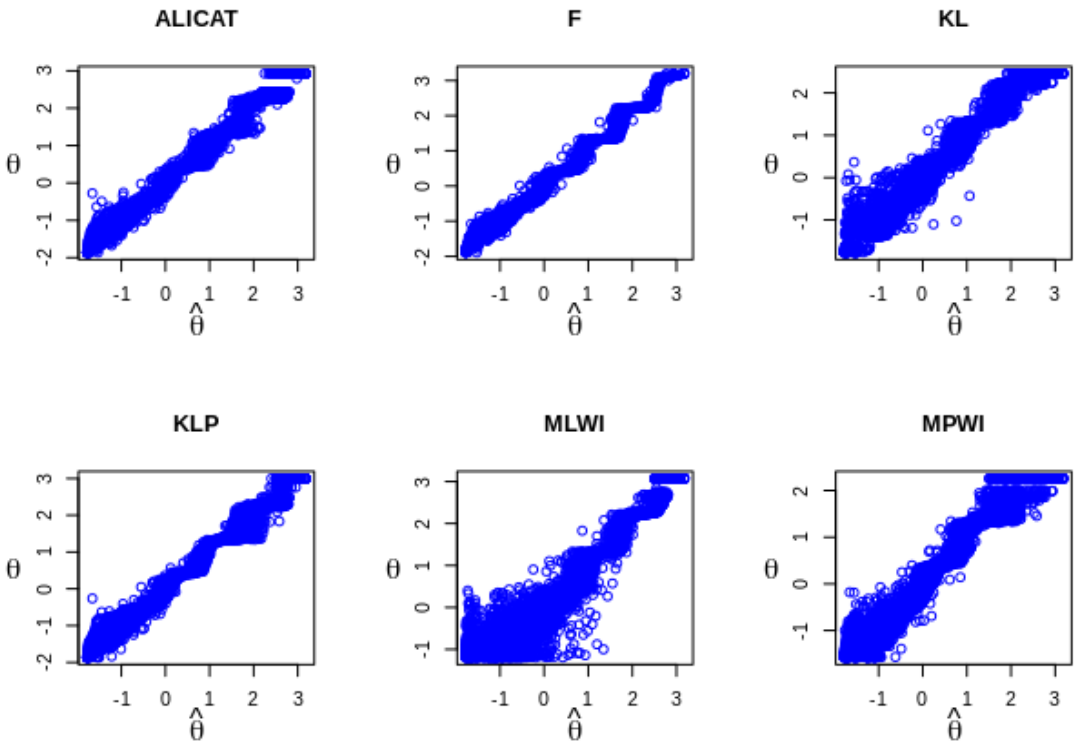
A Figura 12 exibe os resultados dos escores obtidos via CAT versus os *escores verdadeiros* obtidos via prova completa. O ALICAT e o CAT com uso exclusivo da regra $KLP$ tiveram resultados parecidos. Os CATs com as regras $KL$ e $MLWI$ tiveram bastante divergências nas estimativas de escores abaixo de 0 (zero). A diferença maior nas estimativas do $MPWI$ foram para $\theta$s menores que 0 e maiores que 2. A principal diferença entre o ALICAT e a regra $F$, foram nos $\theta$s próximos de 3. Ao mesmo tempo que a $F$ teve um

Tabela 4 – Quantidade de respondentes ($\sigma$) e média de itens selecionados ($\bar{x}$) para cada intervalo de $\hat{\theta}$ na abordagem do ALICAT

| Intervalo dos $\hat{\theta}$ | ALICAT | |
|---|---|---|
| | $\sigma$ | $\bar{x}$ |
| [ -2 ; -1,5 ] | 495 | 7 |
| ] -1,5 ; -1 ] | 471 | 8 |
| ] -1 ; -0,5 ] | 626 | 10 |
| ] -0,5 ; 0 ] | 540 | 11 |
| ] 0 ; 0,5 ] | 507 | 13 |
| ] 0,5 ; 1 ] | 754 | 9 |
| ] 1 ; 1,5 ] | 449 | 18 |
| ] 1,5 ; 2 ] | 100 | 18 |
| ] 2 ; 2,5 ] | 554 | **21** |
| [ 2,5 ; 3 ] | 483 | 10 |

aparente melhor desempenho, ela também foi a que precisou de mais itens. Ao todo foram 35, em comparação aos 21 utilizados na execução do ALICAT.

Figura 12 – Comparação dos escores obtidos via CAT ($\hat{\theta}$) em relação aos escores verdadeiros ($\theta$)




Fonte: Victor Miranda Gonçalves Jatobá, 2017

# 6 Conclusões e trabalhos futuros

As regras $F$ e $KLP$ tiveram bons desempenhos gerais nas estimativas dos $\theta$s considerando o ponto de estabilidade como critério de parada. Contudo a $KLP$ precisou de menos questões (24 contra os 35 da $F$) tornando-se a melhor opção na avaliação geral das ISRs. Ao investigar o comportamento das ISRs nos momentos iniciais dos CATs, a $MLWI$ e a $F$ tiveram os melhores desempenhos para $\theta$s extremos positivos e negativos, respectivamente. Através desses resultados, foi possível configurar e validar o ALICAT. O desempenho desse para o BIAS e RMSE foi próximo do $KLP$, entretanto foram necessários apenas 21 itens. Isso representa uma redução de $54,3\%$ em relação aos 45 itens da prova completa de *Matemática e suas Tecnologias* do ENEM de 2012 sem perda significativa na estimativa das habilidades.

Em síntese o ALICAT conseguiu otimizar ainda mais o processo de seleção de itens, escolhendo dinamicamente a ISR, ao invés da estratégia fixa de uso de apenas uma ISR em todo o teste. Essa melhora no processo de construção de testes adaptativos computadorizados, representa uma redução direta no tempo de resolução da prova. Essa característica engloba vantagens, como: (a) redução de custo pela instituição que está aplicando a prova, pois os usuários passarão menos tempo utilizando os recursos computacionais e os espaços físicos, se estes forem assim aplicados; e (b) diminuição da fadiga e frustração dos respondentes. Estas características podem melhorar a motivação dos respondentes na resolução dos itens, e assim ter uma melhor precisão na estimativa das habilidades pelo CAT. Isso é possível, pois, dentre outros aspectos, os participantes não precisarão responder a itens muito fáceis ou muito difíceis, para seu nível de conhecimento.

No contexto do estudo de caso, o trabalho utilizou apenas dados de uma macroárea do ENEM. Entretanto, o estudo pode ser estendido para as outras áreas do conhecimento ou mesmo para outras provas de natureza similar.

Outra limitação está no uso fixo do maior valor médio dos pontos de estabilidade dentre os 10 grupos de $\hat{\theta}$s como critério de parada no CAT. Isso pode diminuir a precisão na estimativa de respondentes que possuem um maior ponto de estabilidade.

Duas publicações foram produzidas até o momento de depósito dessa dissertação. A primeira foi uma Revisão Sistemática, aceita no XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017), sob o título de: “Testes Adaptativos Computadorizados ba-

seados na Teoria de Resposta ao Item em Sistemas e-learning: Uma revisão sistemática da literatura". A outra, diz respeito aos resultados apresentados na etapa do estudo comparativo entre as ISRs. O artigo recebe o título de: "Regras de Seleção de Itens em Testes Adaptativos Computadorizados: um estudo de caso no ENEM". Ele foi aceito no XXIX SBIE, edição de 2018.

Para trabalhos futuros, sugere-se aplicar o ALICAT em diferentes provas educacionais e utilizar diferentes métodos para a definição do comprimento e a configuração do CAT. Como sugestão podem ser identificados o desvio padrão da média dos pontos de estabilidade para assim aumentar a confiança na escolha do tamanho do teste. Ademais, outros métodos de seleção de itens também podem ser verificados no estudo comparativo das ISRs. Assim será possível validar o impacto nas estimativas dos escores da abordagem do ALICAT em diferentes cenários de testes.

## Referências[1]

AL-A'ALI, M. Implementation of an improved adaptive testing theory. *Journal of Educational Technology & Society*, v. 10, n. 4, 2007. Citado 2 vezes nas páginas 18 e 33.

AL-A'ALI, M. Implementation of an improved adaptive testing theory. *Educational Technology & Society*, v. 10, n. 4, p. 80–94, 2007. Citado 2 vezes nas páginas 18 e 19.

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*, 2000. Citado 2 vezes nas páginas 15 e 21.

ANDRADE, G. G. A metodologia do enem: uma reflexão. *Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB*, n. 33, 2013. Citado na página 34.

BAKER, F. B. *The basics of item response theory*. United States of America: ERIC, 2001. Citado na página 23.

BAKER, F. B.; KIM, S.-H. *Item response theory: Parameter estimation techniques*. Boca Raton, Florida, EUA: CRC Press, 2004. Citado na página 22.

BARRADA, J. R. et al. Incorporating randomness in the fisher information for improving item-exposure control in cats. *British Journal of Mathematical and Statistical Psychology*, Wiley Online Library, v. 61, n. 2, p. 493–513, 2008. Citado 2 vezes nas páginas 16 e 25.

BERNES-LEE, T. *Linked Data - Design Issues*. 2006. Disponível em: ⟨http://www.w3.org/DesignIssues/LinkedData.html⟩. Citado na página 15.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, Addison-Wesley, 1968. Citado 2 vezes nas páginas 22 e 24.

BOCK, R. D.; MISLEVY, R. J. Adaptive eap estimation of ability in a microcomputer environment. *Applied psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 6, n. 4, p. 431–444, 1982. Citado na página 31.

BRASIL. *Instituto Nacional de Educação e Pesquisas Educacionais Anísio Teixeira. Entenda sua nota no ENEM*. 2013. Disponível em: ⟨http://download.inep.gov.br/educacao_basica/enem/guia_participante/2013/guia_do_participante_notas.pdf⟩. Acesso em: 29 mar. 2017. Citado 4 vezes nas páginas 15, 22, 34 e 35.

BUTTERFIELD, M. S. *Comparing item selection methods in computerized adaptive testing using the rating scale model*. Tese (Doutorado), 2016. Citado na página 25.

CHANG, H.-H.; YING, Z. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 3, p. 213–229, 1996. Citado 2 vezes nas páginas 28 e 29.

CHEN, C.-M.; LEE, H.-M.; CHEN, Y.-H. Personalized e-learning system using item response theory. *Computers & Education*, Elsevier, v. 44, n. 3, p. 237–255, 2005. Citado 2 vezes nas páginas 15 e 24.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

CHEN, S.-Y.; ANKENMAN, R. D. Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, Wiley Online Library, v. 41, n. 2, p. 149–174, 2004. Citado 5 vezes nas páginas 16, 25, 26, 30 e 36.

CHEN, S.-Y.; ANKENMANN, R. D.; CHANG, H.-H. A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 24, n. 3, p. 241–255, 2000. Citado 10 vezes nas páginas 16, 24, 25, 26, 27, 28, 29, 30, 36 e 46.

COLLINS, J. A.; GREER, J. E.; HUANG, S. X. Adaptive assessment using granularity hierarchies and bayesian nets. In: SPRINGER. *International Conference on Intelligent Tutoring Systems*. Berlim, Heidelberg, 1996. p. 569–577. Citado na página 20.

COSTA, C. E. S. *Análise da dimensionalidade e modelagem multidimensional pela TRI no ENEM (1998-2008)*. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2015. Citado na página 33.

DOEBLER, A. The problem of bias in person parameter estimation in adaptive testing. *Applied Psychological Measurement*, Sage Publications Sage CA: Los Angeles, CA, v. 36, n. 4, p. 255–270, 2012. Citado 2 vezes nas páginas 32 e 33.

EGGEN, T.; STRAETMANS, G. Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 60, n. 5, p. 713–734, 2000. Citado na página 24.

GALVAO, A. F.; NETO, R. F.; BORGES, C. C. H. Um modelo inteligente para seleç ao de itens em testes adaptativos computadorizados. 2013. Citado na página 22.

GONÇALVES, J. P.; ALUÍSIO, S. M. Teste adaptativo computadorizado multidimensional com propósitos educacionais: Princípios e métodos. *Revista Ensaio: Avaliação e Políticas Públicas em Educação*, v. 23, n. 87, p. 389–414, 2015. Citado na página 33.

GUZMÁN, E.; CONEJO, R. A model for student knowledge diagnosis through adaptive testing. In: SPRINGER. *International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg, 2004. p. 12–21. Citado 3 vezes nas páginas 15, 19 e 20.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. *Fundamentals of item response theory*. Londres, Inglaterra: Sage, 1991. Citado 3 vezes nas páginas 21, 24 e 25.

HANSON, B. A. *IRT command language*. California, EUA: Version, 2002. Citado na página 38.

Heitink, M.; Veldkamp, D. B. P. Computer adaptive assessment for learning in a virtual learning environment. In: *Proceedings of the 18th International Computer Assisted Assessment Conference*. Enschede, Holanda: Springer, 2015. p. 22 – 26. Citado na página 33.

JACOBSON, S. H. Z. *A comparison of early childhood assessments and a standardized measure for program evaluation*. Tese (Doutorado) — Virginia Tech, 1997. Citado 2 vezes nas páginas 18 e 24.

JEONG, H.-Y.; HONG, B.-H. A service component based CAT system with SCORM for advanced learning effects. *Multimedia Tools Appl.*, v. 63, n. 1, p. 217–226, mar. 2013. ISSN 1380-7501. Citado na página 33.

KARINO, C. A.; BARBOSA, M. T. S. *Procedimento de cálculo das notas do ENEM. Nota Técnica.* 2011. Disponível em: ⟨http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_procedimento_de_calculo_das_notas_enem_2.pdf⟩. Acesso em: 15 mai. 2017. Citado na página 34.

KOVATCHEVA, E.; NIKOLOV, R. An adaptive feedback approach for e-learning systems. *IEEE Technology and Engineering Education (ITEE)*, v. 4, n. 1, p. 55–57, 2009. Citado 3 vezes nas páginas 15, 19 e 24.

LENDYUK, T.; RIPPA, S.; SACHENKO, S. Simulation of computer adaptive learning and improved algorithm of pyramidal testing. In: *2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*. Berlin, Alemanha: IEEE, 2013. v. 02, p. 764–769. Citado 2 vezes nas páginas 24 e 33.

LINDEN, W. J. V. D. Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 23, n. 1, p. 21–29, 1999. Citado 2 vezes nas páginas 30 e 32.

LINDEN, W. J. van der. Bayesian item selection criteria for adaptive testing. *Psychometrika*, v. 63, n. 2, p. 201–216, Jun 1998. Citado 4 vezes nas páginas 16, 28, 30 e 36.

LINDEN, W. J. Van der; PASHLEY, P. J. Item selection and ability estimation in adaptive testing. In: *Elements of adaptive testing*. Nova Iorque, EUA: Springer, 2009. p. 3–30. Citado 5 vezes nas páginas 16, 25, 28, 30 e 36.

LÓPEZ-CUADRADO, J. et al. Calibration of an item bank for the assessment of basque language knowledge. *Computers & Education*, Elsevier, v. 55, n. 3, p. 1044–1055, 2010. Citado 2 vezes nas páginas 15 e 20.

LORD, F. M. A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 1, n. 1, p. 95–100, 1977. Citado na página 19.

LORD, F. M. *Applications of item response theory to practical testing problems*. Nova Iorque, EUA: Routledge, 1980. Citado na página 25.

LORD, F. M. Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, Springer, v. 48, n. 2, p. 233–245, 1983. Citado na página 31.

MAGIS, D.; RAÎCHE, G. et al. Random generation of response patterns under computerized adaptive testing with the r package catr. *Journal of Statistical Software*, Foundation for Open Access Statistics, v. 48, n. 8, p. 1–31, 2012. Citado na página 39.

MEIJER, R. R.; NERING, M. L. *Computerized adaptive testing: Overview and introduction*. Illinois, EUA: Sage Publications Sage CA: Thousand Oaks, CA, 1999. Citado 3 vezes nas páginas 31, 32 e 33.

NOVICK, M. R. The axioms and principal results of classical test theory. *Journal of mathematical psychology*, Elsevier, v. 3, n. 1, p. 1–18, 1966. Citado na página 18.

OLEA, J.; PONSODA, V. Tests adaptativos informatizados. *Psicometría*, Editorial Universitas. Madrid, p. 731–783, 1996. Citado 2 vezes nas páginas 16 e 20.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria da resposta ao item: Tri. *Avaliação Psicológica*, Instituto Brasileiro de Avaliação Psicológica. UFRGS, v. 2, n. 2, p. 99–110, 2003. Citado 2 vezes nas páginas 18 e 21.

QUARESMA, E. d. S. *Modelagem para construçao de escalas avaliativas e classificatórias em exames seletivos utilizando teoria da resposta ao item uni e multidimensional*. Tese (Doutorado) — Escola Superior de Agricultura "Luiz de Queiroz, 2014. Citado 2 vezes nas páginas 18 e 21.

RAJAMANI, K.; KATHIRAVAN, V. An adaptive assessment system to compose serial test sheets using item response theory. In: IEEE. *International Conference on Pattern Recognition, Informatics and Mobile Engineering*. Salem, India, 2013. p. 120–124. Citado 3 vezes nas páginas 18, 32 e 33.

RASCH, G. Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*, 1960. Citado na página 22.

RECKASE, M. D. *Multidimensional Item Response Theory*. 1st. ed. Nova Iorque, EUA: Springer Publishing Company, 2009. Citado na página 23.

RECKASE, M. D.; WEISS, D. A procedure for decision making using tailored testing. *New horizons in testing: Latent trait test theory and computerized adaptive testing*, Nova Iorque, EUA, p. 237–255, 1983. Citado na página 20.

RIJN, P. V. et al. Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 26, n. 4, p. 393–411, 2002. Citado na página 25.

RUDNER, L. M. An examination of decision-theory adaptive testing procedures. In: GMAC. *annual meeting of the American Educational Research Association*. Minessota, EUA, 2002. Citado na página 20.

SEGALL, D. O. A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 4, p. 439–460, 2004. Citado na página 15.

SEGALL, D. O. Principles of multidimensional adaptive testing. In: *Elements of adaptive testing*. Nova Iorque, EUA: Springer, 2009. p. 57–75. Citado na página 18.

SIE, H. *Statistical Aspects Of Computerized Adaptive Testing*. Tese (Doutorado) — The Pennsylvania State University, 2014. Citado na página 18.

SPENASSATO, D.; BORNIA, A. C.; TEZZA, R. Computerized adaptive testing: a review of research and technical characteristics. *IEEE Latin America Transactions*, IEEE, v. 13, n. 12, p. 3890–3898, 2015. Citado 3 vezes nas páginas 19, 24 e 32.

SPENASSATO, D. et al. Testes adaptativos computadorizados aplicados em avaliações educacionais. *Revista Brasileira de Informática na Educação*, v. 24, n. 2, 2016. Citado 6 vezes nas páginas 30, 33, 38, 39, 40 e 45.

TEIXEIRA, I. N. de Estudos e P. E. A. *Microdados do ENEM. Brasília: Inep, 2016*. 2016. Disponível em: ⟨http://portal.inep.gov.br/basica-levantamentos-acessar⟩. Acesso em: 29 mar. 2017. Citado na página 38.

THOMPSON, N. A.; WEISS, D. J. A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, Citeseer, v. 16, n. 1, p. 1–9, 2011. Citado 2 vezes nas páginas 21 e 32.

VEERKAMP, W. J.; BERGER, M. P. Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, Sage Publications, v. 22, n. 2, p. 203–226, 1997. Citado 6 vezes nas páginas 16, 26, 27, 28, 30 e 36.

VELDKAMP, B. P.; MATTEUCCI, M. Bayesian computerized adaptive testing. *Ensaio: Avaliação e Políticas Públicas em Educação*, SciELO Brasil, v. 21, n. 78, p. 57–82, 2013. Citado na página 33.

VISPOEL, W. P. Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement*, Wiley Online Library, v. 35, n. 2, p. 155–167, 1998. Citado na página 33.

WAINER, H. et al. *Computerized adaptive testing: A primer*. Nova Iorque, EUA: Routledge, 2000. Citado 4 vezes nas páginas 15, 16, 19 e 20.

WANG, C.; CHANG, H.-H.; HUEBNER, A. Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, Wiley Online Library, v. 48, n. 3, p. 255–273, 2011. Citado 2 vezes nas páginas 16 e 25.

WANG, T.; VISPOEL, W. P. Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, Wiley Online Library, v. 35, n. 2, p. 109–135, 1998. Citado na página 32.

WARM, T. A. Weighted likelihood estimation of ability in item response theory. *Psychometrika*, Springer, v. 54, n. 3, p. 427–450, 1989. Citado na página 31.

WEISS, D. J. Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 6, n. 4, p. 473–492, 1982. Citado na página 19.

WISE, S. L. Overview of practical issues in a cat program. ERIC, 1997. Citado na página 24.

WONG, K. et al. E-learning: Developing a simple web-based intelligent tutoring system using cognitive diagnostic assessment and adaptive testing technology. In: SPRINGER. *International Conference on Hybrid Learning*. Berlin, Heidelberg, 2010. p. 23–34. Citado 2 vezes nas páginas 15 e 20.

ZITNỲ, P. et al. Validity of cognitive ability tests-comparison of computerized adaptive testing with paper and pencil and computer-based forms of administrations. *Studia Psychologica*, Institute of Experimental Psychology, Slovak Academy of Sciences, v. 54, n. 3, p. 181, 2012. Citado na página 33.

ÖZYURT, H. et al. Integrating computerized adaptive testing into UZWEBMAT: Implementation of individualized assessment module in an e-learning system. *Expert Systems with Applications*, v. 39, n. 10, p. 9837 – 9847, 2012. Citado 2 vezes nas páginas 18 e 24.