Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Estatística

Testes Adaptativos Computadorizados Aplicados às Provas do ENEM

Wellington Ferreira de Souza

Ouro Preto - Minas Gerais - Brasil
Junho de 2019

Wellington Ferreira de Souza

Testes Adaptativos Computadorizados Aplicados às Provas do ENEM

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientadora

Prof. Dra. Erica Castilho Rodrigues

Universidade Federal de Ouro Preto – UFOP
Departamento de Estatística – DEEST

Ouro Preto - Minas Gerais - Brasil

Junho de 2019

Ata de Defesa

Ata da sessão pública para julgamento da monografia de Wellington Ferreira de Souza, aluno do curso de Bacharelado em Estatística, do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto.

Aos cinco dias do mês de junho do ano de 2019, às 15h, no Auditório do Departamento de Estatística, reuniu-se a comissão julgadora composta pelos professores Érica Castilho Rodrigues, Eduardo Bearzoti e Marcelo Carlos Ribeiro, para avaliar a monografia do aluno **Wellington Ferreira de Souza**, intitulada ***Testes Adaptativos Computadorizados Aplicados às Provas do ENEM***. A sessão pública foi aberta pela professora Érica Castilho Rodrigues, presidente da comissão julgadora e orientadora da pesquisa, que, após formalidades de praxe, passou a palavra ao aluno para apresentação oral e, em seguida, arguição pelos membros da banca. Terminada a arguição, a comissão reuniu-se em sessão secreta para elaborar o relatório individual de apreciação da monografia e decidiu pela aprovação da mesma com a nota 10 (dez pontos e 0 décimos). Nada havendo mais a tratar, foi encerrada a sessão da qual se lavrou a presente ata que vai assinada pela comissão julgadora.

Ouro Preto, 05 de junho de 2019.

Erica L Rodrigues
Prof.ª Dr.ª Érica Castilho Rodrigues
Departamento de Estatística - DEEST
Universidade Federal de Ouro Preto - UFOP
Presidente

Prof. Dr. Eduardo Bearzoti
Departamento de Estatística - DEEST
Universidade Federal de Ouro Preto - UFOP
Membro da banca examinadora

Prof. Me. Marcelo Carlos Ribeiro
Departamento de Estatística - DEEST
Universidade Federal de Ouro Preto - UFOP
Membro da banca examinadora

Agradecimentos

Agradeço primeiramente a Deus por ter me guiado e dado forças para superar as adversidades. Tenho certeza que diversos fatos em minha vida não foram aleatórios, mas sim sua mão iluminando o meu caminho.

Agradeço e dedico este trabalho inteiramente a Maria de Fatima Ferreira, minha mãe, que sempre me apoiou em meus estudos e acreditou que eu poderia alcançar meus objetivos, por diversas vezes abrindo mão de seus próprios para me auxiliar. Este trabalho é uma vitória sua.

Agradeço a minha namorada Vanessa Grasiele pelos anos de amparo, paciência e companheirismo. Seu apoio em diversos momentos me manteve forte quando estava prestes a desanimar. Sou hoje um homem totalmente diferente, e melhor, graças a você.

Agradeço a Universidade Federal de Ouro Preto pela qualidade do ensino e da estrutura. Além disso, ressalto também a importância das bolsas de assistência estudantil que pude usufruir. Sem estas, não teria condições de permanecer na Universidade e me dedicar à vida acadêmica.

Agradeço a todos os professores do Departamento de Estatística que contribuíram para a minha formação, em especial, a Professora Erica Castilho, minha orientadora, que acreditou no meu potencial e, o Professor Eduardo Bearzoti que durante sua permanência como presidente do colegiado não mediu esforços para melhorar o curso e tornar a relação entre alunos e colegiado mais próxima e amigável, trazendo assim, uma nova perspectiva de curso que se refletiu consequentemente no conceito 5 atribuído na avaliação do MEC.

Por fim, agradeço a todos que fizeram parte e me auxiliaram nesta jornada, desde os amigos que optaram por escolher outra carreira à aqueles que por força do destino tiveram que abandonar o sonho acadêmico. Saliento a importância do coleguismo de alguns que já formaram, da turma 2015.2 que me acolheu e de todos aqueles que pude por meio do curso e da Empresa Júnior conviver, em especial, aos amigos Thiago Atanazio e Pedro Augusto Alves Viana pelas boas conversas e risadas proporcionadas.

Experimenta nascer preto e pobre na comunidade
Você vai ver como são diferentes as oportunidades
E nem venha me dizer que isso é vitimismo
Não bota a culpa em mim pra encobrir o seu ra-cis-mo!
Existe muita coisa que não te disseram na escola!
Cota não é esmola!
Cota não é esmola!
Cota não é esmola!
Eu disse: Cota não é esmola!
Cota não é esmola!
Cota não é esmola!
Cota não é esmola!

Bia Ferreira

Testes Adaptativos Computadorizados Aplicados às Provas do ENEM

Autor: Wellington Ferreira de Souza
Orientadora: Prof. Dra. Erica Castilho Rodrigues

Resumo

O ENEM, Exame Nacional do Ensino Médio, teve início em 1998 e constitui-se hoje como o principal meio de entrada para instituições de ensino superior públicas e programas do governo como o Prouni e o Fies. A partir de 2009, as notas, denominadas habilidades, passaram a ser estimadas via Teoria de Resposta ao Item (TRI). O próximo passo para o aprimoramento dessas provas é a utilização de testes adaptativos computadorizados. Estes adaptam o teste de acordo com a habilidade do respondente. Porém, antes da aplicação desta metodologia vários fatores devem ser analisados, como a regra de seleção do próximo item e o método de estimação da habilidade. Neste trabalho geramos um banco de itens para simulação de testes adaptativos a partir da calibração dos itens das 4 provas mais recentes do ENEM. Com este banco fizemos a simulação de 30 cenários diferentes e verificamos que a modificação do método de seleção não provoca alterações drásticas nos resultados. Estes são mais sensíveis ao método de estimação escolhido. Além disso, os melhores resultados foram obtidos com os métodos de estimação BM (Bayes Modal) e EAP (Expected a Posteriori).

Palavras-chave: Teste Adaptativo Computadorizado, ENEM, Teoria de Resposta ao Item, Avaliação Educacional, Avaliações em Larga Escala.

Computerized Adaptive Testing Applied to the ENEM Exam

Author: Wellington Ferreira de Souza
Advisor: Prof. Dra. Erica Castilho Rodrigues

Abstract

ENEM, National High School Examination, began in 1998 and is today the main means of entry for public higher education institutions and government programs such as Prouni and Fies. As of 2009, the grades, called skills, began to be estimated by Item Response Theory (IRT). The next step in improving these tests is the use of computerized adaptive tests. These tests adapt according to the respondent's ability. However, before applying this methodology several factors should be analyzed, such as the rule of selection of the next item and the method of estimation of the skill. In this work, we generated a bank of items for the simulation of adaptive tests based on the calibration of the items of the 4 most recent ENEM tests. With this bank we simulated 30 different scenarios and verified that the modification of the selection method does not cause drastic changes in the results. These are more sensitive to the chosen method of estimation. In addition, the best results were obtained with BM (Bayes Modal) and EAP (Expected a Posteriori) methods.

Keywords: Computerized Adaptive Testing, ENEM, Item Response Theory, Educational Evaluation, Large Scale Evaluation.

Lista de figuras

Lista de tabelas

Sumário

1 Introdução

O ENEM, Exame Nacional do Ensino Médio, é uma avaliação anual desenvolvida pelo Inep, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Tem entre seus objetivos permitir o acesso à educação superior, a programas governamentais de financiamento e viabilizar o desenvolvimento de estudos e indicadores sobre a educação brasileira. Teve inicio em 1998 e hoje é a principal avaliação em larga escala do país, sendo sua nota utilizada no Sisu, Sistema de Seleção Unificada. Esta é uma plataforma online do Ministério da Educação, MEC, que seleciona estudantes para instituições públicas de ensino superior. Atualmente, o Sisu é a principal forma de ingresso nos cursos superiores, fato este que justifica os mais de 5.5 milhões de inscritos na edição de 2018 do ENEM.

Sendo uma prova de interesse nacional, é evidente que a metodologia para atribuição de notas compõe parte fundamental no desenvolvimento da avaliação. A Teoria de Resposta ao Item (TRI) começou a ser utilizada à partir de 2009 para a estimação das notas dos candidatos. Este método tem como foco o item e não a prova como um todo, e apresenta diversas vantagens em relação à teoria clássica dos testes, que se baseia na quantidade de respostas corretas. Dentre as principais vantagens desta metodologia estão a comparação de populações e diminuição do número de empates. Aspectos estes essenciais em uma prova de aplicação a nível nacional.

Como mencionado, a Teoria de Resposta ao Item apresenta diversas vantagens em relação à metodologia tradicional de atribuição de notas. Porém, ainda assim apresenta alguns inconvenientes:

- o resultado do teste não é apresentado ao candidato imediatamente;
- para ter estimativas precisas para todos os níveis de habilidade é necessário que o teste contenha itens cujas curvas de informação abranjam toda a escala de habilidade, acarretando em testes longos;
- os indivíduos acabam sendo obrigados muitas vezes a responder itens que não correspondem ao seu nível de habilidade; por exemplo, um indivíduo cuja habilidade é

alta e tem que responder itens fáceis, e vice-versa;

- testes muito longos podem provocar fadiga e influenciar as respostas.

Com a evolução tecnológica, uma possibilidade para resolução destes problemas é a abordagem adaptativa. Nesta, a seleção dos itens é feita de maneira dinâmica e se adéqua a realidade e conhecimentos de cada individuo. Ao invés de dar a todos os candidatos o mesmo teste, a seleção de itens se adapta ao nível de habilidade de cada um deles. O indivíduo realiza o teste a partir de algum dispositivo, como um computador ou tablet, e a habilidade é estimada a partir de cada resposta, sendo o próximo item selecionado de acordo com a habilidade provisória. O teste termina quando o critério de parada é satisfeito. Esta metodologia resolve os problemas citados, além de gerar estimativas mais precisas de habilidade. Além disso, os custos com aplicação, armazenamento e correção dos testes são reduzidos.

Os testes adaptativos computadorizados (CATs - Computerized Adaptive Testing) têm sido cada vez mais utilizados ao redor do mundo, inclusive em avaliações renomadas como o TOEFL (Test of English as a Foreign Language). No Brasil, no entanto, a discussão sobre o tema ainda é incipiente, exigindo estudos sobre o assunto. Diversas questões devem ser levantadas e analisadas antes da aplicação desta metodologia:

- o método de seleção do próximo item;
- o método de estimação da habilidade;
- o critério de parada;
- o balanceamento de conteúdo;
- a taxa de exposição do item.

2 Objetivos

Neste trabalho objetiva-se demonstrar as vantagens da abordagem adaptativa em provas de larga escala. A partir de simulações iremos aplicar a abordagem adaptativa em diferentes cenários e verificar os resultados. Espera-se que, ao utilizar esta metodologia, seja possível estimar as habilidades dos alunos com um número menor de itens e com maior precisão e, ainda, determinar os melhores cenários do teste. O ENEM foi escolhido, pois, como mencionado constitui-se como a principal prova em larga escala do país.

3 Referencial Teórico

Neste capítulo apresentamos os conceitos principais desta pesquisa tendo como base outros trabalhos já publicados na literatura e que fundamentaram as análises presentes neste estudo.

3.1 Conceitos centrais

O desenvolvimento de escalas apropriadas para medir características de indivíduos que não podem ser medidas diretamente, as quais são comumente denominadas de traço latente, tem tomado a atenção de pesquisadores das mais diferentes áreas do conhecimento (ANDRADE; ANJOS, 2012).

Existem diversas formas de avaliar o nível de proficiência dos indivíduos. Um modelo bastante utilizado atualmente é o da Teoria de Resposta ao Item. Esta teoria é composta por modelos matemáticos que procuram estabelecer a probabilidade de um respondente qualquer acertar uma determinada questão, dadas as características do item e a habilidade do avaliado. Esse é o modelo adotado na prova objetiva do Exame Nacional do Ensino Médio (ENEM) para calcular o desempenho dos estudantes (JATOBÁ et al., 2018). No cálculo da nota, o modelo matemático da TRI considera a coerência das respostas corretas do participante. Espera-se que participantes que acertaram as questões difíceis devam também acertar as questões fáceis, pois, entende-se que a aquisição do conhecimento ocorre de forma cumulativa, de modo que habilidades mais complexas requerem o domínio de habilidades mais simples (ENEM, 2012).

Salienta-se que a TRI é muito comum no contexto educacional, no entanto, esta define um conjunto de modelos para estimação de variáveis latentes, e este conceito é extremamente amplo, não se restringindo somente ao campo da educação. A TRI é muito utilizada na psicologia e pode ser útil em diversas outras áreas. (ALEXANDRE et al., 2002), por exemplo, apresentam uma aplicação da TRI para avaliação da gestão pela qualidade total de empresas.

Atualmente, em varias áreas do conhecimento, particularmente em avaliação educacional, vem crescendo o interesse na aplicação de técnicas derivadas da Teoria de Resposta ao Item (ANDRADE; TAVARES; VALLE, 2000). Dentre estas técnicas destacam-se os testes adaptativos computadorizados.

A ideia fundamental dos testes adaptativos é ajustar os itens de um teste ao nível de habilidade individual de cada participante. Em outras palavras, é apresentada uma sequência de questões cuja dificuldade mais se aproxima da habilidade de cada estudante, de maneira que um teste não será o mesmo para todos os indivíduos, possibilitando uma medida mais fiel a respeito da competência de cada um, bem como a economia de tempo de aplicação dos testes (OLIVEIRA; ALUÍSIO; PÍTON, 2004). Isso também garante que os alunos fiquem mais motivados durante o teste, visto que, não serão expostos a itens difíceis demais nem fáceis demais para seu nível de habilidade.

3.2 Abordagens sobre o tema

Em (SASSI, 2012) o autor apresenta os conceitos fundamentais relacionados aos testes adaptativos computadorizados. Nesta dissertação se utiliza simulação e abordagem bayesiana para estimação do parâmetro de interesse, além de apresentar os principais algoritmos de seleção do próximo item. Para comparação dos resultados, usou-se o erro quadrático médio e o vício. Também é feita a análise do tempo entre a resposta e a apresentação de um novo item. O autor, além de modificar o critério de seleção do próximo item, varia o tamanho do teste e do banco de itens. Dentre suas conclusões, afirma que o critério da máxima informação de Fisher é um dos mais rápidos e com menores valores de erro quadrático médio.

Maria Eugénia Ferrão e Paula Prata em (FERRÃO; PRATA, 2014) realizaram testes adaptativos tendo como base uma amostra de 300 alunos que responderam itens avaliando as habilidades em matemática do currículo básico de Portugal. Foi feita a variação da taxa de exposição do item e do comprimento máximo do teste. Para avaliação dos resultados examinou-se o viés absoluto, RMSE e correlação entre a habilidade estimada e a real. Além disso, foi feita comparação das distribuições de frequência com a estatística qui-quadrado. Todas as simulações foram realizadas utilizando o software SimuMCAT e concluíram que a abordagem adaptativa permite uma redução de 36% no tamanho original do teste, sem prejuízo da precisão dos resultados.

Em (SPENASSATO et al., 2016) os autores objetivaram apresentar as vantagens do uso

dos testes adaptativos utilizando os dados da prova de matemática do ENEM 2012. Para estimação dos parâmetros e escores foi utilizado o software Bilog. Para as simulações adaptativas empregou-se o pacote *catR*, (MAGIS; RAÎCHE, 2012; MAGIS; BARRADA, 2017), do software R, (R Core Team, 2018). Neste trabalho, os autores optaram pela chamada simulação *post-hoc*, onde a resposta do individuo não é simulada, mas sim coletada do vetor de respostas real utilizado para estimação dos parâmetros. Foi empregado somente um design com método de estimação da habilidade EAP (expected a posteriori) e método de seleção do próximo item MFI (Máxima Informação de Fisher). Com os resultados obtidos constataram que o teste de Matemática do ENEM 2012, se aplicado na forma adaptativa (CAT), poderia ser reduzido em pelo menos 26.6% sem perda significativa de precisão, se comparado ao teste completo com 45 itens.

(JATOBÁ et al., 2018) determinaram o impacto na estimativa dos escores dos respondentes no uso de diferentes regras de seleção utilizando os itens da prova de matemática do ENEM 2012. Este trabalho procura complementar as análises presentes em (SPENASSATO et al., 2016), e utiliza os valores de parâmetros obtidos neste artigo citado acima. Concluem que usando a regra KLP (Kullback-Leibler Information with a Posterior Distribution) é possível reduzir em 46.6% o tamanho da prova sem perda significativa da precisão na estimativa dos escores dos respondentes.

4 Materiais e métodos

Antes de falarmos sobre testes adaptativos computadorizados devemos apresentar os conceitos relativos a teoria de resposta ao item, visto que, ela é a base para todo o desenvolvimento do teste adaptativo.

4.1 Teoria de resposta ao item - TRI

Segundo Dalton *et al.*, (ANDRADE; TAVARES; VALLE, 2000), a TRI é um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e da habilidade do respondente. Essa relação é sempre expressa de tal forma que quanto maior a habilidade, maior a probabilidade de acerto no item.

Existem diversos modelos de TRI, que dependem de fatores como a natureza do item (dicotômico ou não dicotômico), do número de populações envolvidas e da quantidade de traços latentes a serem estimados (unidimensional ou multidimensional). Os modelos unidimensionais para itens dicotômicos são os mais utilizados, sendo que estes se diferenciam pela quantidade de parâmetros exigidos, que podem ser 1, 2, 3 ou 4.

O modelo da TRI utilizado atualmente no ENEM é o modelo logístico de 3 parâmetros. Este modela a probabilidade do indivíduo responder corretamente cada questão de acordo com 3 parâmetros do item e o nível de habilidade do individuo. Este modelo é definido como:

$$P(U_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \tag{4.1}$$

em que,

- $P(U_{ij} = 1|\theta_j, a_i, b_i, c_i)$ é a probabilidade do indivíduo j com habilidade θ_j acertar o item i;

- a_i é o parâmetro de discriminação do item i, com valor proporcional à inclinação da Curva Característica do Item no ponto b_i;
- b_i é o parâmetro de dificuldade do item i, medido na mesma escala da habilidade;
- c_i é o parâmetro do item que representa a probabilidade de indivíduos com baixa habilidade responderem corretamente o item i, denominado parâmetro de acerto casual.

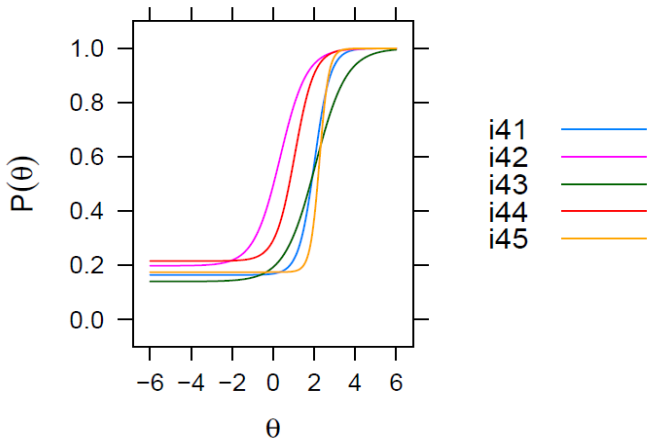
Na Figura 1 apresentamos, para exemplificar o modelo logístico de 3 parâmetros, as curvas características dos itens 41, 42, 43, 44 e 45 da prova de matemática do ENEM 2014. Esta é a representação do item em relação aos seus parâmetros e demonstra a evolução da probabilidade de acerto de acordo com a habilidade do respondente.

Deve-se constatar que o ponto de inflexão das curvas características ocorre no valor do parâmetro de dificuldade, logo, se pode concluir que, por exemplo, o item 42 apresenta parâmetro de dificuldade menor que o item 43, sendo os valores respectivamente 0.346 e 2.062; portanto, o item 42 exige um nível de habilidade menor que o item 43.

Com relação ao parâmetro c_i, este representa a probabilidade de um indivíduo que não possui os conhecimentos avaliados no item, responde-lo corretamente por acaso, ou popularmente "chute". Na curva característica este parâmetro é representado pelo menor valor de probabilidade de acerto assumido na curva. Podemos verificar na Figura 1 por exemplo, que dentre os 5 itens apresentados, o item 43 apresenta o menor valor com relação ao parâmetro de acerto casual, sendo este igual a 0.14, e o item 44 apresenta o maior valor, sendo igual a 0.215.

Por fim, o parâmetro a_i representa o quanto cada item consegue discriminar os indivíduos que possuem as habilidades avaliadas na questão, dos que não possuem. Seu valor é proporcional à derivada da tangente da curva característica no ponto de inflexão. Salienta-se que valores de discriminação negativos não são esperados, visto que, estes representariam que a probabilidade de acerto diminui com o aumento da habilidade, sendo incoerente com a lógica de aprendizado. Na curva característica este parâmetro se manifesta pela inclinação, logo, podemos constatar na Figura 1, que o item 45 apresenta a curva mais "íngreme", e portanto possui o maior valor do parâmetro de discriminação, sendo este 4.47.

Figura 1: Curvas características



4.1.1 Função de informação do item

A função de informação do item permite analisar o quanto de informação o item traz para os diversos níveis de habilidade, sendo definida como:

$$I_i(\theta) = \frac{[\frac{d}{d\theta} P_i(\theta)]^2}{P_i(\theta) Q_i(\theta)} \tag{4.2}$$

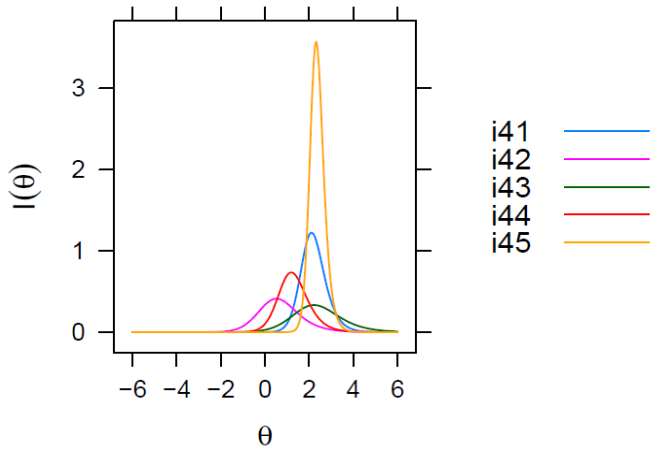
em que,

- $I_i(\theta)$ é a informação fornecida pelo item i no nível de habilidade θ;
- $P_i(\theta) = P(U_{ij} = 1|\theta_j, a_i, b_i, c_i)$;
- $Q_i(\theta) = 1 - P_i(\theta)$.

A quantidade de informação de um item relaciona-se diretamente com os valores dos parâmetros. Um item irá apresentar mais informação para níveis de habilidade θ próximos do valor do parâmetro de dificuldade, b_i, e também quanto maior for o parâmetro de discriminação, a_i, e menor o parâmetro de acerto casual, c_i.

Na Figura 2 apresentamos as curvas de informação dos itens 41, 42, 43, 44 e 45 da prova de matemática do ENEM 2014. Podemos perceber que o item 45 alcança valores de informação mais elevados. Este fato pode ser explicado pela análise da Figura 1, onde a curva característica do item 45 possui inclinação, parâmetro de discriminação, maior que as demais.

Figura 2: Curvas de informação



4.1.2 Função de informação do teste

A função de informação do teste corresponde a soma das informações fornecidas por cada item. Permite analisar a quantidade de informação que a prova apresenta para cada nível de habilidade. A função de informação do teste é definida como:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta) \tag{4.3}$$

em que,

- $I(\theta)$ é a informação fornecida pelo teste no nível de habilidade θ;
- $I_i(\theta)$ é a informação fornecida pelo item i no nível de habilidade θ.

4.1.3 Estimação dos parâmetros usando o algoritmo EM (Expectation-Maximization)

Sejam θ_j a habilidade do indivíduo j e u_{ji} a variável aleatória que representa a resposta binária do indivíduo j ao item i, sendo esta 1 em caso de resposta correta e 0 caso contrario. Sejam $u_{j.} = (u_{j1}, u_{j2}, ..., u_{jI})$ o vetor de respostas do indivíduo j e $u_{..} = (u_{1.}, u_{2.}, ..., u_{n.})$ o conjunto das respostas de todos os n indivíduos. Ainda, $\theta = (\theta_1, ..., \theta_n)$ representa o vetor de habilidades dos n indivíduos.

Segundo Dalton *et al.*, (ANDRADE; TAVARES; VALLE, 2000), o algoritmo EM é um processo iterativo para determinação de estimativas de máxima verossimilhança de parâmetros de modelos de probabilidade na presença de variáveis aleatórias não observadas.

Cada iteração deste processo é feita em dois passos: Esperança (E) e Maximização (M). Denominando ζ como o conjunto de parâmetros dos itens, ou seja, $\zeta_i = (a_i, b_i, c_i)$ corresponde ao vetor de parâmetros do item i, podemos afirmar que no caso da TRI o objetivo do algoritmo EM é obter estimativas de ζ na presença de variáveis não observadas θ. Neste caso, $u_{..}$ representa o vetor de dados incompletos e $(u_{..}, \theta)$ representa o vetor de dados completos. Seja $f(u_{..}, \theta|\zeta)$ a densidade conjunta dos dados completos. Se $\hat{\zeta}^{(t)}$ é uma estimativa de ζ na iteração t, então os passos EM para obter $\hat{\zeta}^{(t+1)}$ são:

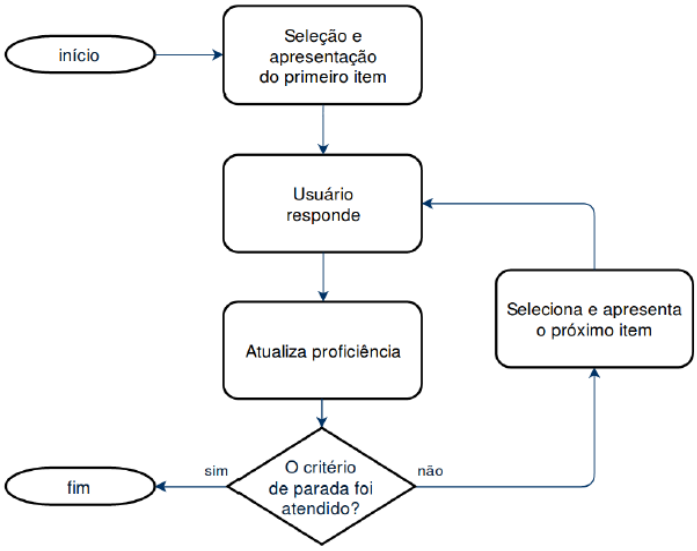
- Passo E: Calcular $E[logf(u_{..}, \theta|\zeta)|u_{..}, \hat{\zeta}^{(t)}]$;
- Passo M: Obter $\hat{\zeta}^{(t+1)}$ que maximiza a função do passo E.

Estes passos compõem cada iteração do algoritmo EM, as quais serão repetidas até que algum critério de parada seja alcançado.

4.2 Testes adaptativos computadorizados

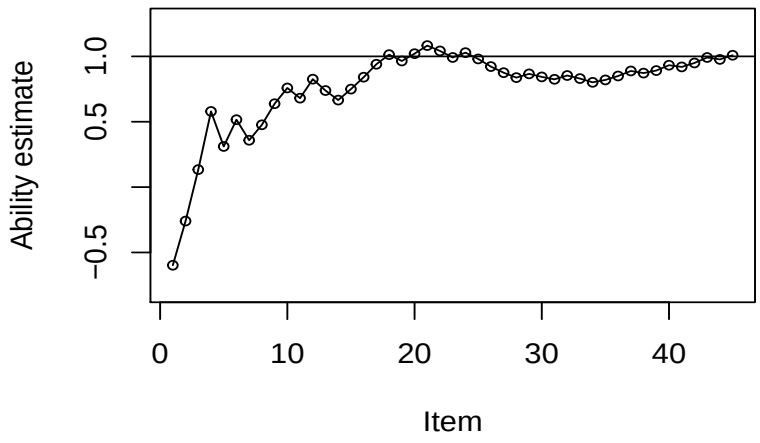
Como já mencionado, os testes adaptativos computadorizados podem ser considerados um aprimoramento da teoria de resposta ao item. Estes adaptam o teste às habilidades de cada respondente, permitindo uma prova mais individualizada. Na Figura 3 apresentamos um fluxograma do processo CAT, retirado de (JATOBÁ et al., 2018). Neste fluxograma identificamos que o teste adaptativo inicia com a seleção de um item para o usuário e após sua resposta, a habilidade provisoria é estimada. Caso o critério de parada seja atendido, o teste encerra; caso não, é selecionado um novo item para o usuário de acordo com a habilidade provisoria atual.

Figura 3: Fluxograma do processo CAT



Na Figura 4 é apresentada a evolução da estimativa da habilidade de um indivíduo com θ real igual a 1, submetido a um teste adaptativo computadorizado (CAT) com 45 itens. À medida que o respondente acerta ou erra os itens, a habilidade é reestimada e um novo item adequado para a habilidade provisoria é aplicado. Com o progresso do teste a habilidade estimada tende a se aproximar cada vez mais da real.

Figura 4: CAT simulado



Pelos conceitos apresentados é nítido que, para definição de um design para um teste adaptativo computadorizado, 3 elementos exercem papel fundamental: o método de sele-

ção do próximo item, o método de estimação da habilidade e o critério de parada. Nas subseções seguintes estes elementos serão melhor detalhados.

4.2.1 Métodos de seleção do próximo item

Os métodos de seleção do próximo item determinam qual item será administrado ao respondente, dada a lista de itens administrados anteriormente, os itens ainda disponíveis e a estimativa da habilidade atual, com vários critérios possíveis. A seguir são apresentados todos os métodos de seleção utilizados para o desenvolvimento das simulações presentes neste trabalho.

1. O método MFI (Máxima Informação de Fisher) seleciona como próximo item aquele que maximiza a função de informação do item com a habilidade estimada atual. Este método tende a ser extremamente rápido sendo um dos mais utilizados em CAT.

2. O método bOpt (Procedimento de Urry ou Critério de Owen) consiste em selecionar o item cujo parâmetro de dificuldade está mais próximo da estimativa da habilidade atual. Suponhamos que temos j-1 itens respondidos e desejamos selecionar o j-ésimo item, estando este presente no subconjunto de itens administráveis do banco (B). O método bOpt seleciona o item por:

$$j = \underset{i}{\operatorname{argmin}}\{|b_i - \hat{\theta}||i \in B\} \tag{4.4}$$

3. O método thOpt consiste em selecionar o item cujo valor ótimo de θ, aquele para o qual a informação do item é máxima, está mais próximo da estimativa de habilidade provisoria atual.

$$j = \underset{i}{\operatorname{argmin}}\{|\hat{\theta} - \theta_i^{max}||i \in B\} \tag{4.5}$$

4. Com o método *progressive* o item selecionado é aquele para o qual a soma ponderada de um componente aleatório com a informação do item é mais alta. No início do teste, quando o erro com relação à estimativa da habilidade é alto, o peso do componente aleatório é máximo e o peso da informação do teste é mínimo. À medida que o número de itens administrados aumenta, o peso do componente aleatório diminui e o peso da informação do teste aumenta. O método progressivo pode ser descrito da seguinte forma:

$$j = \underset{i}{\operatorname{argmax}}\{(1 - W)\, R_i + W\, I_i(\hat{\theta})|i \in B\}, \tag{4.6}$$

onde R_i é um valor aleatório dentro do intervalo [0,$max_{i \in B} I_i(\hat{\theta})$] e $I_i(\hat{\theta})$ e a função de informação do item i para a habilidade provisoria estimada $\hat{\theta}$.

Para CATs de tamanho fixo o valor de W relaciona-se com a quantidade de itens administrados (variando de 1 até Q) da seguinte forma:

$$W = \begin{cases} 0 & se\, \mathrm{q} = 1, \\ \frac{\sum_{f=1}^{q} (f-1)^t}{\sum_{f=1}^{Q} (f-1)^t} & se\, \mathrm{q} \neq 1 \end{cases} \tag{4.7}$$

O parâmetro t representa a velocidade com a qual o peso do componente aleatório é reduzido e, portanto, a velocidade com a qual a importância da informação do item aumenta. Valores mais altos implicam uma maior relevância do componente aleatório na seleção do item.

5. O método KLP (Kullback-Leibler divergency weighted by the posterior distribution) baseia-se na informação de Kullback-Leibler, que avalia a capacidade de discriminação do item entre quaisquer pares possíveis de níveis de habilidade, sendo portanto, uma medida de informação global.

$$j = \underset{i}{\operatorname{argmin}} \left\{ \int_{-\infty}^{+\infty} KL_i(\theta||\hat{\theta})\, f(\theta)\, L(\theta)\, d\theta | i \in B \right\}, \tag{4.8}$$

onde $f(\theta)$ é a distribuição a priori para a habilidade, $L(\theta)$ é a verossimilhança e $KL_i(\theta||\hat{\theta})$ corresponde a informação de Kullback-Leibler, sendo esta calculada como segue:

$$KL_i(\theta||\hat{\theta}) = E\left[log\, \frac{L(\hat{\theta}|X_i)}{L(\theta|X_i)} \right] = \sum_{k=0}^{g_i} P_{ik}(\hat{\theta})\, log\left[\frac{P_{ik}(\hat{\theta})}{P_{ik}(\theta)} \right], \tag{4.9}$$

onde $L(\theta|X_i)$ representa o termo de contribuição do item i para a verossimilhança total $L(\theta)$.

6. Com o critério KL a distribuição a priori não é considerada, sendo a seleção do próximo item feita por:

$$j = \underset{i}{\operatorname{argmin}} \left\{ \int_{-\infty}^{+\infty} KL_i(\theta||\hat{\theta})\, L(\theta)\, d\theta | i \in B \right\} \tag{4.10}$$

4.2.2 Métodos de estimação da habilidade

O método de estimação da habilidade a ser utilizado representa parte fundamental do processo CAT. Existem vários métodos para estimar o nível de habilidade de um indivíduo, dados os parâmetros dos itens e o padrão de resposta correspondente. Nesta

subseção apresentamos os métodos utilizados neste estudo.

1. O método ML (maximum likelihood) define como valor da habilidade aquele que maximiza a função de verossimilhança $L(\theta)$:

$$L(\theta) = \prod_{j=1}^{J} P_j(\theta)^{X_j} \left(1 - P_j(\theta)\right)^{1-X_j}, \tag{4.11}$$

   ou mais frequentemente utilizado, o logaritmo da função de verossimilhança $log\, L(\theta)$:

$$log\, L(\theta) = \sum_{j=1}^{J} \left[X_j\, log\, P_j(\theta) + (1 - X_j)\, log\, (1 - P_j(\theta))\right] \tag{4.12}$$

2. O método BM (Bayes modal) é similar ao ML, exceto que a função a ser maximizada é a distribuição a *posteriori* $g(\theta)$, resultante da combinação da distribuição a *priori* $f(\theta)$ com a verossimilhança $L(\theta)$, $g(\theta) \propto f(\theta)\, L(\theta)$.

   A escolha de uma distribuição a *priori* é geralmente conduzida por alguma crença do pesquisador ou estudo anterior. Quando não se tem qualquer informação disponível, a opção é por uma *priori* pouco informativa. A escolha mais comum de distribuição a *priori* é a normal.

3. O terceiro método de estimação da habilidade a ser apresentado é o EAP (expected a *posteriori*). Enquanto o método BM calcula a moda da distribuição a *posteriori* o método EAP calcula a média *posteriori*:

$$\hat{\theta}_{EAP} = \frac{\int_{-\infty}^{+\infty} \theta\, f(\theta)\, L(\theta)\, d\theta}{\int_{-\infty}^{+\infty} f(\theta)\, L(\theta)\, d\theta} \tag{4.13}$$

   Como a distribuição dos níveis de habilidade é geralmente simétrica em torno do nível médio de habilidade, os estimadores BM e EAP frequentemente retornam estimativas similares.

4. O método WL (weighted likelihood) foi desenvolvido para reduzir e quase cancelar o viés do estimador ML, usando uma ponderação apropriada da função de verossimilhança. O estimador $\hat{\theta}_{WL}$ deve satisfazer a seguinte relação:

$$\frac{J(\theta)}{2\, I(\theta)} + \frac{d\, log\, L(\theta)}{d\, \theta} = 0, \tag{4.14}$$

   onde

$$J(\theta) = \sum_{j=1}^{J} \frac{P_j^{'}(\theta)\, P_j^{''}(\theta)}{P_j(\theta)\, (1 - P_j(\theta))}, \tag{4.15}$$

e $P_j^{''}(\theta)$ é a derivada segunda de $P_j(\theta)$ em relação a θ.

5. O ultimo método de estimação da habilidade a ser apresentado é o ROB (robust), proposto por (SCHUSTER; YUAN, 2011). Estes desenvolveram esta metodologia objetivando uma melhoria no método ML permitindo uma estimativa mais precisa na presença dos chamados distúrbios de resposta, como descuidos e "chutes". Em vez de definir a estimativa da habilidade como a solução para $\frac{d\,log\,L(\theta)}{d\,\theta} = 0$ temos agora:

$$\sum_{j=1}^{J} w(r_j) \left(\frac{d\,log\,L_j(\theta)}{d\,\theta} \right) = 0, \tag{4.16}$$

onde r_j é o resíduo que permite especificar quais itens serão considerados "potencialmente inconsistentes" e w é a função de peso Huber, definida como:

$$w(r) = \begin{cases} 1 & se\,|\mathrm{r}| \leq H, \\ \frac{H}{|r|} & se\,|\mathrm{r}| \geq H \end{cases} \tag{4.17}$$

H é chamada constante de sintonização, sendo que grandes valores significam que há pouca perda de peso, e pequenos que há uma considerável perda de peso do item. O resíduo, r_j, é influenciado pelo valor dos parâmetros e da habilidade, mas basicamente, itens com parâmetro de dificuldade distante da habilidade recebem valores altos de r_j. Portanto, este método atribui pesos aos itens, de forma que, aqueles que fornecem pouca informação para o nível de habilidade θ recebem pesos menores.

4.2.3 Critério de parada

O critério de parada determina a regra pela qual o teste será encerrado, ou seja, tendo o critério sido satisfeito, nenhum novo item será administrado e tem-se a estimativa final da habilidade. Há dois critérios de parada principais:

1. Definindo-se um tamanho de teste fixo: Todos os indivíduos responderam um teste adaptativo com o mesmo número de itens N.

2. Analisando a precisão da estimativa da habilidade: Encerra o teste quando o erro padrão provisório da habilidade se tornar menor ou igual a um valor pre-especificado, nesta caso, os indivíduos serão submetidos a testes de tamanhos distintos.

Spenassato destaca em seu artigo (SPENASSATO et al., 2016) que, em avaliações com objetivo de classificação dos participantes, como é o caso do ENEM, costuma-se utilizar

um número fixo de itens, pois um comprimento variável, nesses casos, pode não ser viável, porque os respondentes podem perceber o teste como injusto, dado que cada respondente poderá receber um comprimento de teste diferente.

5 Resultados e discussão

Para a aplicação de um teste adaptativo é necessário um banco de itens calibrado de acordo com algum modelo da Teoria de Resposta ao Item. Como o Inep não disponibiliza os valores dos parâmetros das questões tornou-se necessário realizar o download dos microdados do ENEM para estimar os parâmetros dos itens a partir das respostas dos participantes.

Os microdados do ENEM disponibilizados pelo Inep consistem no menor nível de desagregação dos dados do exame. Estes apresentam todo conteúdo das avaliações realizadas e questionários socioeconômicos respondidos, respeitando obviamente, o sigilo dos participantes. Os microdados possibilitam aos pesquisadores, com conhecimentos de linguagens de programação, realizar o desenvolvimento de estudos e indicadores sobre a educação brasileira, um dos principais objetivos do ENEM.

5.1 Estimação dos parâmetros e criação do banco de itens

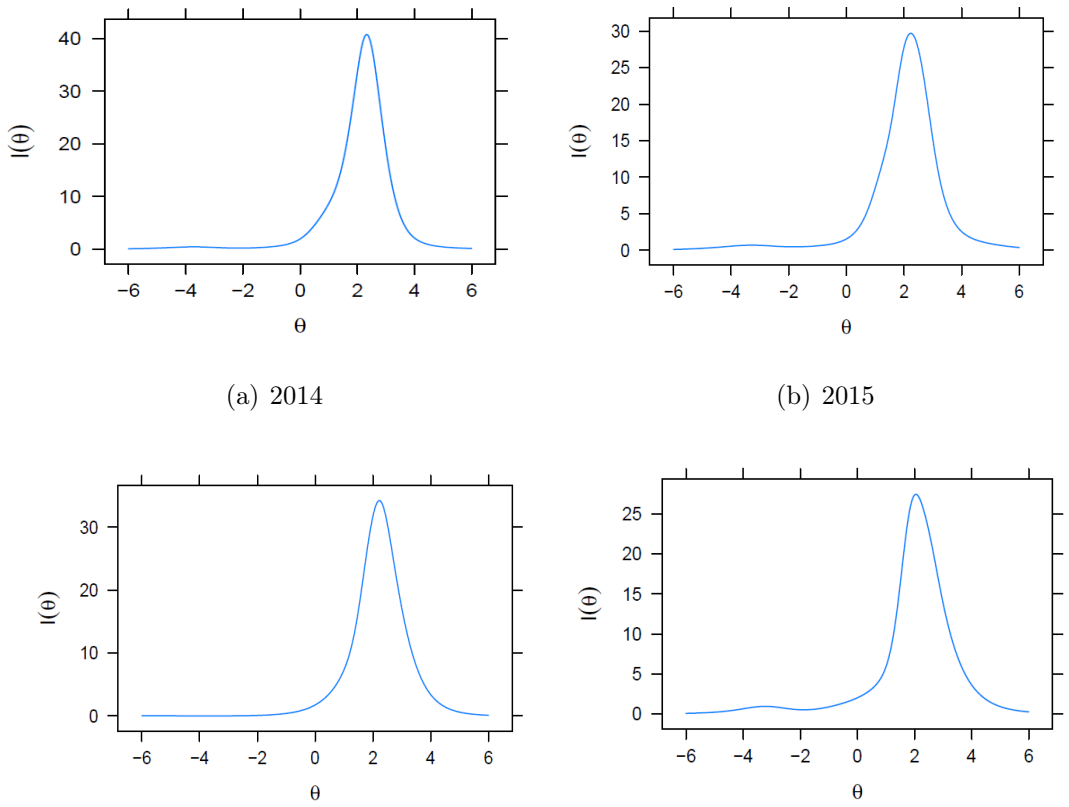
Um tópico importante de ser analisado é o tamanho do banco de itens. Na literatura encontram-se trabalhos que realizam alguns estudos adaptativos com os dados do ENEM como em (JATOBÁ et al., 2018); porém, trabalhando somente com uma prova, ou seja, 45 itens. Do ponto de vista realista, não tem sentido um banco para um teste adaptativo com este tamanho. Neste trabalho, optou-se portanto, por calibrar os itens de 4 provas do ENEM, para obtenção de um banco de itens maior e consequentemente mais realista. As provas escolhidas foram as 4 mais recentes com microdados disponíveis: 2014, 2015, 2016 e 2017.

Neste estudo coletamos amostras aleatórias de 300000 respostas às provas de matemática de cada um dos 4 anos. As amostras foram compostas por indivíduos que responderam às provas amarelas, cinzas, azuis ou rosas.

Para o desenvolvimento das análises presentes neste trabalho foi utilizada a linguagem de programação estatística R, (R Core Team, 2018). Caso seja de interesse de algum pesquisador, os códigos elaborados serão disponibilizados via solicitação. Com relação a estimação dos parâmetros dos itens, utilizou-se o algoritmo EM (Expectation-Maximization) implementado no pacote *mirt* (CHALMERS, 2012), sendo este um pacote R específico para análise de dados de respostas dicotômicas e politômicas usando modelos de traços latentes unidimensionais e multidimensionais sob o paradigma da Teoria de Resposta ao Item.

Como mencionado anteriormente, o parâmetro de dificuldade é medido na mesma escala da habilidade. O Inep atribui notas baseando-se em uma distribuição normal com média 500 e desvio padrão 100. Para evitar problemas de identificabilidade do modelo, uma das suposições é de que as habilidades seguem uma distribuição normal padrão e essa é a distribuição implementada no pacote *mirt*.

A partir dos parâmetros estimados pode-se fazer a plotagem das curvas de informação dos testes, que correspondem à soma das informações fornecidas por cada item. Na Figura 5 são apresentadas as curvas de informação dos testes correspondentes as provas do ENEM de 2014, 2015, 2016 e 2017. É possível perceber que estas provas se caracterizam por trazer mais informação para níveis de habilidade acima da média 0. Este aspecto do ENEM já havia sido percebido por Spenassato (SPENASSATO et al., 2016), em artigo publicado em 2016 e é compreensível, visto que hoje o ENEM se destaca por ser uma prova de caráter principalmente classificatório. Logo, é coerente que a prova tenha maior precisão para indivíduos com maior habilidade, pois, indivíduos com habilidades baixas dificilmente poderão almejar alguma vaga em universidades ou programas governamentais como o Prouni. Além disso, construir uma prova que também abranja de forma eficiente estes níveis acarretaria em um teste muito extenso.



(a) 2014

(b) 2015

(c) 2016

(d) 2017

Figura 5: Curvas de informação dos testes

Esta característica evidencia a importância da abordagem definida neste trabalho de agregar itens de 4 provas distintas, já que a aplicação de um teste adaptativo cujo banco de itens possui esta estrutura de curva de informação perde o sentido para indivíduos de baixa habilidade. Para estes, o teste iria em algum momento começar a selecionar itens que não acrescentariam praticamente nenhuma informação sobre as suas respectivas habilidades.

As habilidades dos indivíduos foram estimadas pelo método EAP (expected a posteriori) implementado na função *fscores*, do pacote *mirt*. Verificou-se a correlação entre as habilidades estimadas e a nota obtida no ENEM para os 4 anos. Se as estimativas dos parâmetros estiverem corretas é esperado que estas correlações sejam próximas de 1, o que foi confirmado, visto que, a menor correlação obtida foi 0.9758141.

A alta correlação é um indicador da qualidade na estimação dos parâmetros; porém, ainda assim foram encontrados itens com parâmetros considerados inadequados. Como pressupõem-se que a distribuição das habilidades é *normal(0,1)*, parâmetros de dificuldade muito distantes desta distribuição não são convenientes. Além disso, não há sentido em

parâmetros de discriminação negativos, pois estes indicariam que a curva característica do item decresce com o aumento da habilidade, ou seja, quanto maior é a habilidade do indivíduo menor é a probabilidade deste acertar o item, sendo discordante com a coerência pedagógica.

Para formação do banco de itens definiu-se, portanto, que seriam utilizados itens cujo parâmetro de discriminação fosse positivo e o de dificuldade estivesse no intervalo [-5,5]. Após essa filtragem, o banco de itens a ser utilizado para aplicação e simulação dos testes adaptativos ficou formado por 165 itens.

Optou-se por realizar as análises seguintes utilizando apenas as habilidades estimadas para o ano de 2016, pois este ano teve menos parâmetros excluídos, somente dois. Estas estimativas obtidas via TRI foram definidas como as verdadeiras. O valor estimado via teste adaptativo foi então comparado com este.

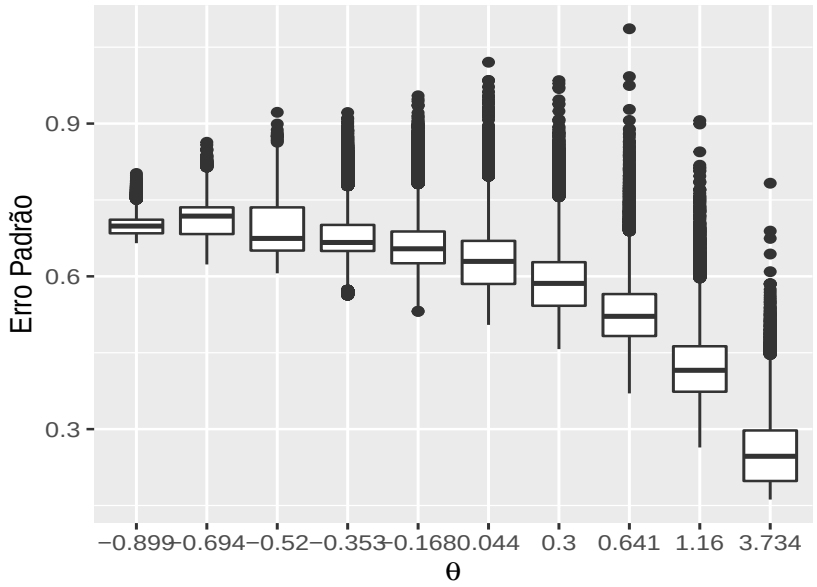
Inicialmente classificou-se as habilidades em 10 grupos de acordo com os decis e consequentemente cada grupo composto por 30000 valores. Esta classificação é apresentada na Tabela 1.

Tabela 1: Classificação das habilidades

Grupos	Quantidade
[-1.2044 ,-0.8988]	30000
]-0.8988 ,-0.6943]	30000
]-0.6943 ,-0.5197]	30000
]-0.5197 ,-0.3532]	30000
]-0.3532 ,-0.1682]	30000
]-0.1682 , 0.0442]	30000
] 0.0442 , 0.3003]	30000
] 0.3003 , 0.6414]	30000
] 0.6414 , 1.1598]	30000
] 1.1598 , 3.7342]	30000

A partir da classificação gerada apresenta-se na Figura 6 os boxplots dos erros padrões obtidos via TRI de cada grupo. Esse gráfico mostra que é nítida a influência da curva de informação do teste. Como o teste apresenta mais informação para níveis de habilidade acima da média, os erros padrões diminuem consequentemente com o aumento da habilidade. Os erros padrões influenciam nos intervalos de confiança e, portanto, quanto menores mais precisa é a estimativa da habilidade.

Figura 6: Boxplots dos erros padrões TRI



5.2 Definição do design dos cenários

Para a realização das análises adaptativas, foi gerada uma amostra de tamanho 10000 uniformemente distribuída em todas as faixas definidas na Tabela 1.

Salienta-se que para a realização dos teste adaptativos foi utilizada a função *randomCAT* do *catR*, (MAGIS; RAÎCHE, 2012; MAGIS; BARRADA, 2017), pacote do R próprio para o desenvolvimento de testes adaptativos computadorizados. Porém, ao analisar o código fonte foi identificado um erro que afeta a aleatoriedade do processo e, portanto, foram realizadas pequenas modificações nas funções *genPattern, nextItem, randomCAT* e *startItems* com o intuito basicamente de eliminar a expressão *set.seed(NULL)*.

Destaca-se que, como mencionado, em (SPENASSATO et al., 2016) e (JATOBÁ et al., 2018) os autores optaram pela denominada simulação *post-hoc*, onde a resposta do indivíduo não é simulada, mas sim coletada do vetor de respostas real utilizado para estimação dos parâmetros. Porém, neste trabalho como agregamos itens de quatro provas distintas este tipo de simulação não era possível, além do que, ao realizar a coleta da resposta real do indivíduo, esta pode ter influência da metodologia utilizada, *paper and pencil*. Visando obter resultados fieis à metodologia adaptativa, as respostas dos indivíduos neste trabalho não foram coletadas do vetor de respostas real, mas sim simuladas.

É necessário frisar que neste trabalho não será abordado balanceamento de conteúdo e nem o controle da taxa de exposição do item. O balanceamento visa garantir que, mesmo

com a diminuição do tamanho do teste e os itens sendo selecionados de forma aleatória, a prova aborde todo o conteúdo programático. O controle da taxa de exposição do item visa impedir que itens sejam expostos a uma quantidade muito grande de indivíduos, bem como a uma quantidade muito pequena, esse controle é necessário devido a questão de segurança e também devido ao custo de produção de um item. Estes são pontos fundamentais na aplicação real de um teste adaptativo, porém, o Inep não disponibiliza a qual conteúdo corresponde cada item, inviabilizando o balanceamento, e com relação a taxa de exposição do item é um quesito que foge ao escopo deste estudo de simulação.

Nas seções seguintes iremos analisar os resultados das simulações dos 10000 testes adaptativos nos diversos cenários explorados, sendo que, cada um destes foi construído a partir da escolha de 1 método de estimação da habilidade dentre 5 possíveis e 1 método de seleção do próximo item dentre 6 possíveis, totalizando 30 cenários. Ressalta-se que, o pacote *catR* permite a utilização de 11 métodos de seleção do próximo item, no entanto, a utilização de todos tornaria este estudo muito extenso e de difícil análise dos resultados. Portanto, optamos pela utilização dos 6 métodos de seleção com execução mais ágil.

Além disso, a seleção do primeiro item do teste só pode ser feita pelos métodos MFI (Máxima Informação de Fisher) ou bOpt (Procedimento de Urry), deste modo, foi definido que em todos os cenários o primeiro item é selecionado pelo método MFI para um nível de habilidade θ igual a 0, exceto, quando o método de seleção do próximo item for o bOpt.

Com relação ao critério de parada definimos como o número máximo de itens a serem administrados igual a 45.

5.3 Cenários com método de estimação BM (Bayes modal)

A partir da simulação dos 10000 testes adaptativos em cada cenário, passa-se a análise dos resultados. Nesta seção investigamos os cenários cujo método de estimação da habilidade foi o BM.

Na Figura 7 e Tabela 2 podemos verificar como se comporta o erro padrão médio com a evolução do teste. É possível verificar que os cenários analisados apresentam resultados distintos no inicio; no entanto, com o progresso do teste tornam-se extremamente similares.

Examinando a Tabela 2 podemos concluir que o cenário que obteve os menores erros padrões utilizando o método de estimação da habilidade BM foi o que utilizou o MFI como método de seleção do próximo item. Nas 4 etapas do teste destacadas na tabela, este ce-

nário gerou os menores valores; porém, como mencionado, com resultados muito similares aos demais. Além disso, destaca-se que com 25 itens administrados todos os cenários que utilizaram o método de estimação BM produziram erros padrões médios inferiores a 0.4, sendo que um teste de 25 itens representaria uma diminuição de aproximadamente 44% no tamanho da prova do ENEM, que com a metodologia atual é de 45 itens.

Figura 7: Erro padrão médio por etapa - cenários com método BM

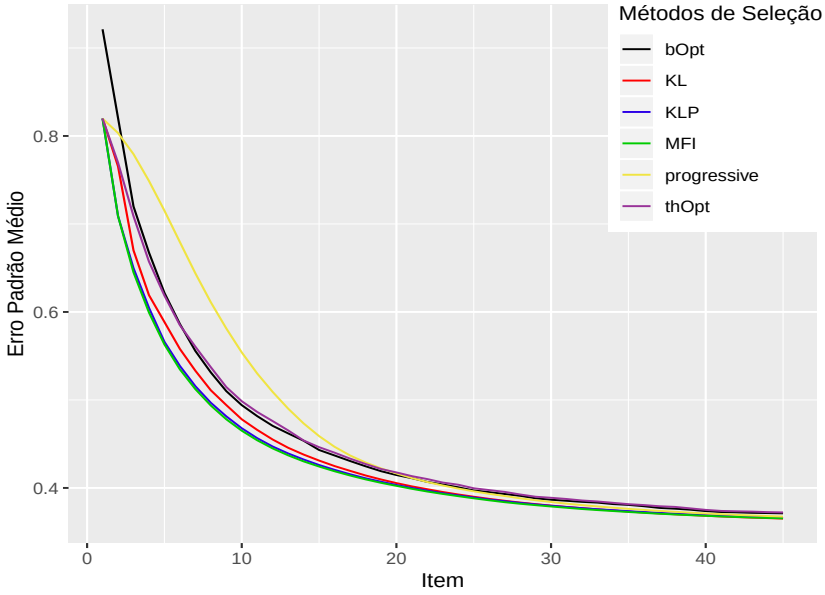


Tabela 2: Erro padrão médio por etapa - cenários com método BM

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.4024	0.4147	0.4176	0.4165	0.4055	0.4034
25	0.3882	0.3972	0.3996	0.3962	0.3900	0.3892
30	0.3789	0.3866	0.3888	0.3841	0.3799	0.3798
35	0.3727	0.3807	0.3816	0.3762	0.3731	0.3737

Na Figura 8 e Tabela 3 apresentamos a variação do erro padrão médio em cada etapa do teste adaptativo, sendo esta definida como:

$$\text{VEP}_i = \frac{|\text{se}_{i-1} - \text{se}_i|}{\text{se}_{i-1}} \tag{5.1}$$

em que,

- VEP_i corresponde a variação do erro padrão médio na etapa i;
- se_i corresponde ao erro padrão médio na etapa i.

A Figura 8 mostra que o cenário que utiliza o método de seleção do próximo item *progressive* tem um comportamento inicial distinto dos demais, no entanto, à medida que o teste progride ele se equipara aos outros. Na Tabela 3 verifica-se que com a aplicação de 20 itens a maior variação do erro padrão médio foi de 0.0137, ou seja, uma variação percentual de somente 1.37%.

Figura 8: Variação do erro padrão médio por etapa - cenários com método BM

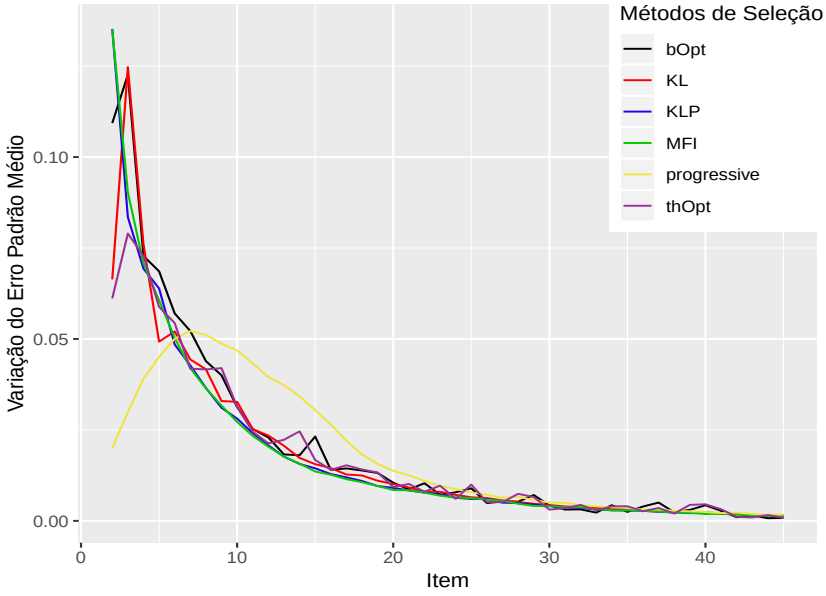


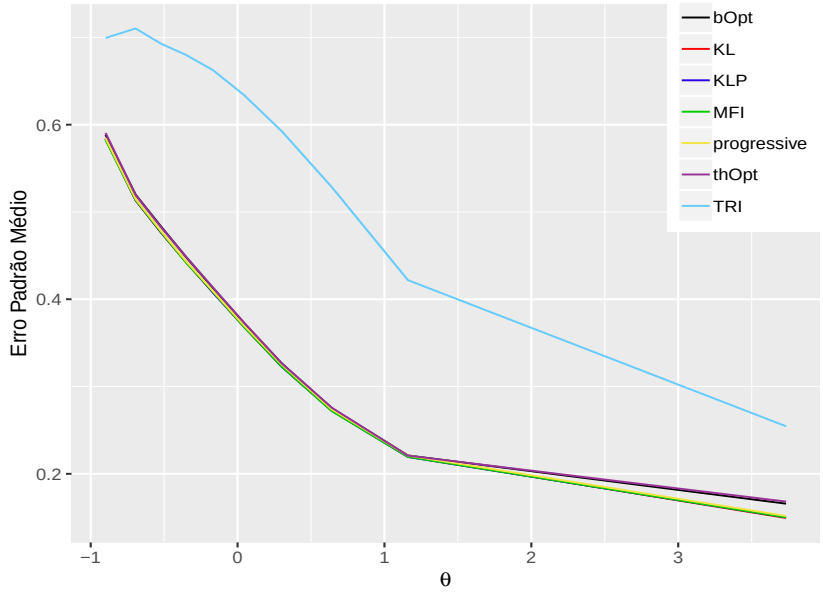
Tabela 3: Variação do erro padrão médio por etapa - cenários com método BM

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0085	0.0105	0.0093	0.0137	0.0101	0.0091
25	0.0062	0.0090	0.0100	0.0078	0.0065	0.0060
30	0.0040	0.0042	0.0031	0.0050	0.0043	0.0041
35	0.0029	0.0025	0.0040	0.0035	0.0035	0.0029

Na Figura 9 se compara os erros padrões médios obtidos via TRI e cenários adaptativos, estes com 45 itens administrados, em cada uma das faixas de habilidade definidas na Tabela 1. É nítido que em todos os níveis de habilidade a abordagem adaptativa gera erros padrões menores em todos os cenários; no entanto, ainda mantém maior precisão para níveis de habilidade maiores. Mesmo com a união de 4 provas do ENEM não foi possível garantir a mesma precisão em todas as faixas de habilidades, visto que esta é uma característica intrínseca do teste. Porém, as vantagens do teste adaptativo ficam evidentes e caso o banco fosse composto por itens que distribuíssem de forma mais homogênea a informação nos níveis de habilidade esse comportamento seria minimizado ou mesmo

eliminado.

Figura 9: Erro padrão médio por θ - cenários com método BM e TRI



Passamos agora para a análise do RMSE (Root Mean Square Error), sendo este definido como:

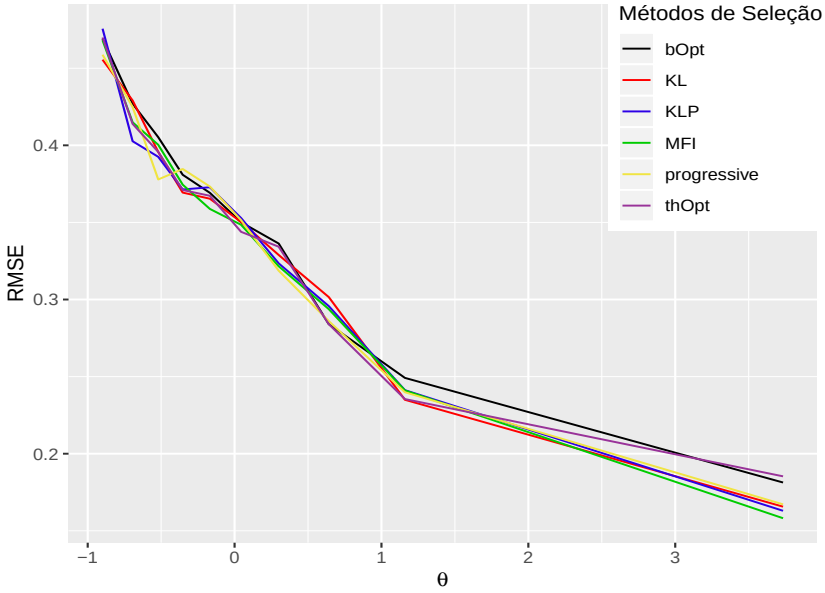
$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2}{N}} \tag{5.2}$$

em que,

- θ_j é a habilidade real do indivíduo j;
- $\hat{\theta}_j$ é a habilidade estimada do indivíduo j.

A Figura 10 mostra os valores do RMSE em função do nível de habilidade, para um teste de 45 itens. Como esperado pelos resultados obtidos com o erro padrão médio, onde verificou-se que os testes aumentam a precisão à medida que aumenta a habilidade, o RMSE diminui com o aumento da habilidade para todos os métodos de seleção.

Figura 10: RMSE por θ - cenários com método BM



Na Figura 11 e Tabela 4 podemos verificar como o RMSE se comporta com a evolução do teste adaptativo. O cenário que utilizou o MFI como método de seleção do próximo item obteve o menor RMSE para número de itens igual a 20, 25 e 35. Contudo, nota-se novamente resultados extremamente similares com a evolução do teste.

Figura 11: RMSE por etapa - cenários com método BM

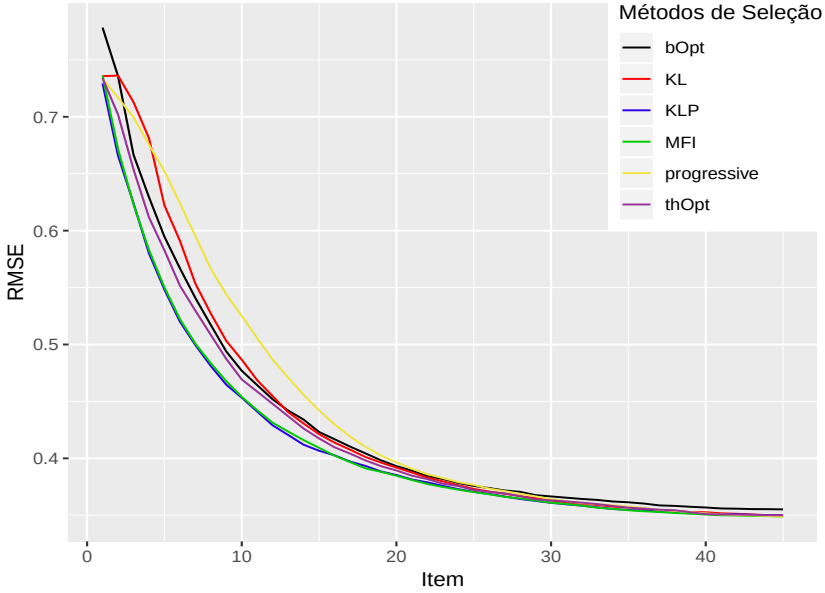


Tabela 4: RMSE por etapa - cenários com método BM

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.3847	0.3932	0.3893	0.3964	0.3918	0.3854
25	0.3704	0.3757	0.3721	0.3767	0.3731	0.3706
30	0.3610	0.3665	0.3631	0.3642	0.3629	0.3609
35	0.3544	0.3614	0.3567	0.3577	0.3568	0.3546

Na Figura 12 e Tabela 5 verifica-se a variação do RMSE durante o progresso do teste adaptativo, sendo esta definida como:

$$\text{VRMSE}_i = \frac{|\text{RMSE}_{i-1} - \text{RMSE}_i|}{\text{RMSE}_{i-1}} \tag{5.3}$$

em que,

- VRMSE_i corresponde a variação do RMSE na etapa i;
- RMSE_i corresponde ao RMSE na etapa i.

Pode-se constatar pela análise da Tabela 5 que com 25 itens administrados a maior variação foi de 0.0095, ou seja, menos de 1%.

Figura 12: Variação do RMSE por etapa - cenários com método BM

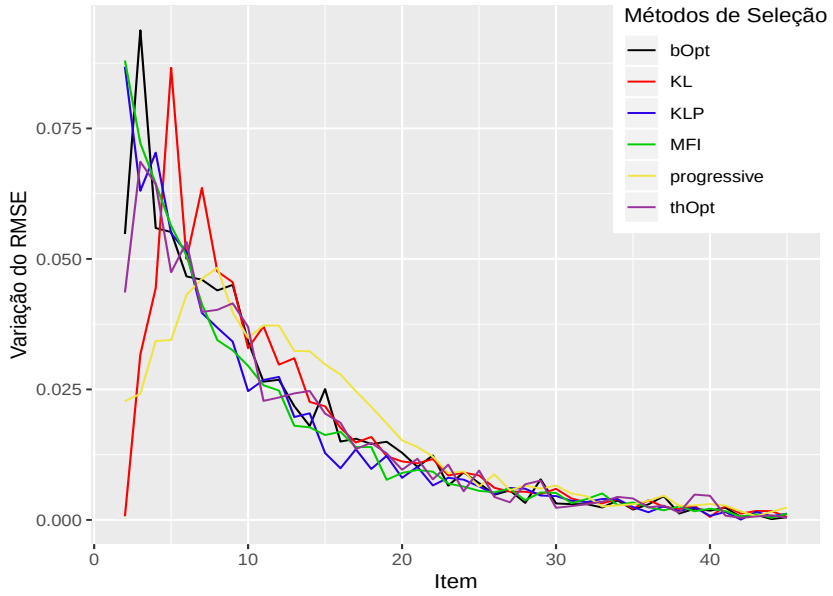


Tabela 5: Variação do RMSE por etapa - cenários com método BM

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0090	0.0129	0.0096	0.0152	0.0112	0.0081
25	0.0056	0.0070	0.0095	0.0064	0.0085	0.0064
30	0.0052	0.0032	0.0023	0.0066	0.0060	0.0046
35	0.0034	0.0020	0.0041	0.0030	0.0021	0.0025

5.4 Cenários com método de estimação ML (maximum likelihood)

Passamos agora a análise dos resultados obtidos com os cenários cujo método de estimação da habilidade foi definido como ML.

Na Figura 13 e Tabela 6 podemos verificar como se comporta o erro padrão médio durante cada etapa do teste. Estes cenários iniciaram com erros padrões extremamente elevados, em comparação com os resultados dos cenários cujo método de estimação foi o BM (Figura 7). Porém, é notório que há um decrescimento de forma vertiginosa com poucos itens aplicados, com exceção do cenário que utiliza o *progressive* como método de seleção do próximo item. Este só assume valores similares aos demais após a aplicação de mais de 10 itens. Além disso, nota-se que o cenário que utiliza o método de seleção do próximo item KLP gerou os menores erros padrões em todas as 4 etapas apontadas na tabela, porém, com valores muito similares aos demais.

Comparando os resultados das Tabelas 6 e 2 concluímos que os cenários com método de estimação BM produzem erros padrões menores que os com método de estimação ML.

Figura 13: Erro padrão médio por etapa - cenários com método ML

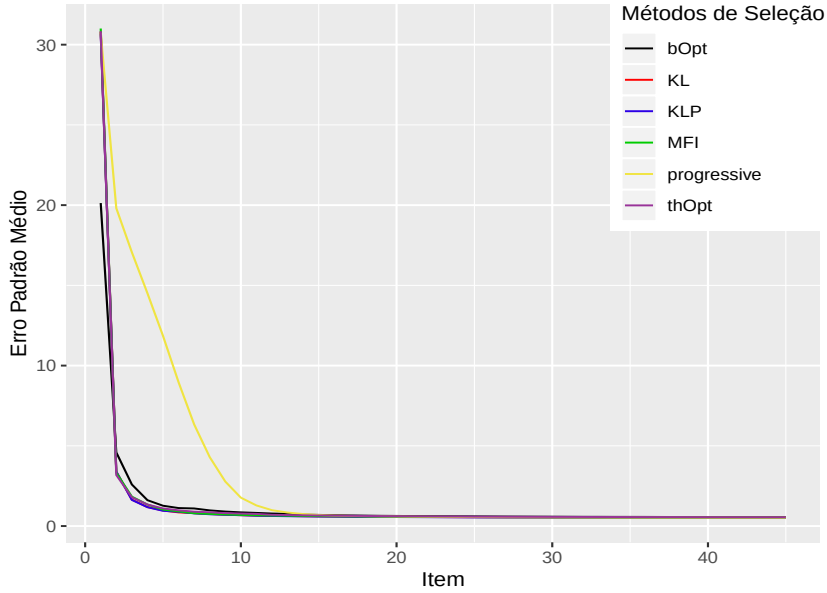


Tabela 6: Erro padrão médio por etapa - cenários com método ML

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.5834	0.6321	0.6145	0.5970	0.5723	0.5635
25	0.5606	0.6034	0.5875	0.5637	0.5513	0.5432
30	0.5466	0.5831	0.5722	0.5458	0.5374	0.5324
35	0.5377	0.5721	0.5613	0.5328	0.5302	0.5229

Na Figura 14 e Tabela 7 observamos a variação do erro padrão médio durante a aplicação do teste. O cenário que utiliza o método de seleção *progressive*, como esperado, apresenta um comportamento distinto e se assemelha aos demais somente em torno da aplicação do vigésimo item. Contudo, após a administração de 25 itens só o cenário com método de seleção thOpt apresenta variação percentual maior que 1%.

Figura 14: Variação do erro padrão médio por etapa - cenários com método ML

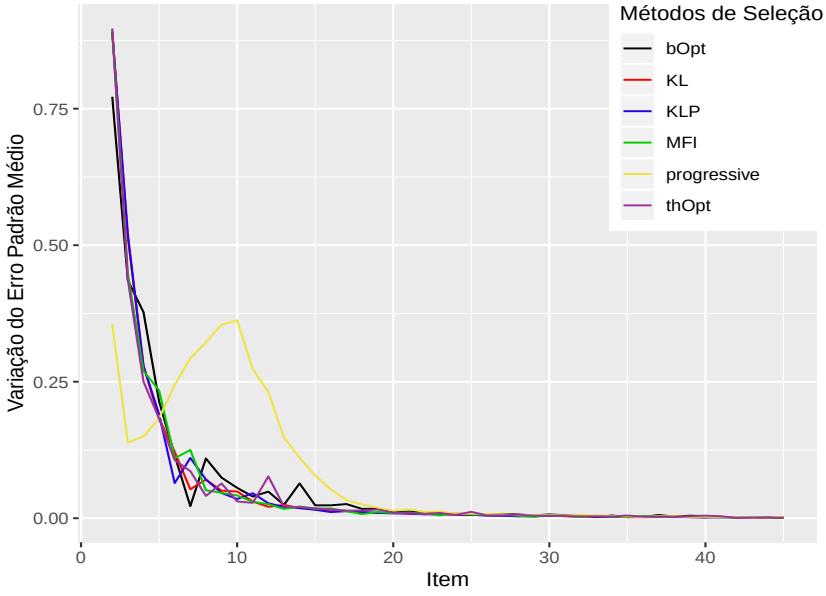
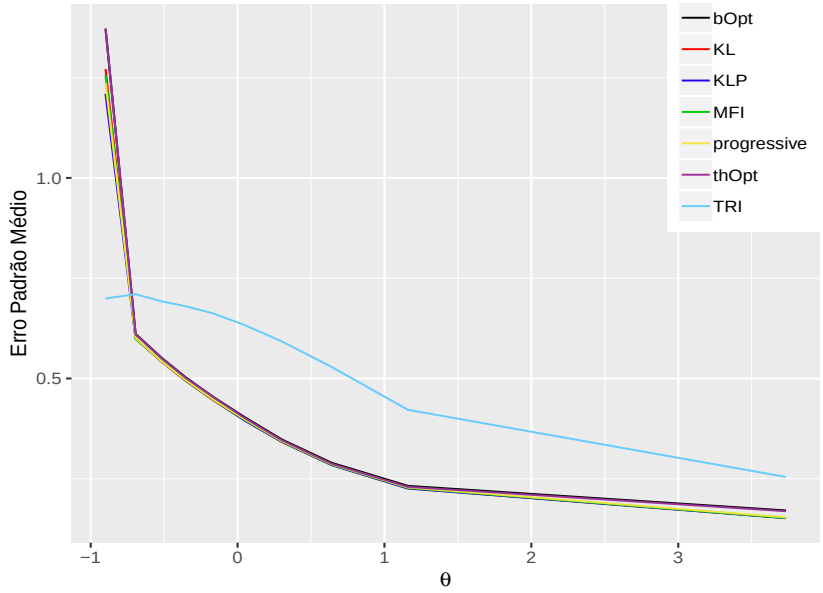


Tabela 7: Variação do erro padrão médio por etapa - cenários com método ML

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0093	0.0111	0.0101	0.0136	0.0101	0.0087
25	0.0072	0.0060	0.0117	0.0085	0.0061	0.0065
30	0.0062	0.0069	0.0046	0.0055	0.0050	0.0040
35	0.0025	0.0019	0.0050	0.0035	0.0020	0.0030

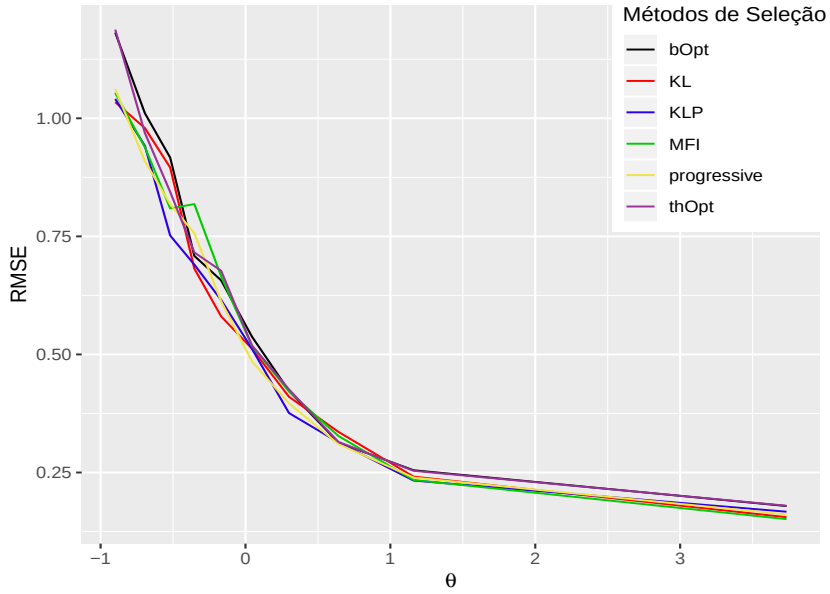
Na Figura 15 mostramos os erros padrões obtidos com os 6 cenários e utilizando o modelo TRI tradicional, todos com 45 itens. O aspecto de maior precisão para níveis de habilidade maiores se mantem nestes cenários. Nota-se que, para a primeira faixa os cenários adaptativos com método de estimação ML geraram erros padrões elevados, inclusive maiores que os obtidos com TRI.

Figura 15: Erro padrão médio por θ - cenários com método ML e TRI



Na Figura 16 apresentamos o RMSE obtido após a aplicação de 45 itens com relação aos níveis de habilidade. Tal como esperado, o RMSE diminui com o aumento do nível de habilidade. Comparando com a Figura 10 percebemos que para os níveis de habilidade menores, os RMSE's gerados com os cenários que utilizaram o método de estimação ML foram muito maiores que os obtidos pelos cenários cujo método de estimação foi o BM.

Figura 16: RMSE por θ - cenários com método ML



Na Figura 17 e Tabela 8 expomos como o RMSE prossegue com a administração dos itens. Em todas as 4 etapas analisadas na Tabela 8 o cenário que utiliza o KLP como

método de seleção do próximo item obtêm os menores valores. No entanto, todos estes resultados apresentados com relação ao RMSE foram maiores que os obtidos com o método de estimação BM, Tabela 4.

Figura 17: RMSE por etapa - cenários com método ML

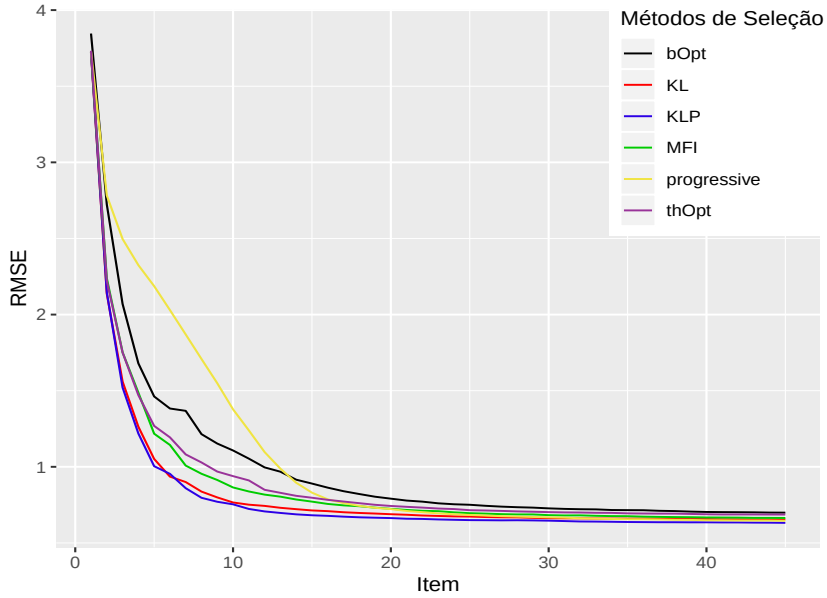


Tabela 8: RMSE por etapa - cenários com método ML

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.7241	0.7909	0.7429	0.7223	0.6887	0.6632
25	0.6955	0.7507	0.7150	0.6838	0.6731	0.6498
30	0.6825	0.7271	0.7024	0.6677	0.6622	0.6463
35	0.6757	0.7155	0.6944	0.6559	0.6598	0.6375

Na Figura 18 e Tabela 9 verifica-se a variação do RMSE por etapa. Com a evolução do teste os resultados se tornam similares, sendo que com 30 itens administrados todos os cenários apresentaram variação percentual inferior a 1%.

Figura 18: Variação do RMSE por etapa - cenários com método ML

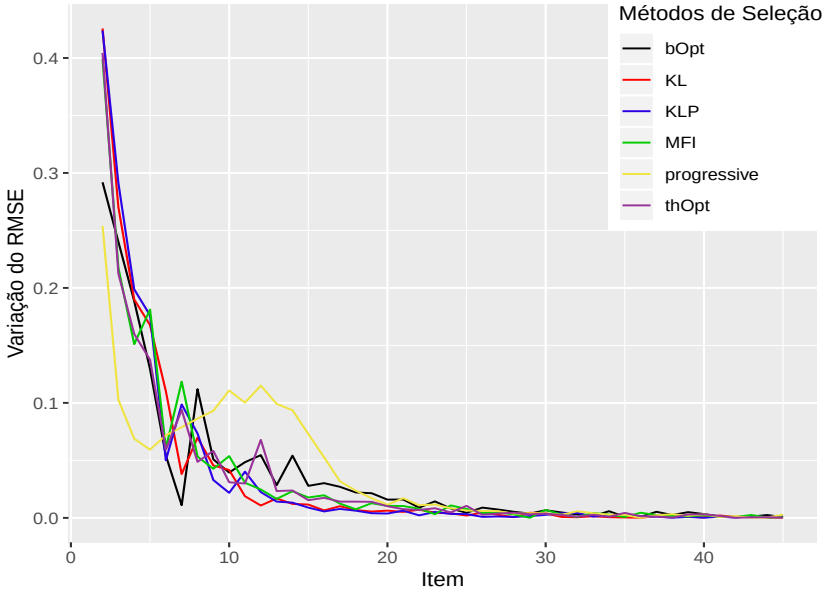


Tabela 9: Variação do RMSE por etapa - cenários com método ML

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0105	0.0159	0.0101	0.0115	0.0063	0.0038
25	0.0078	0.0045	0.0105	0.0063	0.0020	0.0033
30	0.0069	0.0066	0.0037	0.0040	0.0039	0.0027
35	0.0007	0.0010	0.0044	0.0026	0.0003	0.0021

5.5 Cenários com método de estimação EAP (expected a posteriori)

Verificaremos agora os resultados gerados pelos cenários onde foi utilizado o EAP como método de estimação da habilidade.

Na Figura 19 e Tabela 10 identificamos que após a aplicação de 20 itens os resultados com relação ao erro padrão médio nos cenários analisados tornam-se extremamente semelhantes. Contudo, percebemos ainda que o cenário que utilizou o MFI como método de seleção do próximo item gerou os menores valores nas 4 etapas do teste apresentadas na tabela. Observando a Tabela 2, percebemos que os resultados obtidos com o método de estimação BM foram um pouco melhores que os apresentados na Tabela 10, porém, muito próximos.

Figura 19: Erro padrão médio por etapa - cenários com método EAP

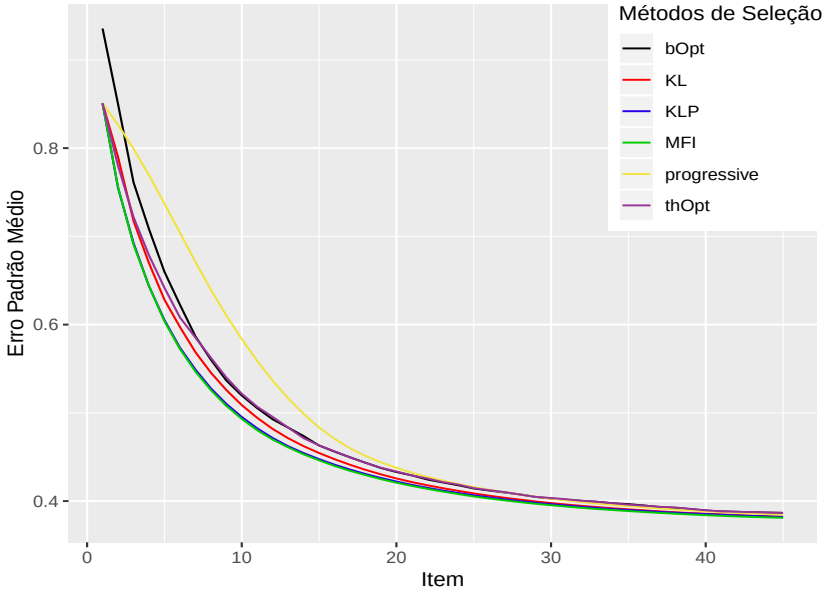


Tabela 10: Erro padrão médio por etapa - cenários com método EAP

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.4208	0.4329	0.4335	0.4377	0.4256	0.4222
25	0.4052	0.4143	0.4147	0.4159	0.4087	0.4066
30	0.3953	0.4029	0.4035	0.4027	0.3978	0.3965
35	0.3887	0.3966	0.3960	0.3943	0.3906	0.3899

Na Tabela 11 e Figura 20 exploramos a variação do erro padrão médio durante o progresso dos testes. Com 25 itens aplicados, todos os cenários apresentam variação percentual menor que 1%.

Figura 20: Variação do erro padrão médio por etapa - cenários com método EAP

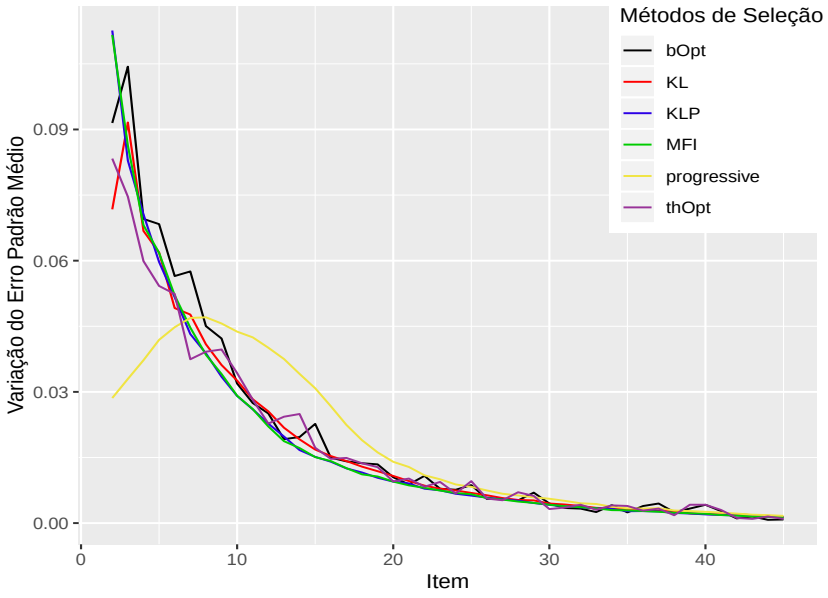
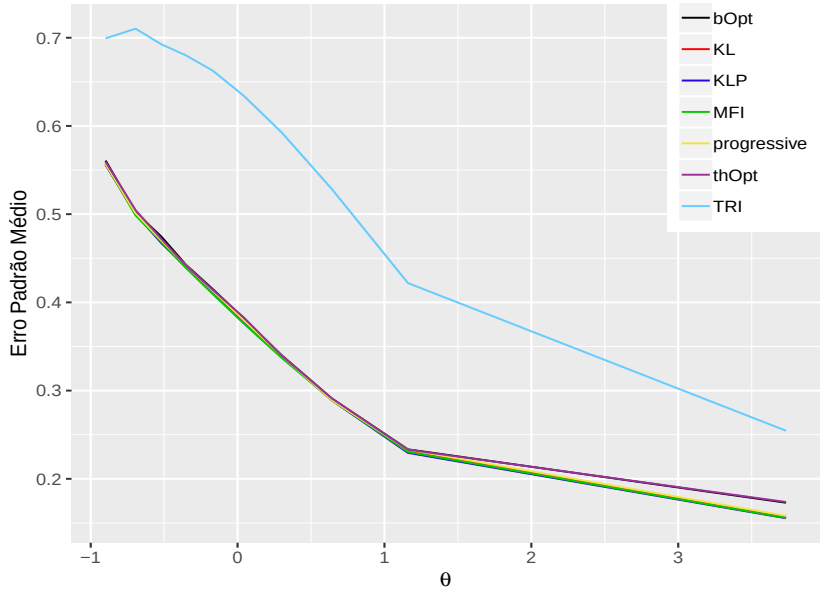


Tabela 11: Variação do erro padrão médio por etapa - cenários com método EAP

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0095	0.0105	0.0095	0.0140	0.0108	0.0095
25	0.0066	0.0087	0.0096	0.0083	0.0069	0.0063
30	0.0042	0.0045	0.0032	0.0056	0.0045	0.0042
35	0.0029	0.0024	0.0039	0.0034	0.0032	0.0027

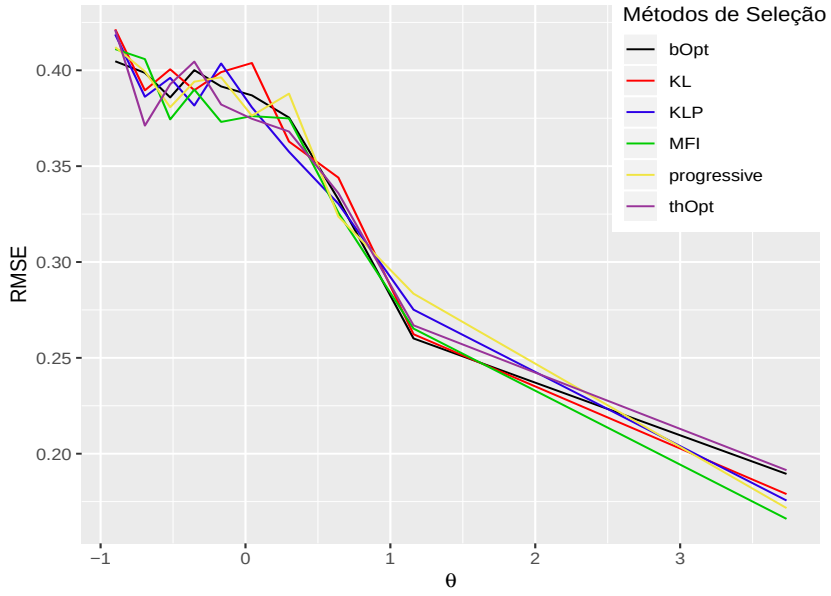
Na Figura 21 verificamos que após 45 itens administrados todos os cenários apresentam erros padrões médios menores que os obtidos via TRI, em todos os níveis de habilidade. Novamente os erros padrões diminuem à medida que aumenta a habilidade.

Figura 21: Erro padrão médio por θ - cenários com método EAP e TRI



Na Figura 22 é analisado o RMSE gerado com 45 itens nas diversa faixas de habilidade. Os cenários em estudo apresentam resultados muito próximos que, mais uma vez, melhoram com o aumento da habilidade.

Figura 22: RMSE por θ - cenários com método EAP



Na Tabela 12 e Figura 23 vemos como progride o RMSE com a aplicação dos itens e percebemos que com 25 itens todos os cenários apresentam valores inferiores a 0.4. Novamente notamos que após 20 itens os resultados dos cenários são extremamente próximos; contudo, nas 4 etapas da Tabela 12 o método de seleção do próximo item MFI

gera o menor RMSE. Além disso, confrontando estes resultados com os obtidos via método de estimação BM, Tabela 4, verificamos que os cenários que utilizaram o método EAP alcançaram valores muito similares aos com o método BM, porém, ainda um pouco inferiores.

Figura 23: RMSE por etapa - cenários com método EAP

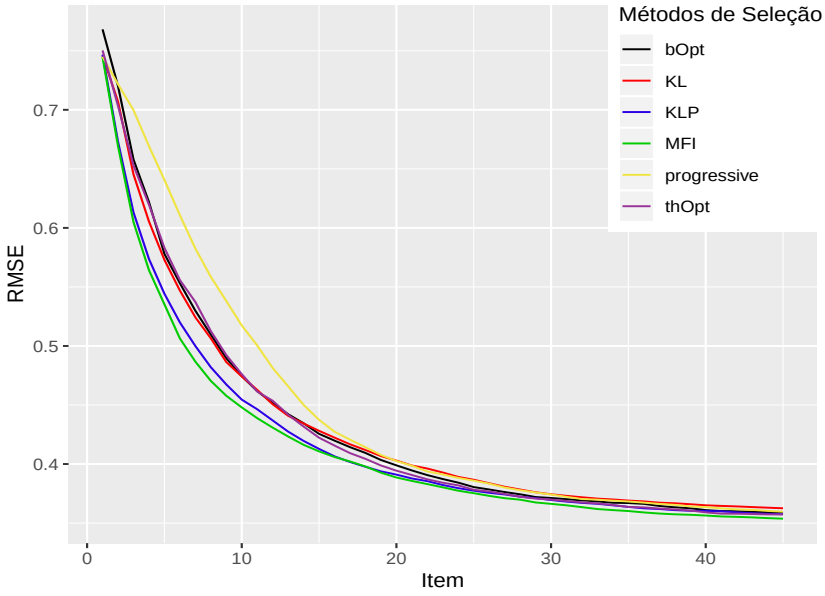


Tabela 12: RMSE por etapa - cenários com método EAP

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.3886	0.3989	0.3944	0.4025	0.4029	0.3910
25	0.3753	0.3805	0.3782	0.3859	0.3867	0.3774
30	0.3663	0.3712	0.3699	0.3742	0.3744	0.3695
35	0.3602	0.3669	0.3637	0.3683	0.3690	0.3637

Na Figura 24 e Tabela 13 podemos constatar que com 25 itens aplicados todos os cenários com método de estimação EAP apresentam variação percentual menor que 1%, visto que, o máximo foi 0.0098.

Figura 24: Variação do RMSE por etapa - cenários com método EAP

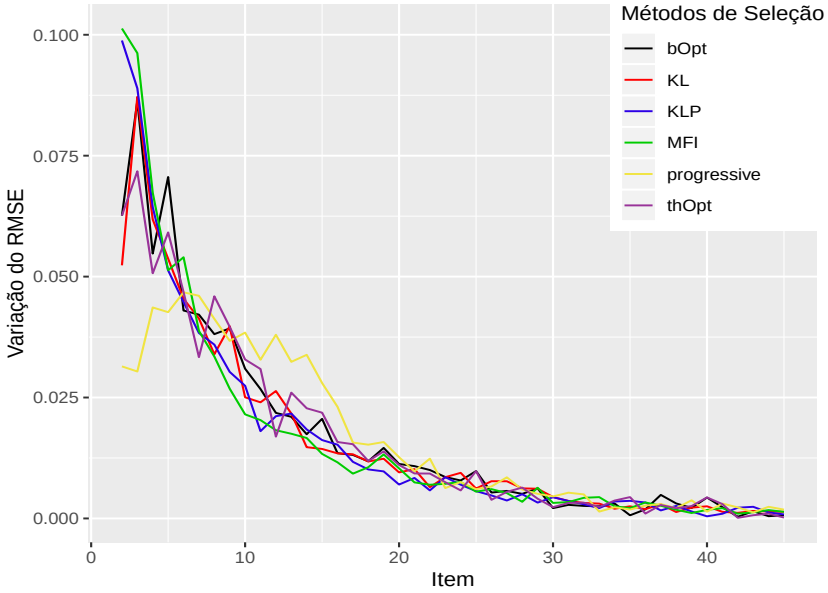


Tabela 13: Variação do RMSE por etapa - cenários com método EAP

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0103	0.0113	0.0111	0.0127	0.0095	0.0070
25	0.0055	0.0098	0.0098	0.0060	0.0062	0.0056
30	0.0032	0.0021	0.0023	0.0045	0.0044	0.0044
35	0.0021	0.0007	0.0044	0.0018	0.0025	0.0036

5.6 Cenários com método de estimação WL (weighted likelihood)

Nesta seção analisaremos os resultados obtidos com os cenários que utilizaram o WL como método de estimação da habilidade.

Analisando a Figura 25 e Tabela 14 verificamos que o cenário que utiliza o método de seleção do próximo item bOpt inicia com um erro padrão médio mais elevado, porém, rapidamente se estabiliza com os demais. Além do mais, em torno do item 20, todos os cenários começam a apresentar resultados similares e após o item 25 todos os cenários apresentam valores menores que 0.48. Os resultados apresentados não são melhores que os obtidos com os métodos de estimação BM e EAP, porém, são melhores que os obtidos com o ML.

Figura 25: Erro padrão médio por etapa - cenários com método WL

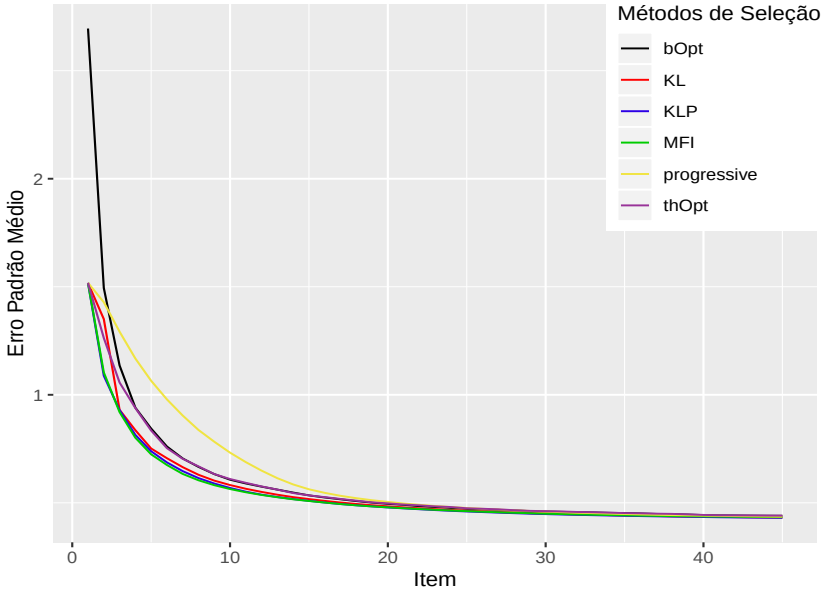


Tabela 14: Erro padrão médio por etapa - cenários com método WL

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.4791	0.4953	0.4975	0.5041	0.4839	0.4785
25	0.4615	0.4724	0.4748	0.4760	0.4636	0.4600
30	0.4496	0.4597	0.4614	0.4595	0.4500	0.4478
35	0.4420	0.4526	0.4525	0.4482	0.4418	0.4400

Com a Figura 26 e Tabela 15 constatamos que com 25 itens aplicados praticamente todos os cenários apresentam variação percentual do erro padrão inferior a 1%, com exceção do cenário que utilizou o método de seleção do próximo item thOpt. Com 30 itens administrados somente um cenário, que utiliza o método de seleção do próximo item *progressive*, apresenta variação percentual superior a 0.5%.

Figura 26: Variação do erro padrão médio por etapa - cenários com método WL

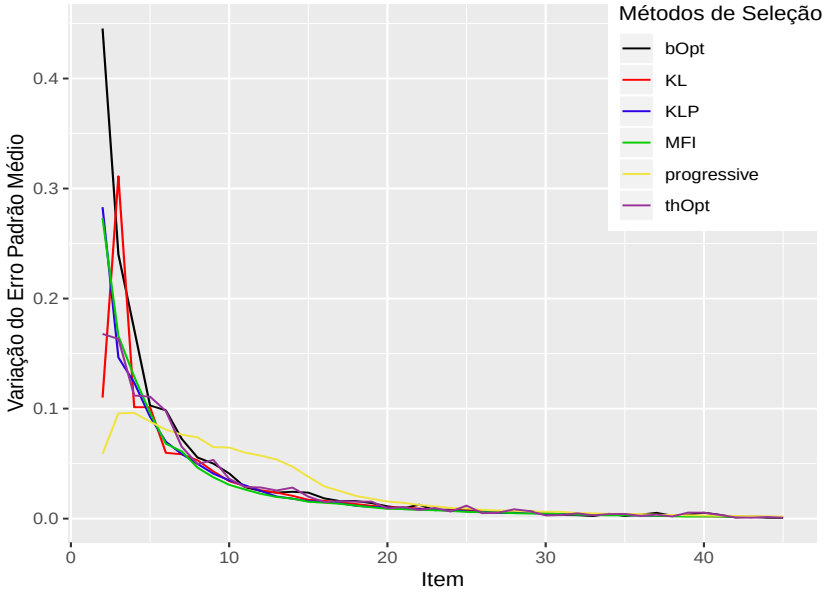
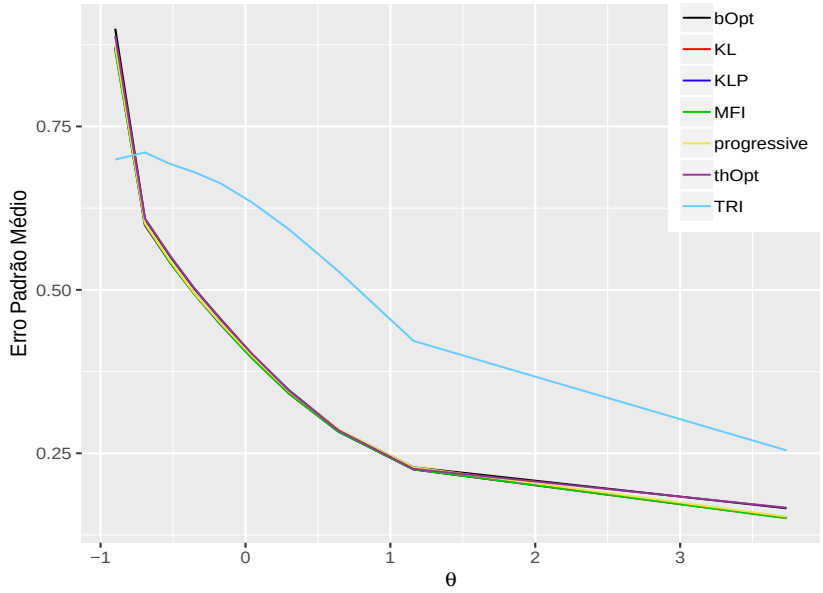


Tabela 15: Variação do erro padrão médio por etapa - cenários com método WL

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0089	0.0112	0.0093	0.0155	0.0104	0.0102
25	0.0062	0.0090	0.0119	0.0092	0.0073	0.0066
30	0.0049	0.0042	0.0029	0.0063	0.0049	0.0043
35	0.0031	0.0024	0.0042	0.0043	0.0033	0.0032

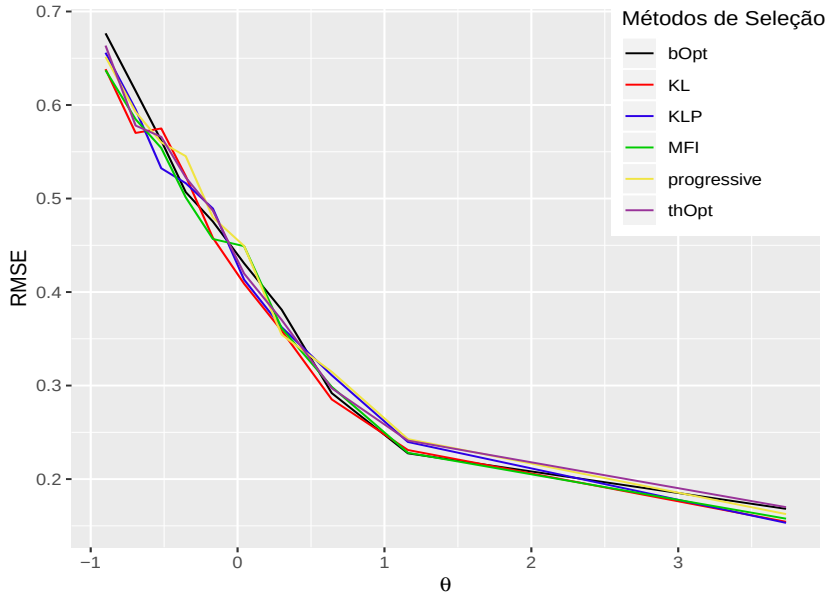
Analisando a Figura 27 percebemos que, após 45 itens administrados, para a primeira faixa de habilidade o erro padrão médio obtido pelos cenários que utilizaram o método de estimação WL foi maior que o obtido via somente TRI, sem abordagem adaptativa, assim como havia sido notado nos cenários cujo método de estimação foi o ML.

Figura 27: Erro padrão médio por θ - cenários com método WL e TRI



Na Figura 28 estudamos o RMSE alcançado após 45 itens nas faixas de θ e constatamos a sua diminuição com o aumento da habilidade. Notamos, outra vez, resultados similares entre os cenários.

Figura 28: RMSE por θ - cenários com método WL



Na Figura 29 e Tabela 16 identificamos que os cenários que utilizaram o método de estimação WL apresentam aumento do RMSE no inicio do teste. Além disso, comparando com os resultados anteriores percebemos que estes são melhores que os obtidos com o método ML, porém, inferiores aos obtidos com os métodos BM e EAP.

Figura 29: RMSE por etapa - cenários com método WL

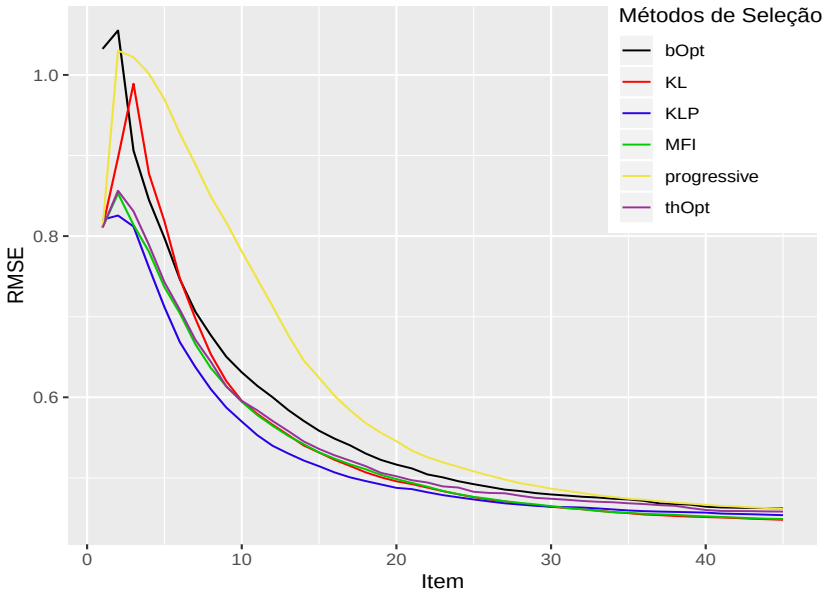


Tabela 16: RMSE por etapa - cenários com método WL

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.4985	0.5165	0.5020	0.5456	0.4958	0.4877
25	0.4764	0.4922	0.4827	0.5081	0.4761	0.4733
30	0.4650	0.4795	0.4741	0.4867	0.4642	0.4642
35	0.4566	0.4732	0.4685	0.4741	0.4563	0.4596

Examinando a Figura 30 e Tabela 17 percebemos que com 25 itens aplicados 2 cenários ainda apresentam variação percentual do RMSE maior que 1%.

Figura 30: Variação do RMSE por etapa - cenários com método WL

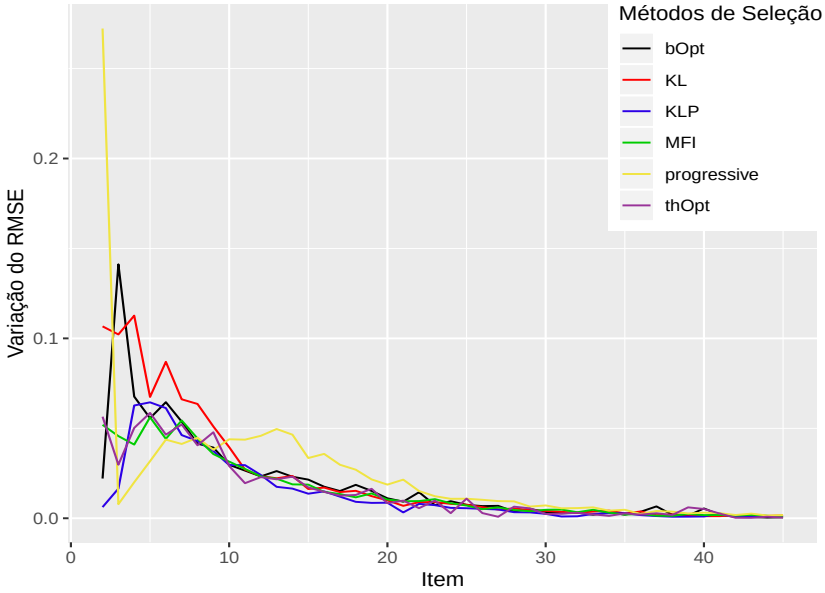


Tabela 17: Variação do RMSE por etapa - cenários com método WL

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0104	0.0111	0.0084	0.0187	0.0099	0.0086
25	0.0071	0.0075	0.0109	0.0109	0.0075	0.0056
30	0.0047	0.0035	0.0023	0.0071	0.0041	0.0025
35	0.0020	0.0020	0.0027	0.0047	0.0027	0.0028

5.7 Cenários com método de estimação ROB (robust)

Nesta seção analisamos os últimos cenários restantes, gerados com o método de estimação ROB.

Verificamos na Figura 31 e Tabela 18 o erro padrão médio por etapa do teste e percebemos que estes cenários iniciam com valores elevados, bem como os cenários que utilizaram o método de estimação ML. Em todas as 4 etapas analisadas na Tabela 18 o cenário que utilizou o método de seleção KLP apresenta os menores erros padrões, no entanto, também nota-se que os valores dos cenários em estudo não apresentam grande discrepância.

Figura 31: Erro padrão médio por etapa - cenários com método ROB

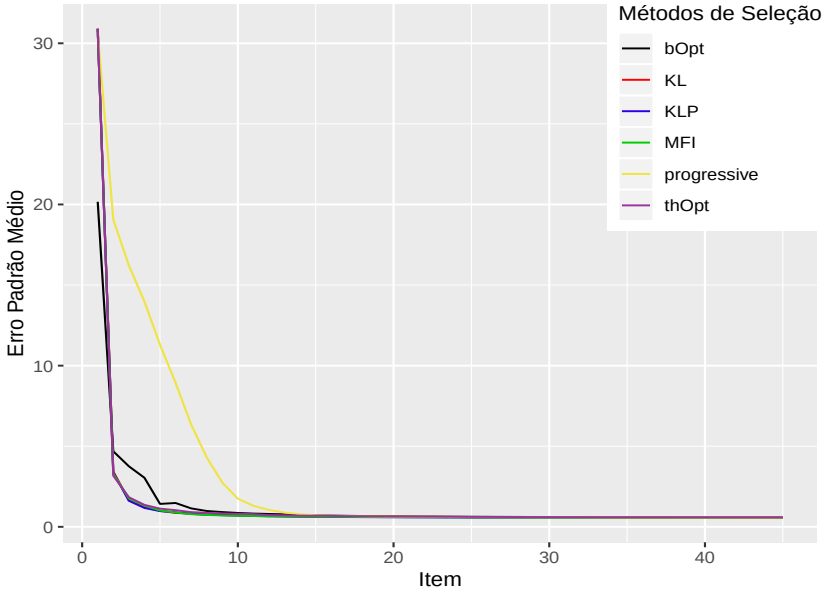


Tabela 18: Erro padrão médio por etapa - cenários com método ROB

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.6022	0.6538	0.6415	0.6358	0.6169	0.5925
25	0.5856	0.6234	0.6168	0.6094	0.5999	0.5764
30	0.5747	0.6087	0.6040	0.5934	0.5902	0.5688
35	0.5666	0.5995	0.5935	0.5855	0.5828	0.5638

Na Figura 32 e Tabela 19 examinamos a variação do erro padrão médio no decorrer do teste e com 25 itens aplicados a maior variação encontrada foi 0.0101, com a utilização do método de seleção thOpt, ou seja, somente 1.01%.

Figura 32: Variação do erro padrão médio por etapa - cenários com método ROB

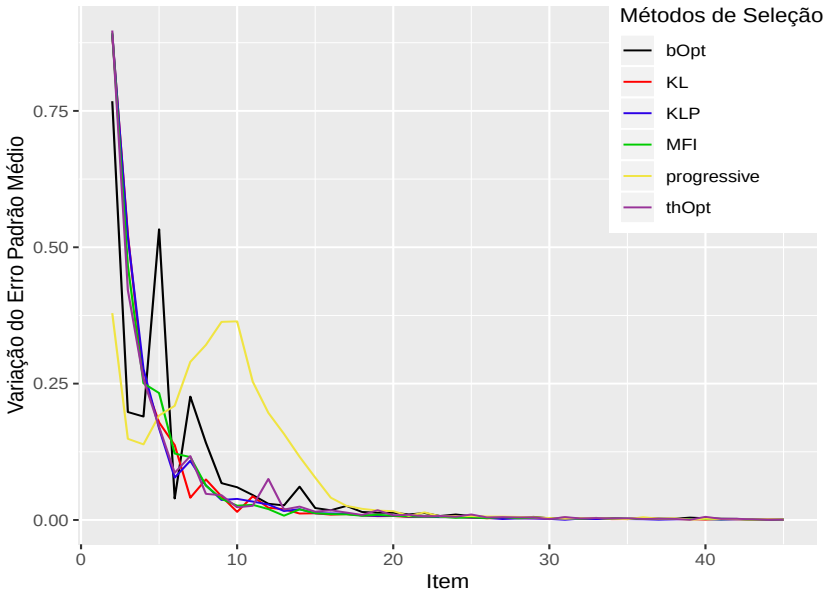
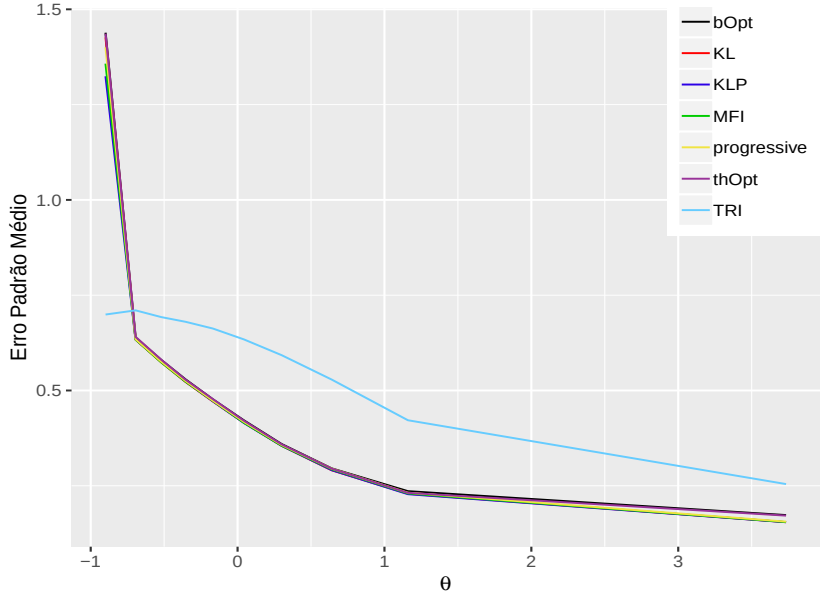


Tabela 19: Variação do erro padrão médio por etapa - cenários com método ROB

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0075	0.0134	0.0085	0.0161	0.0073	0.0081
25	0.0051	0.0072	0.0101	0.0071	0.0042	0.0045
30	0.0029	0.0031	0.0016	0.0036	0.0030	0.0019
35	0.0024	0.0021	0.0031	0.0014	0.0017	0.0016

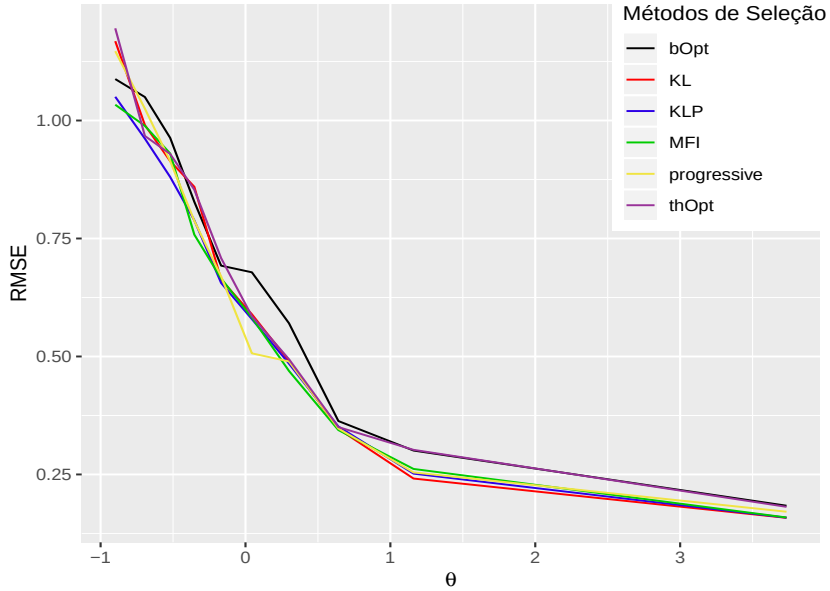
Na Figura 33 percebemos que, assim como outros cenários já analisados, na primeira faixa de habilidade os cenários que utilizaram o método de estimação ROB geraram erros padrões maiores que os produzidos com a utilização de TRI sem abordagem adaptativa.

Figura 33: Erro padrão médio por θ - cenários com método ROB e TRI



Na Figura 34 temos o RMSE por nível de habilidade e novamente identificamos que o RMSE diminui com o aumento da habilidade.

Figura 34: RMSE por θ - cenários com método ROB



Na Figura 35 e Tabela 20 verificamos o RMSE por etapa. Com 25 itens aplicados todos os cenários apresentam RMSE menor que 0.78. Em todas as 4 etapas da Tabela os melhores resultados foram obtidos pelo cenário que utilizou o método de seleção KLP e os piores o método bOpt. Estes valores encontrados, fazendo comparação com os métodos de estimação anteriores, são inferiores aos obtidos com os métodos EAP e BM.

Figura 35: RMSE por etapa - cenários com método ROB

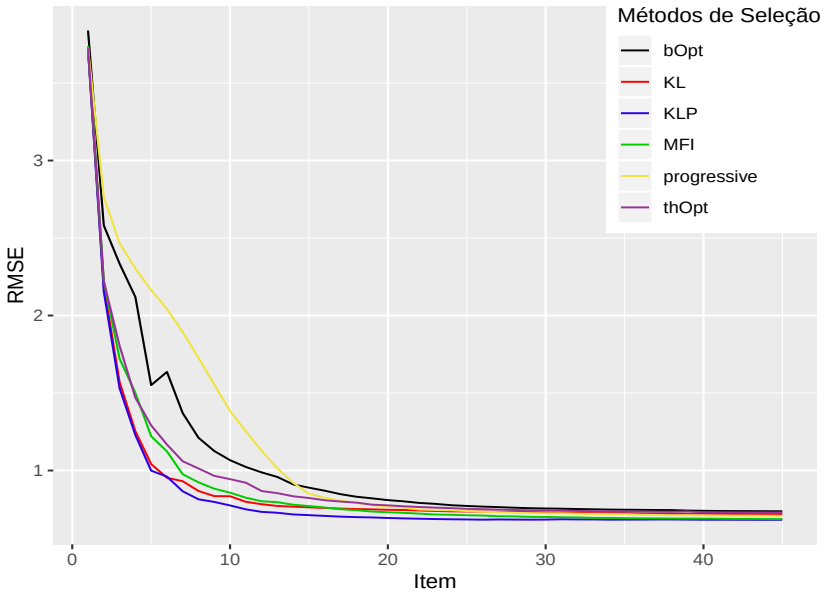


Tabela 20: RMSE por etapa - cenários com método ROB

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.7308	0.8091	0.7755	0.7656	0.7465	0.6938
25	0.7112	0.7714	0.7528	0.7385	0.7356	0.6844
30	0.7003	0.7551	0.7438	0.7239	0.7302	0.6833
35	0.6929	0.7471	0.7354	0.7185	0.7252	0.6832

Fazendo a análise da variação do RMSE nota-se que após 25 itens administrados, todos os cenários que utilizaram o método de estimação ROB apresentam valores menores que 0.01.

Figura 36: Variação do RMSE por etapa - cenários com método ROB

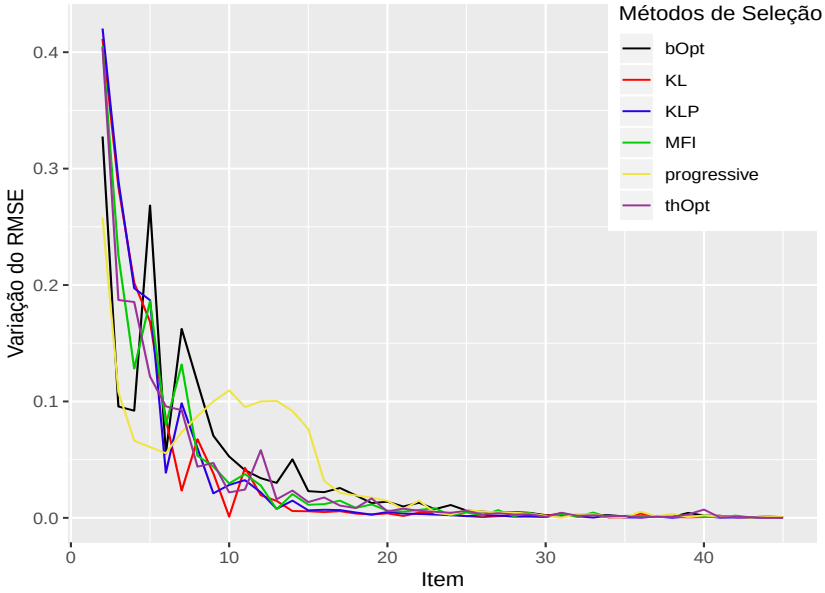


Tabela 21: Variação do RMSE por etapa - cenários com método ROB

Item	MFI	bOpt	thOpt	progressive	KL	KLP
20	0.0062	0.0139	0.0055	0.0146	0.0039	0.0050
25	0.0046	0.0063	0.0061	0.0065	0.0015	0.0015
30	0.0011	0.0023	0.0009	0.0018	0.0018	0.0007
35	0.0017	0.0013	0.0015	0.0004	0.0004	0.0004

6 Conclusões

Com as análises do erro padrão médio e RMSE apresentadas percebemos que os melhores resultados foram obtidos pelos cenários cujo método de estimação da habilidade foi definido como BM ou EAP, sendo que os cenários com método BM apresentaram resultados ligeiramente melhores. Isso demonstra que as abordagens Bayesianas para os métodos de estimação se comportaram melhor do que as clássicas.

O melhor cenário pode ser considerado o que utilizou o método de estimação da habilidade BM e o método de seleção do próximo item MFI, visto que, com relação ao erro padrão médio por etapa, este gerou os menores valores nas 4 etapas analisadas e, com relação ao RMSE por etapa, este gerou os melhores resultados em 3 das 4 etapas. Contudo, é necessário salientar que com a aplicação de pelos menos 20 itens, percebeu-se que, tendo fixado um método de estimação da habilidade a alteração do método de seleção do próximo item não provoca alterações drásticas nos resultados.

Os resultados que dizem respeito à variação do RMSE e erro padrão demonstram que os testes podem ser encerrados com menos que 45 itens sem grandes prejuízos nos resultados. O cenário com método de estimação BM e seleção MFI, por exemplo, com 20 itens aplicados já apresenta variação percentual inferior a 1% tanto para o RMSE como para o erro padrão.

Verificou-se que mesmo com a união de 4 provas do ENEM para criação do banco de itens, a maior precisão para níveis de habilidade maiores se manteve em todos os cenários adaptativos. Este aspecto foi identificado em todas as análises de RMSE e erro padrão. Porém, como já mencionado, caso o banco fosse composto por itens que distribuíssem de forma mais homogênea a informação nos níveis de habilidade esse comportamento seria minimizado ou mesmo eliminado. Além disso, foi mostrado que os cenários cujo método de estimação da habilidade foi definido como BM ou EAP geraram erros padrões menores que os obtidos via TRI, sem abordagem adaptativa, em todos os níveis de habilidade. Os demais cenários geraram erros padrões maiores somente para a primeira faixa de habilidade.

Todos estes resultados apresentados mostram que uma abordagem adaptativa pode reduzir bastante o tamanho do teste a ser aplicado, além de gerar estimativas mais precisas. Outra vantagem já mencionada é a redução de custos com aplicação, armazenamento e correção de provas. Portanto, este trabalho demonstra como a abordagem adaptativa é promissora para o aprimoramento das provas de larga escala aplicadas no Brasil e amplia os estudos sobre este assunto tão incipiente ainda no país; contudo, ainda há aspectos que devem ser analisados detalhadamente em estudos futuros antes de qualquer aplicação em situação real, como a taxa de exposição do item e o balanceamento de conteúdo.

Referências

ALEXANDRE, J. W. C. et al. Teoria da resposta ao item: aplicação do modelo de escala gradual na gestão pela qualidade. *Anais do Encontro Nacional de Engenharia de Produção, Curitiba-PR*, v. 22, 2002.

ANDRADE, D.; ANJOS, A. Teoria da resposta ao item com uso do r. *Joao Pessoa-PB*, 2012.

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

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

ENEM, E. a. S. N. no. Guia do participante. *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Brasília*, 2012.

FERRÃO, M. E.; PRATA, P. Item response models in computerized adaptive testing: a simulation study. In: SPRINGER. *International Conference on Computational Science and Its Applications*. [S.l.], 2014. p. 552–565.

JATOBÁ, V. et al. Comparação de regras de seleção de itens em testes adaptativos computadorizados: um estudo de caso no enem. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2018. v. 29, n. 1, p. 1453.

MAGIS, D.; BARRADA, J. R. Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, Code Snippets*, v. 76, n. 1, p. 1–19, 2017.

MAGIS, D.; RAÎCHE, G. Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, v. 48, n. 8, p. 1–31, 2012.

OLIVEIRA, L. H. M. de; ALUÍSIO, S. M.; PÍTON, J. Criação e aplicação de testes adaptativos informatizados: um estudo de caso. 2004.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>.

SASSI, G. P. *Teoria e a prática de um teste adaptativo informatizado*. Tese (Doutorado) — Universidade de São Paulo, 2012.

SCHUSTER, C.; YUAN, K.-H. Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, SAGE Publications Sage CA: Los Angeles, CA, v. 36, n. 6, p. 720–735, 2011.

SPENASSATO, D. et al. Testes adaptativos computadorizados aplicados em avaliações educacionais. *Revista Brasileira de Informática na Educação*, v. 24, n. 2, 2016.