UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

UMA SOLUÇÃO BAYESIANA PARA SE CONSIDERAR A INCERTEZA ASSOCIADA À CALIBRAÇÃO DE ITENS NA TEORIA DE RESPOSTA AO ITEM

Ana Carolina Fernandes Dias

Belo Horizonte

Junho, 2019

UMA SOLUÇÃO BAYESIANA PARA SE CONSIDERAR A INCERTEZA ASSOCIADA À CALIBRAÇÃO DE ITENS NA TEORIA DE RESPOSTA AO ITEM

Ana Carolina Fernandes Dias

Dissertação submetida ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais - UFMG, como parte dos requisitos necessários à obtenção do grau de Mestre em Estatística.

Orientador: Prof. Marcos Oliveira Prates
Coorientador: Prof. Flávio Bambirra Gonçalves

Belo Horizonte

Junho, 2019

FICHA CATALOGRÁFICA

Dias, Ana Carolina Fernandes

Uma solução Bayesiana para se considerar a incerteza associada à calibração de itens na Teoria de Resposta ao Item.

Belo Horizonte: UFMG, ICEX, JUNHO DE 2019.

Dissertação - Universidade Federal De Minas Gerais.

Aos meus pais e ao Michel.

AGRADECIMENTOS

Agradeço a Deus pela minha família, pela minha saúde, e todas as coisas maravilhosas que tenho em minha vida que me permitem e me dão forças para lutar e realizar os meus sonhos e objetivos.

Agradeço à minha família, meus pais (Vanda Lúcia e Roberto Dias) e meu irmão (Luis Gustavo), por serem minha base, por acreditarem em mim e na minha capacidade, por fazerem de tudo para que eu chegasse até aqui e por sempre me acompanharem aonde quer que eu esteja.

Ao amor da minha vida, Michel Ferreira que me apoia e me dá forças para continuar, por me lembrar sempre do que sou capaz e entender o tempo que precisei dedicar aos estudos.

Aos meus amigos pelos momentos compartilhados, ajuda no crescimento profissional, pelas conversas, risadas e por todos os desafios que conseguimos ultrapassar juntos.

Aos meus orientadores pelo voto de confiança, por me direcionar, por me permitir desenvolver este trabalho e por todo conhecimento transmitido, além da disposição e paciência em ajudar.

Agradeço aos colegas de curso Gabriel Oliveira Assunção e Juliane Venturelli Silva Lima por fornecerem dados e códigos necessários para que o meu trabalho fosse realizado.

A todos os professores que de algum modo contribuíram para o meu conhecimento e crescimento profissional e fizeram com que eu tivesse orgulho e muito amor em exercer essa profissão.

A todos aqueles que fazem parte da minha vida, e torceram para que eu conseguisse chegar até aqui.

O conhecimento é o processo de acumular dados; a sabedoria reside na sua simplificação.

Martin H. Fischer

RESUMO

Nos métodos convencionais da Teoria de Reposta ao Item (TRI) comumente estima-se o traço latente dos indivíduos fundamentado em um teste previamente calibrado, ou seja, assume-se que os parâmetros dos itens são conhecidos após os valores serem estimados através de um pré-teste, dessa forma existe uma incerteza na estimação dos parâmetros dos itens, uma vez que usamos uma amostra para estimá-los. Ignorar essa incerteza, pode levar a erros inferenciais das estimativas dos traços latentes, particularmente quando a amostra de calibração não é suficientemente grande. A partir da necessidade de incluir a incerteza existente na calibração dos itens nas estimativas das habilidades dos indivíduos, este trabalho propõe uma abordagem Bayesiana para tratar do problema de estimar a habilidade levando em consideração a incerteza quanto aos parâmetros dos itens pré-calibrados. É proposto um algoritmo que aproxima a distribuição *a posteriori* de um indivíduo submetido ao teste pré-calibrado a partir da amostra da distribuição *a posteriori* dos parâmetros dos itens obtida via MCMC. Por fim, o algoritmo proposto é estendido para o contexto de testes adaptativos, permitindo a estimação da habilidade a cada item respondido. Neste contexto, são propostos novos métodos de escolhas de itens e regra de parada. A metodologia proposta é investigada em análises de dados simulados e ilustrada na análise de um conjunto de dados do Enem 2017.

Palavras-chave: Teoria de Resposta ao Item. Teste Adaptativo Informatizado. TRI. TAI. Estatística Bayesiana. Incerteza na estimação dos parâmetros.

ABSTRACT

In the conventional methods of the Item Response Theory (IRT), the latent trait of the individuals is usually estimated based on a previously calibrated test, that is, it is assumed that the parameters of the items are known after the values are estimated through a pre-test, thus there is an uncertainty in the estimation since we use a sample to estimate them. Ignoring this uncertainty can lead to inferential errors of estimates of latent traits, particularly when the calibration sample is not large enough. This paper proposes a Bayesian approach to deal with the problem of estimating the ability taking into account the uncertainty regarding the parameters of the pre-calibrated items. An algorithm that approximates the posterior distribution of an individual submitted to the pre-calibrated test from the sample of the posterior distribution of the parameters of the items obtained via MCMC is proposed. Finally, the discussed algorithm is extended to the context of adaptive tests, allowing the estimation of the ability to each item answered. In this context, new methods of item choices and stop rules are proposed. The proposed methodology is investigated in simulated data analysis and illustrated in the analysis of a data set related to the Enem
.

Keywords: Item Response Theory. Computer Adaptive Test. IRT. CAT. Bayesian Statistics,Uncertainty in parameter estimation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3PL – Three parameter logistic model

3PNO – Three parameter normal ogive model

CAT – Computerized Adaptive Testing ou Computering Adaptive Testing ou Computer Adaptive Test

CCI – Curva Característica do Item

ENEM – Exame Nacional do Ensino Médio

EAP – Estimador Bayesiano da média a posteriori (Esperança a posteriori)

EMV – Estimador de Máxima Verossimilhança

IEAP – Informação Esperada a posteriori

IF – Informação de Fisher

IO – Informação Observada

KL – Kullback Leibler

MAP – Moda a posteriori

MC – Monte Carlo

MCMC – Markov Chain Monte Carlo

MI – Máxima Informação

MIE – Máxima Informação Esperada

MMV – Métodos de Máxima Verossimilhança

MV – Máxima Verossimilhança

SAEB – Sistema de Avaliação da Educação Básica

SARESP – Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo

TAI – Teste Adaptativo Informatizado

TCT – Teoria Clássica dos Testes

TRI – Teoria de Resposta ao Item

SUMÁRIO

1 INTRODUÇÃO

1.1 INTRODUÇÃO

A avaliação educacional é uma tarefa necessária e constitui um dos pontos importantes das políticas educacionais visto que, colabora para a redefinição dessas políticas, fazendo o acompanhamento de todos os passos do processo de ensino e aprendizagem.

A Teoria de Resposta ao Item (TRI) é uma metodologia estatística sofisticada e precisa, que permite não só avaliar o conhecimento do respondente em um teste, como também acompanhar o progresso do seu conhecimento adquirido ao longo do tempo. Essa metodologia é mais adequada que a teoria clássica que utiliza os escores brutos (onde as notas são dadas pelo total de questões respondidas corretamente em um teste) ou escores padronizados. A TRI trata cada resposta como resultado de um experimento aleatório cuja probabilidade depende da característica do item e da habilidade do respondente. Por meio da TRI, é possível responder a várias questões de interesse prático no ambiente educacional, como o desenvolvimento de uma determinada série de um ano para o outro e a comparação do desempenho entre escolas públicas e privadas, por exemplo.

A utilização dos modelos da TRI é bastante consolidada nas avaliações educacionais. No Brasil, é utilizada em importantes avaliações, como no Enem (Exame Nacional do Ensino Médio), na Prova Brasil (Avaliação Nacional do Rendimento Escolar), no SAEB (Sistema Nacional de Ensino Básico), no SARESP (Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo) etc., e assim muitos estudos

técnicos são realizados para verificar a adequação da aplicação da TRI em avaliações educacionais (CHILDS; OPPLER, 2000; BARBETTA; ANDRADE; BORGATTO, 2011; PRIMI et al., 2013).

Para que as proficiências dos indivíduos sejam obtidas é necessário a estimação dos parâmetros dos itens da prova aplicada e, o que se faz na maioria das vezes é determinar essa proficiência baseando-se em um teste onde o conjunto de itens foram previamente calibrados (estimados), ou seja, após serem formulados, os itens são aplicados a uma amostra moderadamente grande de indivíduos e dados os vetores de respostas, as estimativas dos parâmetros dos itens são obtidas.

Comumente, assume-se que os parâmetros dos itens são conhecidos após seus valores serem estimados através de um pré-teste. Contudo, mesmo quando o modelo assumido está corretamente especificado, existe uma incerteza na estimação dos parâmetros dos itens, uma vez que usamos uma amostra para estimá-los. Ignorar essa incerteza assumindo que os parâmetros dos itens são conhecidos pode levar a erros inferenciais das estimativas das proficiências, essencialmente quando a amostra de calibração não é suficientemente grande (TSUTAKAWA; JOHNSON, 1990). Como na realidade, grandes amostras para realizar o pré-teste dos itens podem não estar disponíveis e as leis de divulgação normalmente exigem a exposição pública dos testes, torna-se necessário a construção de mais itens. No entanto, a quantidade de pessoas disponíveis para participar da amostra de calibração e os recursos para captação dessas informações é limitado.

É de suma importância ter-se uma metodologia eficiente para estimar as habilidades de indivíduos submetidos a um teste com itens previamente calibrados, levando-se em consideração, de forma robusta, a incerteza envolvida na calibração. Este é o objetivo principal desta

dissertação. Além disso, a metodologia proposta é adaptada para o contexto de Testes Adaptativos, incluindo a proposta de um novo critério de seleção de itens e um novo critério de parada do teste.

A forma mais natural e robusta de se quantificar e considerar a incerteza em um contexto de inferência estatística é através do Paradigma Bayesiano. Esta abordagem é particularmente mais atrativa no contexto da TRI e de análises sequenciais, como é feito ao se considerar testes adaptativos. Portanto, todos os objetivos descritos anteriormente serão realizados e implementados sob o Paradigma Bayesiano, permitindo estimar de forma eficiente a habilidade dos indivíduos e quantificar de forma robusta a incerteza da estimação dos parâmetros dos itens.

1.2 TEORIA DE RESPOSTA AO ITEM

A teoria de resposta ao item (TRI) tem suas origens no trabalho pioneiro de Thurstone na década de 1920, um conjunto de autores como Lawley, Mosier e Richardson na década de 1940, e os trabalhos mais decisivos de Birnbaum, Lord e Rasch nos anos 1950 e 1960 (LINDEN, 2016).

A TRI é um conjunto de modelos matemáticos que representa a teoria psicométrica amplamente utilizada nas áreas de avaliação educacional e psicologia cognitiva. Esses modelos procuram representar a probabilidade de um indivíduo acertar um item como função dos parâmetros do item respondido e da habilidade do respondente. Essa relação é expressa de tal forma que, quanto maior a habilidade (ou seja, quanto maior o conhecimento adquirido pelo aluno em determinado assunto), maior a probabilidade de acerto no item referente àquele conteúdo.

A variável de interesse é não observável, ou seja, não pode ser medida diretamente, como por exemplo, a proficiência em determinada área do conhecimento como matemática ou nível de satisfação do indivíduo. Essa variável é denominada traço latente, proficiência ou habilidade do indivíduo no contexto de avaliação educacional (BAKER, 2001). Atualmente existem diversos modelos propostos na literatura para diferentes tipos de itens: para itens dicotômicos, ou seja, o indivíduo acerta ou não o item; para itens politômicos, em que há um certo grau dependendo da resposta escolhida, e modelos para resposta contínua, isto é, a resposta pode assumir qualquer valor em um certo intervalo.

Nessa dissertação vamos considerar apenas o modelo unidimensional de três parâmetros, onde a unidimensionalidade do modelo é referente ao traço latente que está sendo medido. Essa suposição diz que apenas uma habilidade é necessária para realizar todos os itens da prova, uma vez que como no Enem as provas são separadas por área de conhecimento, como Ciências Humanas ou Ciências da Natureza, por exemplo. Para satisfazer a essa suposição, visto que comumente mais de um traço latente é necessário para executar qualquer tarefa humana, é suficiente admitir que haja uma habilidade dominante responsável pela realização do conjunto de itens que está sendo respondido. Esta habilidade então é a que se supõe estar sendo medida pelo teste. Além de unidimensional, vamos considerar somente o modelo para itens dicotômicos ou dicotomizados (casos de teste de múltipla escolha), ou seja, aqueles que consideram duas únicas respostas possíveis para o item: a certa ou a errada.

1.2.1 O Modelo de três parâmetros

Dos modelos propostos pela TRI, o modelo unidimensional de três parâmetros é um dos mais utilizados, principalmente quando se trata de avaliações educacionais. Ele é dado por:

$$p(Y_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)F(a_i(\theta_j - b_i Y_{ij})), \qquad (1.1)$$

em que Y_{ij} é uma variável dicotômica que assume o valor 1, quando o indivíduo j responde ao item i corretamente, ou 0 quando o indivíduo j não responde corretamente ao item i, com $i = 1, ..., I$ e $j = 1, ..., J$; $\theta_j \in (-\infty, \infty)$ representa a habilidade (traço latente) do indivíduo j; $a_i \in (-\infty, \infty)$ é o parâmetro de discriminação (ou de inclinação) do item i; $b_i \in (-\infty, \infty)$ é o parâmetro de dificuldade (ou de posição) do item i; $c_i \in [0, 1]$ é o parâmetro de acerto casual do item i e F é uma função de distribuição, ou seja, F é monótona não decrescente, isto implica que estamos assumindo na prática, que quanto maior a proficiência do indivíduo, maior a probabilidade de acerto no item.

O parâmetro de discriminação mede a capacidade deste item diferenciar indivíduos com proficiências distintas, baixos valores deste parâmetro indicam que indivíduos com habilidades diferentes têm aproximadamente a mesma probabilidade de acertar o item. Já o parâmetro de dificuldade posiciona os itens ao longo da escala de proficiência (visto que a proficiência e a dificuldade do item estão na mesma escala), quanto maior a dificuldade, maior é a proficiência necessária para que o indivíduo tenha alta probabilidade de responder corretamente ao item. O parâmetro de acerto casual descreve a probabilidade mínima que todo indivíduo tem de responder ao item de forma correta, que na prática representa a probabilidade de indivíduos com baixa habili-

dade responderem corretamente o item (popularmente conhecido como chute).

Um teste ótimo é aquele que possui itens com parâmetros de dificuldade distribuídos em toda a escala, para que tenhamos informação sobre indivíduos de diferentes níveis de proficiência e todos os itens com a discriminação alta, indicando que indivíduos com habilidades diferentes têm probabilidades marcadamente distintas de responder corretamente ao item.

As escolhas mais comuns na literatura de TRI para F são: O modelo da ogiva normal (também conhecido como modelo probito) de 3 parâmetros (3PNO) , onde F é a acumulada do normal padrão, dado por:

$$p(Y_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\Phi(a_i(\theta_j - b_i)), \qquad (1.2)$$

em que $\Phi(.)$ é função distribuição acumulada da N(0,1) (função ogiva normal).

E o modelo logístico de três parâmetros (3PL), onde F é uma função logística, dada por:

$$p(Y_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + \mathrm{e}^{-Da_i(\theta_j - b_i)}}, \qquad (1.3)$$

em que D é um fator de escala, definido em geral como uma constante igual a 1, ou igual a 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal.

Em qualquer uma das duas escolhas, temos um modelo TRI de três parâmetros. Podemos obter modelos mais simples: se fixarmos $c_i = 0$, obtemos o modelo de dois parâmetros e se fixarmos $c_i = 0$ e $a_i = 1$,

obtemos o modelo de um parâmetro (BAKER; KIM, 2004; LORD, 2012).

Pode-se notar que os modelos em (1.2) e (1.3) são não identificáveis. Qualquer transformação do tipo $\theta_j^* = \theta_j + r$ e $b_i^* = b_i + r, r \in \mathbb{R}$, com $j = 1, ..., J$, levam a uma mesma probabilidade de acerto, pois $\Phi(a_i(\theta_j - b_i)) = \Phi(a_i(\theta_j^* - b_i^*))$. Uma solução bastante utilizada na prática é fixar a distribuição das proficiências, visto que, o problema da não-identificabilidade é eliminado ao definir as métricas (unidades de medida) para o traço latente. Usualmente adota-se que a proficiência segue uma distribuição normal com média e variância conhecidas. Este pressuposto estabelece que as habilidades são uma amostra aleatória dessa distribuição e, adicionalmente, isso estabelece uma métrica para as estimativas (PATZ; JUNKER, 1999; HABERMAN, 2005; GONÇALVES; DIAS; SOARES, 2018).

1.2.2 Curva característica do item

Vimos anteriormente que o modelo é definido para cada item de forma separada, dessa maneira é possível construir a curva característica para todos os itens de uma prova. A curva característica do item (CCI) é uma ferramenta gráfica usada para descrever o comportamento dos parâmetros da TRI. Pelas características dadas no modelo (1.1) já vimos que quanto maior a proficiência do indivíduo, maior a probabilidade de acertar ao item (TUCKER, 1946). Este comportamento é completamente descrito pela CCI, além disso, é possível analisar a relação existente entre os parâmetros dos itens e a relação desses parâmetros com a proficiência do indivíduo (GONÇALVES; DIAS; SOARES, 2018).

A escala da habilidade é invariante a transformações afins, onde o importante é a ordenação existente entre seus pontos que permite

posicionar os indivíduos de acordo com o nível de habilidade que ele possui, permitindo a classificação de indivíduos considerando o seu posicionamento na escala de habilidade definida.

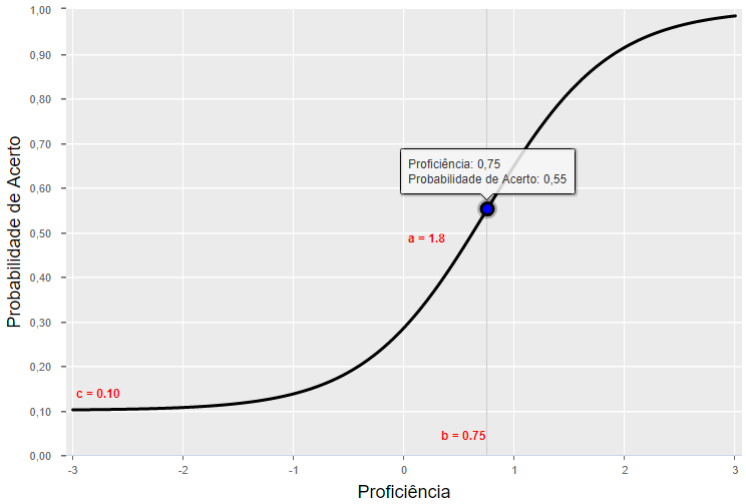

Figura 1 – Curva característica do item para o modelo logístico de três parâmetros com $a = 1{,}80$, $b = 0{,}75$ e $c = 0{,}10$.

Observe, pela Figura 1, que para o modelo logístico de três parâmetros quando a proficiência do indivíduo é igual a dificuldade do item, a probabilidade de acerto no item é $(1 + c)/2$. Se a proficiência do indivíduo é maior que a dificuldade do item ele tem mais chance de acertar do que de errar o item, veja, por exemplo, para um indivíduo que possui proficiência de 1,65 a probabilidade de ele acertar esse item é de aproximadamente 85%, visto que a dificuldade desse item é de 0,75, a discriminação é 1,80 e a chance de acerto ao acaso é 0,10.

Devido ao fato de a inclinação da curva ser definida pelo parâmetro de discriminação, temos que quanto maior a discriminação do item, maior a inclinação da curva característica e, consequentemente, maior

será a diferença entre as probabilidades de acerto de indivíduos com diferentes proficiências, ou seja, maior será a capacidade do item de diferenciar (discriminar) os indivíduos. Assim, no contexto de avaliação educacional, não são esperados itens com o parâmetro a_i negativo, uma vez que eles indicam que a probabilidade de um indivíduo acertar o item, diminui com o aumento da sua habilidade. Valores muito altos do parâmetro de discriminação indicam itens com curvas características muito inclinadas, que discriminam os alunos basicamente em dois grupos: os que possuem habilidades abaixo do valor do parâmetro b_i e os que possuem habilidades acima do valor do parâmetro b_i (TUCKER, 1946; LORD; NOVICK, 2008; ANDRADE; TAVARES; VALLE, 2000; LORD, 2012).

Note que o parâmetro b_i é um parâmetro de posição dos itens na escala. Itens com menor dificuldade possuem maior probabilidade de acerto entre os examinandos, incluindo aqueles com baixa habilidade. E itens com maior dificuldade implicam em probabilidades baixas de acerto entre os examinandos para boa parte da escala, exceto para aqueles com altos níveis de proficiência (LORD, 2012; GONÇALVES; DIAS; SOARES, 2018).

O parâmetro c_i define a assíntota inferior da CCI, e pode ser interpretado como a probabilidade de acerto ao item dos indivíduos com baixa habilidade. Por exemplo, no caso em que o item possui cinco alternativas, espera-se $c_i = 0{,}2$, indicando que o respondente fornece a resposta correta ao item escolhendo o gabarito de forma aleatória entre as alternativas apresentadas.

1.2.3 A Escala de Medida das Habilidades

Os escores brutos ou padronizados dos testes clássicos com I questões dicotômicas corrigidas como certo ou errado assumem valores inteiros entre 0 e I. Já na TRI diferentemente da teoria clássica a habilidade pode teoricamente assumir qualquer valor real entre $-\infty$ e ∞. Assim, precisa-se estabelecer, o valor médio e o desvio-padrão das habilidades dos indivíduos da população em estudo (ANDRADE; TAVARES; VALLE, 2000).

Devido à facilidade computacional, tanto a calibração de itens quanto a de habilidades é feita na escala (0,1), ou seja, numa escala com média igual a zero e desvio padrão igual a um (JÚNIOR, 2011). Em termos práticos, não existe diferença ao estabelecer estes valores ou outros quaisquer, o importante é a ordenação existente entre os pontos, de forma a definir a posição de cada item ou habilidade na escala (LORD; NOVICK, 2008). Por exemplo, nas avaliações em larga escala no Brasil, o SARESP e o SAEB, utilizam a escala (250, 50) e o ENEM utiliza a escala (500, 100). Observa-se que essas são apenas formas de representar a habilidade que tornam o entendimento mais fácil de ser interpretado, uma vez que, existe uma dificuldade em compreender os valores negativos e decimais que existem na escala (0,1), levando a interpretações inadequadas dos valores das proficiências (VALLE, 2001).

Quando estimamos os parâmetros pela escala (0,1) é fácil transformar para uma outra escala qualquer (μ, σ) e essa transformação não muda a relação de ordem entre os indivíduos e itens na escala e nem a probabilidade de resposta correta ao item do indivíduo (a habilidade do indivíduo é invariante à escala de medida) (LORD, 2012). Essa transformação pode ser obtida da seguinte forma:

- Parâmetro a: $a = \frac{a_{(0,1)}}{\sigma}$.
- Parâmetro b: $b = (\sigma b_{(0,1)}) + \mu$.
- Proficiência: $\theta = (\sigma \theta_{(0,1)}) + \mu$.

A construção da escala de habilidade é feita após a calibração e equalização dos itens possibilitando, a interpretação pedagógica dos valores das habilidades. Como na TRI os parâmetros dos itens vindos de provas distintas ou proficiências de examinandos de diferentes grupos estão em uma escala comum, os itens e/ou as proficiências são comparáveis. Dessa forma, é possível a construção de escalas de conhecimento interpretáveis, ou seja, pode-se atribuir um significado pedagógico aos valores obtidos, bem como o acompanhamento do conhecimento adquirido por alunos ao longo do tempo (ANDRADE; TAVARES; VALLE, 2000).

1.3 ESTIMAÇÃO DOS PARÂMETROS DO MODELO DA TRI

Para utilizar o modelo proposto pela TRI, temos uma etapa muito importante que é a estimação dos parâmetros dos itens e das habilidades dos respondentes. A probabilidade de um indivíduo responder corretamente a um determinado item depende da sua habilidade e dos parâmetros que caracterizam o item respondido, e em geral, ambos são desconhecidos.

Podemos dividir então o problema em três situações:

1. Quando já conhecemos os parâmetros dos itens e temos apenas que estimar as habilidades.

2. Quando conhecemos previamente as habilidades dos respondentes e estamos interessados apenas na estimação dos parâmetros dos itens.

3. Situação mais usual, em que desejamos estimar tanto os parâmetros dos itens quanto as habilidades dos indivíduos simultaneamente.

Na TRI, o processo de estimação dos parâmetros dos itens é conhecido por calibração. Consideramos uma boa calibração quando as estimativas dos parâmetros dos itens forem adequadas à área de conhecimento avaliada e seus respectivos erros padrões forem baixos.

Existe uma extensa literatura a respeito dos métodos para estimação dos parâmetros do modelo. Considerando os métodos de abordagem clássica, temos os Métodos de Máxima Verossimilhança (MMV) que não permitem a estimação simultânea dos parâmetros dos itens e habilidades, uma vez que envolve um número muito grande de parâmetros a serem estimados simultaneamente (3 parâmetros para cada item no teste mais um parâmetro de habilidade para cada respondente), levando a grandes problemas computacionais que envolvem a inversão de matrizes dessa ordem. Já na abordagem Bayesiana, a estimação pode ser feita de forma conjunta sem ignorar as fontes de incerteza. Nesse caso a estimação pontual dos parâmetros dos itens pode ser feita, por exemplo, via a Moda *a posteriori* (MAP) ou a Média *a posteriori* (EAP).

Tanto os métodos de MV quanto os métodos Bayesianos podem resultar em equações sem solução explícita, o que torna necessária a utilização de algum método numérico iterativo. Além disso, devido à dificuldade de integração das equações presentes nesses métodos, tam-

bém é muito comum o uso dos métodos de integração numérica como o de quadratura Gaussiana, o qual consiste em aproximar as integrais que não apresentam solução analítica através de retângulos.

Na situação em que desejamos estimar tanto os parâmetros dos itens, quanto as habilidades dos indivíduos, há duas abordagens mais comuns: estimação conjunta dos parâmetros dos itens e habilidades, ou em duas etapas, primeiro a estimação dos parâmetros dos itens e, posteriormente, das habilidades.

Nosso trabalho considera o contexto onde os parâmetros dos itens foram previamente estimados via método MCMC *(Markov Chain Monte Carlo)* e, portanto, uma amostra (aproximada) da distribuição *a posteriori* dos parâmetros dos itens está disponível. Não entraremos em detalhes sobre o processo de calibração dos parâmetros dos itens, uma vez que a teoria de MCMC é extensa e complexa. Uma discussão mais geral sobre o tema pode ser encontrada por exemplo em Gamerman e Lopes (2006) e Robert e Casella (2013). Como o objetivo do nosso trabalho é encontrar uma solução para o problema de estimar a habilidade de respondentes que não fizeram parte da amostra de calibração, sem ignorar a incerteza existente desse processo e sem optar pelo retrabalho de calibrar novamente os parâmetros dos itens utilizando os novos vetores de respostas, partiremos do pressuposto de que tem-se um teste construído com itens previamente calibrados via Inferência Bayesiana, e utilizaremos a amostra (aproximada) da distribuição *a posteriori* dos parâmetros dos itens para estimar, sob o paradigma Bayesiano, a habilidade de novos respondentes submetidos a este teste.

1.3.1 Métodos de Estimação das Habilidades

Nesta seção vamos tratar da estimação das habilidades quando os parâmetros dos itens já foram estimados. Na prática, essa situação ocorre quando os itens já foram calibrados através de uma amostra de respondentes. Como a calibração dos itens deve ser feita com um número suficientemente grande de indivíduos, a estimação das habilidades de um grupo pequeno de indivíduos, por exemplo, deve ser feita utilizando itens já calibrados. Os métodos mais utilizados são os métodos de Máxima Verossimilhança (MV) e os métodos Bayesianos. O método de MV consiste em encontrar os valores da proficiência considerando os parâmetros dos itens conhecidos que fazem com que a probabilidade d indivíduo ter dado a resposta observada seja a maior possível. Contudo, esse método possui várias limitações como ausência de solução explícita para θ_j, nem sempre existe um único máximo da função de verossimilhança, o método não está definido para os padrões de respostas constantes dos respondentes (indivíduos que acertam ou erram todos os itens respondidos) e além disso a utilização do método MV produz um viés na estimação de valores altos e baixos da habilidade: valores altos são superestimados e valores baixos são subestimados (SAMEJIMA, 1973; KIM; NICEWANDER, 1993). Os métodos Bayesianos serão apresentados com mais detalhes a seguir.

1.3.1.1 Métodos Bayesianos

Nos métodos Bayesianos a incerteza é descrita através de probabilidade, com isso além de serem mais intuitivos, conseguem contornar problemas que acontecem nos procedimentos de máxima verossimilhança,

tais como: problemas de estimação dos parâmetros dos itens e das proficiências onde todos os indivíduos respondem corretamente (ou incorretamente) o item, estimativas fora do esperado como discriminação negativa (considerando os casos em que os itens estão bem construídos e não possuem problemas na sua formulação), entre outros.

Além dos dados amostrais, na inferência Bayesiana é necessário a utilização de uma distribuição *a priori* sobre os parâmetros de interesse, onde essa distribuição representa o conhecimento do pesquisador sobre esse parâmetro. Seja Ψ um escalar, matriz ou vetor que representa os parâmetros de interesse. A distribuição *a priori* é definida pela densidade de probabilidade $\pi(\Psi)$.

A distribuição *a posteriori* é denotada por: $\pi(\Psi|x)$. Ela contém toda a informação probabilística de interesse a respeito de Ψ, onde x representa os dados amostrais. Por esse motivo a distribuição *a posteriori* é usada para se fazer inferências sobre os parâmetros desconhecidos.

A distribuição *a posteriori* é obtida pelo Teorema de Bayes. Construímos:

$$\pi(\Psi|x) = \frac{\pi(x|\Psi)\,\pi(\Psi)}{\pi(x)}, \tag{1.4}$$

onde $\pi(x)$ é a distribuição marginal de x.

Na maioria das vezes, é inviável obter-se a forma analítica dessa distribuição, especialmente pela dificuldade em se obter a distribuição marginal dos dados.

Neste cenário, métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC) tem sido amplamente utilizados. A ideia básica deste método é que se conhecemos o núcleo de $\pi(\Psi|x)$, então é possível obter uma amostra (aproximada) da distribuição *a posteriori* usando teoria de cadeias de Markov. Para isso, estipula-se um *burn-in* que é

um número de iterações que se julgue serem necessárias para a convergência da cadeia, e, a partir daí, gera-se uma amostra suficientemente grande para obter-se uma boa aproximação da distribuição de interesse. Com uma amostra da distribuição *a posteriori* de Ψ, características de $\pi(\Psi|x)$ podem ser estudadas empiricamente através de técnicas descritivas. Mais informações sobre o procedimento podem ser encontradas em (BÉGUIN; GLAS, 2001; GAMERMAN; LOPES, 2006; CARLO, 2004).

A partir disso entre os métodos Bayesianos mais utilizados para estimação das habilidades dos respondentes considerando os parâmetros dos itens fixos, são:

- **Estimação pela moda *a posteriori* – MAP:** a estimação pela moda *a posteriori* (ou MAP: *maximum a posteriori*) consiste em encontrar estimativas pontuais que maximizam a distribuição *a posteriori* com respeito aos parâmetros dos itens. Diferentemente do Estimador de Máxima Verossimilhança (EMV), o procedimento de estimação MAP sempre converge independentemente do padrão de resposta dos indivíduos (MISLEVY; STOCKING, 1989). Uma vez que a equação deste método de estimação não possui solução explícita, é necessário a utilização de algum método iterativo para resolvê-la.

- **Estimação pela média *a posteriori* – EAP**: a estimação pela média d*a posteriori* (ou EAP: expected *a posteriori*) consiste em obter a esperança da distribuição *a posteriori*. Este método possui integrais sem solução explícita, contudo essas integrais podem ser diretamente calculadas sem a necessidade de utilização de métodos iterativos ao utilizar os métodos de Quadratura Gaussiana, visto que dado os pontos de quadratura, não é necessário calcular

> as integrais (ANDRADE; TAVARES; VALLE, 2000). O que o torna bastante vantajoso do ponto de vista computacional.

Vale ressaltar que na abordagem Bayesiana é possível fazer a estimação conjunta das habilidades dos respondentes e dos parâmetros dos itens através de um algoritmo MCMC que amostra da distribuição *a posteriori* conjunta das quantidades desconhecidas do modelo. Podemos gerar amostras aproximadas da distribuição conjunta dos parâmetros de interesse *a posteriori*, a partir das distribuições condicionais completas *a posteriori* de cada parâmetro, por exemplo. Para mais detalhes, ver Gonçalves, Dias e Soares (2018). Entretanto, o que esses métodos geralmente fazem, assim como os métodos clássicos, é dividir esse processo em duas etapas, como proposto por Bock e Lieberman (1970). Estes métodos baseiam-se na existência de uma distribuição (latente) associada à habilidade dos indivíduos da população em estudo. Isso possibilita que a estimação dos itens seja feita considerando uma determinada distribuição para a habilidade dos indivíduos e após a estimação dos parâmetros dos itens, as habilidades são estimadas individualmente pela moda ou média da distribuição condicional, considerando os parâmetros dos itens fixos (ANDRADE; TAVARES; VALLE, 2000).

Apesar da possibilidade da estimação conjunta dos parâmetros dos itens e das habilidades dos indivíduos, que leva em consideração toda a variabilidade do problema, na prática isso não é feito e os métodos propostos na literatura, para a estimação em duas etapas, desconsideram a incerteza do processo de calibração dos itens, tratando-os como fixos ao estimar as habilidades dos indivíduos.

1.4 TESTES ADAPTATIVOS INFORMATIZADOS

Nas últimas décadas o uso do computador tornou-se imprescindível no cotidiano da população, e o uso da tecnologia tornou-se fundamental nos mais diversos setores de atividades. Na educação, por exemplo, existe uma grande preocupação em inovar o processo de aprendizagem e em investir em novas tecnologias dentro das salas de aula, bem como criar tecnologias mais sofisticadas de forma a melhorar a medição e a precisão das proficiências adquiridas pelos alunos ao longo do tempo.

Atualmente existe uma variedade enorme de conteúdos educacionais, bem como testes para avaliar os conhecimentos adquiridos pelos respondentes administrados de forma online, apresentados como alternativa para as avaliações do tipo "papel e caneta". Muitas avaliações nacionais e internacionais já utilizam os chamados testes informatizados, uma vez que é uma iniciativa com grandes vantagens, tais como a criação de itens em formatos multimídias, a verificação automática que reduz o tempo de correção dos testes, eliminação da possibilidade de erros de transcrição de gabarito, diminuição dos gastos com material impresso, além de permitir a realização da prova em diferentes dias e horários para candidatos de diferentes lugares.

Uma das implementações mais elegantes no campo da avaliação informatizada são os denominados Testes Adaptativos Informatizados. Um Teste Adaptativo Informatizado – TAI é um teste baseado na TRI onde os itens são administrados e apresentados pelo computador, em que o teste procura apresentar apenas itens adequados a habilidade do indivíduo que o realiza, ou seja, ele apresenta o teste ótimo para cada respondente. Nesse tipo de teste a proficiência do indivíduo é estimada de forma iterativa, no qual são selecionados somente os itens

que mensuram de forma eficiente a proficiência do indivíduo. O objetivo é buscar uma melhor estimação da habilidade do indivíduo junto com a redução do número de perguntas que precisam ser respondidas.

Geralmente, os itens destes testes são selecionados de acordo com o modelo da TRI. Como é um teste personalizado, diferentes respondentes podem receber diferentes testes de tamanhos variados. Como citado por Wainer et al. (2000), a ideia de um teste adaptativo é imitar automaticamente o que um examinador faria, a medida que um indivíduo erra os itens apresentados, escolhe-se itens mais fáceis que os aplicados anteriormente e a medida que um indivíduo acerta os itens apresentados escolhe-se itens mais difíceis para serem aplicados. Esse autor ainda destaca que o resultado traz uma medição mais precisa da proficiência, além da redução do tamanho do teste (geralmente em 50%).

As primeiras pesquisas sobre testes adaptativos computacionais foram realizadas nas décadas de 70 por Lord (1971) e Owen (1975). Desde então diversos testes adaptativos informatizados têm sido implementados, como por exemplo o Graduate Record Examination (GRE).

Para ilustrar como funciona o processo adaptativo desses testes a Figura 2 apresenta um teste hipotético com cinco itens. Como no início do teste não existe nenhuma informação sobre o nível de habilidade do respondente, no caso em que ele realiza a prova pela primeira vez, usualmente assume-se como nível inicial a proficiência média ($\theta = 0$) considerando a escala (0,1). Inicialmente um item de dificuldade média $b_i = 0$ é escolhido e administrado. Suponha que o indivíduo responda corretamente ao primeiro item. Dessa forma, a estimativa da habilidade é atualizada, e como neste caso foi um acerto a habilidade estimada é aumentada para $\theta = 0{,}6$ e um segundo item mais difícil é escolhido

para ser administrado.

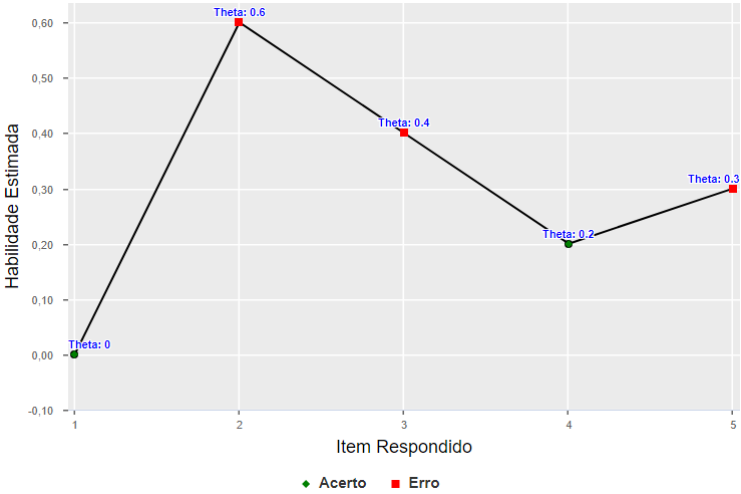

Figura 2 – Teste adaptativo hipotético com cinco itens.

Suponha agora que o indivíduo erre esse item indicando que ele não possui o conhecimento necessário, ou seja, o item é difícil para o seu nível de proficiência. Assim, o computador atualiza a estimativa da habilidade e ela diminui para $\theta = 0{,}4$. O próximo item a ser administrado será mais difícil que o primeiro item apresentado e mais fácil que o segundo, considerando a última estimativa da proficiência do indivíduo. Se o indivíduo também responde incorretamente a este item, a estimativa da habilidade ao ser atualizada diminui novamente e é estimado no valor de $\theta = 0{,}2$. O quarto item será escolhido de tal maneira que seja ainda mais fácil que o terceiro item. Se o examinando responder corretamente a este item, a estimativa de sua habilidade aumentará para $\theta = 0{,}3$ e um item mais difícil para esse nível de habilidade será apresentado como o último item do teste adaptativo.

Nota-se que a cada acerto do indivíduo a sua estimativa da habi-

lidade aumenta e a cada erro essa estimativa diminui. Esse processo de seleção e administração de itens e atualização das estimativas das proficiências, são feitas iterativamente até que algum critério de parada seja satisfeito, como por exemplo, um número máximo de itens a serem respondidos ou o desvio padrão *a posteriori* da habilidade ser menor que um valor pré-estabelecido.

A utilização da TRI em testes adaptativos apresenta uma grande vantagem devido ao fato de que é possível a criação de uma escala de proficiência, possibilitando que tanto os itens quanto as habilidades dos indivíduos sejam colocadas em uma mesma métrica. Essa propriedade é muito importante, uma vez que, embora cada estudante possa responder a diferentes itens, os resultados são comparáveis entre si.

A TRI pode estar presente em todas as fases de um TAI. Desde a construção do banco de itens, uma vez que ela permite avaliar as características desses itens por meio da estimação dos parâmetros (discriminação, dificuldade e acerto casual), até a fase de administração do teste, onde está envolvida na escolha do item a ser administrado, na estimação iterativa da proficiência do indivíduo e no critério de parada do teste.

Podemos então escrever uma estrutura geral de um teste adaptativo. A maioria dos TAIs utiliza uma estratégia que necessita estabelecer:

- Um critério de partida, para determinar o primeiro item a ser apresentado. Como normalmente não temos informação sobre a habilidade do indivíduo, considera-se um nível de proficiência média $\theta = 0$, na escala (0,1).

- Um método estatístico (Bayesiano ou Clássico) para estimar a

proficiência do indivíduo e a precisão associada.

- Um procedimento para selecionar o próximo item.
- Um critério de parada para finalizar o teste.

1.4.1 Métodos de seleção adaptativa dos itens

Um dos componentes essenciais dos testes adaptativos consiste nos procedimentos de seleção dos itens ao longo do teste. Para que um teste adaptativo seja aplicado é necessário que os parâmetros dos itens sejam previamente estimados e dessa forma conhecemos o nível de dificuldade de todos os itens do banco, permitindo o desenvolvimento de um algoritmo para seleção de itens.

Segundo Rudner (1998), em geral, a seleção dos itens é feita por um algoritmo formado por um processo iterativo com os seguintes passos:

1. Todos os itens que ainda não foram exibidos são avaliados para verificar qual será o próximo item a ser apresentado, dado o nível de habilidade do respondente atualmente estimado.
2. O próximo item é disponibilizado e o indivíduo responde.
3. Uma nova estimativa da habilidade do indivíduo é calculada baseada nas respostas de todos os itens solucionados até então.
4. Os passos 1, 2 e 3 são repetidos até que o critério de parada seja alcançado.

Se forem utilizados itens inadequados ao respondente, será necessária uma maior administração de itens para obter o mesmo resultado caso fossem utilizados apenas itens adequados (WIBERG, 2003). Isso

quer dizer que apresentar itens ao candidato extremamente fáceis ou difíceis considerando a sua habilidade estimada, apenas tem um impacto negativo com relação ao número de itens administrados, uma vez que estes itens não acrescentam informações significativas a sua habilidade.

1.4.1.1 Critério de Máxima Informação

Esse método consiste em selecionar o próximo item com base na medida de Informação de Fisher (IF) avaliada na proficiência atual estimada, também conhecida como Função de Informação Local (CHANG; YING, 1996).

A Informação de Fisher é um dos conceitos mais conhecidos na literatura estatística. Ela é usada para mensurar o grau de informação que uma variável aleatória observável Y carrega sobre um parâmetro desconhecido θ. Define-se a Informação de Fisher da seguinte maneira:

Definição 1.4.1 (Informação de Fisher): Seja Y um vetor de variáveis aleatórias com função densidade de probabilidade $f(Y|\theta)$, sendo esta a função de verossimilhança para θ que descreve a probabilidade de observarmos uma amostra Y, dado um valor conhecido de θ. A Informação de Fisher esperada de θ através de Y é dada por:

$$IF_Y(\theta) = -E_{Y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log(f(Y|\theta))\right]. \tag{1.5}$$

A Informação de Fisher (IF) é o valor médio da curvatura da verossimilhança. Quanto maior é esta curvatura, maior é a informação sumarizada na função de verossimilhança e, consequentemente, maior será o valor de $IF(\theta)$.

Na TRI, a Informação de Fisher permite analisar o quanto um item contém de informação para a medida de habilidade θ. Ela é calculada individualmente para cada item a partir dos seus respectivos parâmetros estimados.

Para o modelo probito de três parâmetros (ver Apêndice A), a Informação de Fisher é dada por:

$$\begin{aligned} IF_{Y_i}(\theta) &= \frac{[P_i'(\theta)]^2}{P_i(\theta)[1-P_i(\theta)]} \\ &= \frac{[(1-c_i)a_i\phi(a_i\theta_j - b_i^*)]^2}{[c_i+(1-c_i)\Phi(a_i\theta_j - b_i^*)][1-c_i-(1-c_i)\Phi(a_i\theta_j - b_i^*)]} \end{aligned} \tag{1.6}$$

em que $P_i(\theta)$ e $P_i'(\theta)$ é o modelo probito de três parâmetros apresentado em (1.2) reparametrizado e a sua primeira derivada, respectivamente e $b_i^* = a_i b_i$, $i = 1, ..., I$.

Sob o modelo da TRI, maximizar a Informação de Fisher significa selecionar um item de dificuldade que seja compatível com o nível de proficiência do respondente. Além disso itens com maior discriminação são preferencialmente selecionados pelo algoritmo, uma vez que quanto maior é o valor do parâmetro a_i, maior é a diferença entre as probabilidades de resposta correta de dois indivíduos com habilidades distintas (COSTA, 2009).

Nos TAIs, a IF serve como referência para seleção de itens quando existe conhecimento suficiente sobre a proficiência do respondente. O algoritmo de seleção de máxima informação usa a estimativa pontual da habilidade. Defina $\hat{\theta}_{i-1}$ como o valor estimado de θ após $i-1$ respostas. O i-ésimo item selecionado por este método será o que possuir o valor máximo da IF avaliada em $\hat{\theta}_{i-1}$ (proficiência estimada no passo

anterior) (COSTA, 2009).

1.4.1.2 Critério da Máxima Informação Esperada

O Critério da Máxima Informação Esperada (MIE), é um dos procedimentos Bayesianos mais utilizados em testes adaptativos para seleção de itens. Os testes adaptativos são naturalmente ajustados por uma abordagem Bayesiana empírica ou sequencial. Onde a distribuição *a posteriori* de θ, estimada após $i-1$ itens respondidos, pode ser usada para selecionar o próximo item a ser administrado e então considerada como distribuição *a priori* para a obtenção da próxima distribuição *a posteriori* (COSTA, 2009).

O método MIE baseia-se na análise preditiva. Em Estatística, a análise preditiva consiste em fazer inferências probabilísticas sobre uma quantidade a ser observada no futuro (MIGON; GAMERMAN; LOUZADA, 2014). Nos TAI, deseja-se prever a resposta aos itens ainda não administrados no teste depois de $i-1$ respostas e, então, escolher o próximo item de acordo com as atualizações da distribuição *a posteriori* de θ, da estimativa pontual da proficiência $\hat{\theta}$, e da variância *a posteriori* de θ (COSTA, 2009).

Como destaca Linden (1998), se o i-ésimo item é selecionado, respostas para os $i-1$ itens já são conhecidas. Dessa forma, os dados não podem ser considerados como variáveis aleatórias, eles devem ser tratados como valores (fixos) da realização dessa variável aleatória. A escolha do próximo item pelo critério MIE que será administrado levará em conta a medida de IO dos itens no ponto $\hat{\theta}$:

$$J_{y_1,\ldots,y_{i-1}}(\theta) = \left. \frac{-\partial^2 \log L(\theta; y_1, \ldots, y_{i-1})}{\partial \theta^2} \right| \theta = \hat{\theta}, \tag{1.7}$$

que reflete a curvatura da função de verossimilhança observada para o θ relativo à métrica escolhida. Devido ao fato de que a estimativa pontual $\hat{\theta}$ é atualizada de forma iterativa, a medida de IO não deve ser somente atualizada para a resposta ao i-ésimo item selecionado, mas sim, para todas as respostas anteriores, ou seja, para $y_1, ..., y_{i-1}$.

Esse critério prevê a distribuição de probabilidade das respostas do indivíduo em cada item e seleciona o item com a informação máxima esperada sobre essa distribuição de probabilidade. Para cada novo item administrado, a distribuição de Y_i é considerada como a distribuição preditiva *a posteriori* após serem obtidas as respostas aos itens anteriores. Para mais detalhes sobre o método, ver Linden (1998) e Costa (2009).

Existem outros critérios de seleção adaptativa de itens propostos na literatura. Contudo, o entendimento da motivação e proposta dos métodos apresentados, são suficientes para entender o método de seleção de itens proposto no próximo capítulo. Desta forma, outros métodos abordados na literatura não serão apresentados neste trabalho.

1.4.2 Critério de parada

Outro componente essencial dos testes adaptativos consiste na seleção do critério de parada do teste, uma vez que esse critério determina até quando os itens serão aplicados ao respondente. Uma importante característica que deve ser previamente estabelecida é qual tipo de teste vai ser aplicado aos respondentes: testes de tamanho fixo ou tamanho variável.

Os testes de tamanho fixo não são muito recomendados, uma vez que todos os indivíduos respondem a mesma quantidade de itens e,

consequentemente, a precisão da habilidade estimada não será a mesma para todos os indivíduos. Para ter a mesma precisão, indivíduos com diferentes proficiências necessitam de quantidades diferentes de itens a serem respondidos.

Nos testes de tamanho variado, assumindo que o banco de itens tenha um número suficiente de itens distribuídos em toda escala de proficiência, um teste adaptativo pode ser finalizado de acordo com Linacre (2000) quando: o desvio padrão a posteriori for menor que um valor pré-estabelecido. Isso indica que a habilidade foi estimada com precisão suficiente e assim o teste é interrompido. Caso o objetivo do teste seja classificar um indivíduo como aprovado ou reprovado, o teste pode ser interrompido quando a habilidade estimada estiver seguramente longe do ponto de corte, com uma distância de pelo menos dois desvios padrão ou quando não houver mais itens suficientes para o indivíduo alcançar o ponto de corte.

Vale ressaltar que todos os critérios apresentados de seleção de itens e de parada do teste, são baseados nas estimativas pontuais dos parâmetros dos itens, ou seja, esses métodos não consideram a incerteza presente no processo de calibração, tratando os itens como conhecidos após serem calibrados. Portanto, para o contexto apresentado neste trabalho, vamos propor uma nova metodologia para realizar o processo de seleção de itens no decorrer do TAI e uma nova metodologia como critério de parada do teste, ambos levando em consideração a incerteza existente no processo de calibração dos itens aplicados aos respondentes.

2 CONSIDERANDO A INCERTEZA ASSOCIADA À CALIBRAÇÃO DOS PARÂMETROS DOS ITENS NA TRI

Uma prática comum nos métodos convencionais da TRI é fazer a estimação dos parâmetros do modelo em duas etapas, estima-se em primeira etapa os parâmetros dos itens e em seguida o traço latente dos indivíduos. Desta forma, assumir que os parâmetros dos itens são conhecidos, ignorando essa incerteza pode levar a erros inferenciais das estimativas dos traços latentes, particularmente quando a amostra de calibração não é suficientemente grande (TSUTAKAWA; JOHNSON, 1990).

Este trabalho considera o problema de se incorporar a incerteza associada à estimação dos parâmetros dos itens na estimação das proficiências. O objetivo principal é propor uma metodologia que incorpore esta incerteza e que seja computacionalmente eficiente. Além disto, esta metodologia é trabalhada para ser utilizada no contexto de testes adaptativos, incluindo a proposta de um novo critério de seleção de itens e um novo critério de parada do teste.

A modelagem e a proposta de incluir a incerteza nas estimativas dos parâmetros dos itens na estimação das habilidades foi discutida anteriormente por Tsutakawa e Johnson (1990). A proposta desses autores para o 3PL é construir a distribuição *a posteriori* da proficiência em termos dos dados $z = (x, y)$, onde x é o vetor de respostas de um novo indivíduo, y são os dados da calibração e $\xi = (a, b, c)$ são os parâmetros desconhecidos dos itens aplicados no teste. Assim é possível obter através de uma abordagem Bayesiana a função densidade

de probabilidade dada por $\int_{\xi} p(\theta|z,\xi)p(\xi|z)d\xi = p(\theta|z)$.

Uma vez que as expressões da média e variância *a posteriori* da habilidade são difíceis de serem obtidas o artigo sugere o cálculo desses valores através de aproximações e da decomposição da variância *a posteriori*. Essa aproximação é um caso especial da aproximação de Lindley (1980) para a média *a posteriori* de uma função de hiperparâmetros, e nesse caso a distribuição de ξ é normal. A principal conclusão do artigo é que, quando há incerteza nos parâmetros dos itens, tanto a máxima verossimilhança quanto o Bayes empírico subestimam a variância da habilidade e, portanto, produzem estimativas intervalares que são muito estreitas e enganosas.

2.1 MODELO UTILIZADO

Iremos considerar neste trabalho o modelo dicotômico de 3 parâmetros, em particular o 3PNO, contudo o modelo pode estendido para outros contextos. Considera-se:

$$P_{\xi_i}(\theta_j) = c_i + (1 - c_i)\Phi(a_i\theta_j - b_i^*), \tag{2.1}$$

onde $b_i^* = a_i b_i \ \forall i = 1, ..., I$ e $\xi_i = (a_i, b_i, c_i)$ são os parâmetros do item i. $P_{\xi}(\theta_j)$ é a probabilidade de um indivíduo j com habilidade θ_j responder corretamente ao item i. Esta é chamada de Função de Resposta do Item.

Diferentemente do trabalho de Tsutakawa e Johnson (1990) que utilizam o modelo 3PL, utilizamos o 3PNO, onde o objetivo da parametrização apresentada em (2.1) é facilitar o MCMC para calibração dos parâmetros dos itens via Inferência Bayesiana, ver Gonçalves, Dias e Soares (2018).

Assumiremos neste trabalho que um teste composto por I itens já foi aplicado a uma amostra de calibração de J indivíduos e os parâmetros dos itens foram estimados através de uma abordagem Bayesiana via MCMC. Portanto, temos uma amostra (aproximada) de tamanho N da distribuição *a posteriori* dos parâmetros dos itens, ver Gonçalves, Dias e Soares (2018), para detalhes sobre o algoritmo MCMC.

Considere um teste de I itens, tal que X_{ij} é a resposta ao item i do aluno j e os itens são corrigidos de forma dicotômica:

$$\begin{cases} X_{ij} = 1, \text{se a resposta dada pelo aluno j ao item i estiver correta,} \\ i = 1, ..., I. \\ X_{ij} = 0, \text{caso contrário.} \end{cases}$$

Assume-se independência local do modelo, o que significa que, condicionado nas habilidades e nos parâmetros dos itens, todas as respostas são independentes, entre respondentes e itens (LORD; NOVICK, 2008).

Portanto, temos que a probabilidade do vetor de resposta $X_i = (x_1, ..., x_I)$ para um indivíduo com habilidade θ_j é dada por:

$$p_\xi(X_i|\theta_j) = \prod_{i=1}^{I} (p_{\xi_i}(\theta_j))^{x_i} (1 - p_{\xi_i}(\theta_j))^{1-x_i}. \tag{2.2}$$

Como a calibração dos itens é feita no passo anterior à estimação das habilidades dos indivíduos, o principal problema a ser abordado neste trabalho é a estimativa da habilidade de um novo indivíduo com vetor de resposta X, quando recebemos Y, os dados obtidos através da amostra utilizada para calibração dos itens.

Portanto temos que X é o vetor de respostas do indivíduo a ter a proficiência estimada e Y os vetores de resposta obtidos pela amostra de calibração dos itens.

A identificação dos parâmetros do modelo será feita fixando-se uma distribuição Normal padrão para as proficiências dos alunos no processo de calibração.

2.2 METODOLOGIA PROPOSTA

Sob o Paradigma Bayesiano, a proficiência de um novo aluno respondente do teste é estimada por sua distribuição *a posteriori*, ou seja, para o aluno j, o objetivo é se obter a distribuição condicional de $(\theta\,|\mathrm{x}\,, y)$. Assim, como demonstrado em Tsutakawa e Johnson (1990), temos que:

$$p(\theta|\mathrm{x,y}) = \int p(\theta|\mathrm{x,y},\xi)p(\xi|\mathrm{x,y})d\xi, \tag{2.3}$$

no qual, da independência condicional de x e y dado ξ (ver Apêndice B), temos:

$$p(\xi|\mathrm{x,y}) = \frac{p(x|\xi)p(\xi|\mathrm{y})}{p(x|y)}, \tag{2.4}$$

$$p(\theta|\mathrm{x,y},\xi) = \frac{p_\xi(x|\theta)\phi(\theta)}{p(x|\xi)}. \tag{2.5}$$

O termo $\phi(\theta)$ representa a distribuição fixada inicialmente para θ definida como uma N(0,1) e $(x|\theta)$ é a verossimilhança obtida por (2.2).

Substituindo (2.4) e (2.5) em (2.3) temos:

$$p(\theta|\mathrm{x,\,y}) = \frac{\phi(\theta)}{p(x|y)} \int p_\xi(x|\theta)p(\xi|\mathrm{y})d\xi. \tag{2.6}$$

O processo de calibração fornece uma amostra (aproximada) da distribuição de $(\xi|y)$, o que nos permite aproximar a integral em (2.6) via Monte Carlo (MC). O único termo desconhecido em (2.6) é $p(x|y)$ sendo uma constante de normalização.

A metodologia proposta consiste em aproximar pontualmente a integral da densidade em (2.6) via MC e utilizar a quadratura Gaussiana (ver Apêndice C) para estimar a constante de normalização da densidade e os seus momentos. Vale ressaltar que através da utilização da quadratura gaussiana podemos usar diferentes funções e calcular qualquer momento desejado da variável aleatória.

Definição 1.4.3 (Métodos de Monte Carlo): A ideia geral por trás de métodos de MC é que características de uma distribuição podem ser eficientemente aproximadas se dispusermos de uma amostra suficientemente grande desta ou de outra distribuição. A base dos métodos é a técnica de integração clássica de MC, que consiste em avaliar a integral:

$$\mathbb{E}_f[h(X)] = \int_{\chi} h(x)f(x)dx. \tag{2.7}$$

a partir de uma amostra $(X_1, ..., X_M)$ arbitrariamente grande, gerada de f (função densidade de probabilidade da variável aleatória X) .

Dessa forma, o estimador de MC de (2.7) é dado pela média empírica:

$$\hat{h}_M = \frac{1}{M}\sum_{j=1}^{M} h(X_j), \tag{2.8}$$

uma vez que h seja integrável ($\mathbb{E}_f[h(X)]$ é finita), o estimador (2.8) converge quase certamente para (2.7), pela Lei dos Grandes Números.

Note que, diferentemente dos métodos existentes, a nossa abordagem envolve apenas erro de Monte Carlo e quadratura unidimensional.

Além disso, a metodologia proposta permite aproximar a densidade *a posteriori* marginal de θ_j e, a partir desta, obter estimativas da esperança de diferentes funções, por exemplo, média, variância, quantis e etc.

2.2.1 Estimação da proficiência

A média e variância da distribuição *a posteriori* de θ_j são dadas por:

$$\mu_\theta = \hat{\theta} = E(\theta|x, \mathrm{y}) = \int \theta p(\theta|x, y)\partial\theta, \tag{2.9}$$

$$\hat{\sigma}^2_\theta = \mathrm{V}(\theta|x, \mathrm{y}) = \int (\theta - \hat{\theta})^2 p(\theta|x, y)\partial\theta. \tag{2.10}$$

Assumindo que a amostra *a posteriori* dos parâmetros dos itens está disponível, temos a seguinte matriz:

$$\begin{bmatrix} (a_{11}, b_{11}, c_{11}) & (a_{12}, b_{12}, c_{12}) & \dots & (a_{1I}, b_{1I}, c_{1I}) \\ (a_{21}, b_{21}, c_{21}) & (a_{22}, b_{22}, c_{22}) & \dots & (a_{2I}, b_{2I}, c_{2I}) \\ \vdots & \vdots & \ddots & \vdots \\ (a_{N1}, b_{N1}, c_{N1}) & (a_{N2}, b_{N2}, c_{N2}) & \dots & (a_{NI}, b_{NI}, c_{NI}) \end{bmatrix},$$

onde I é o número de itens calibrados para o teste e N é o tamanho da amostra *a posteriori* dos parâmetros obtidos através da calibração via Inferência Bayesiana. Considere θ_k, $k = 1, ..., K$, sendo k cada ponto de quadratura definido. A metodologia proposta é implementada da seguinte maneira:

Passo 1: Dado o novo vetor de respostas $X = (x_1, ..., x_I)$ calcule a probabilidade desse vetor para um indivíduo com habilidade θ_k através

da equação (2.2):

$$p_k = \left[p_{\xi_1}(x|\theta_k) p_{\xi_2}(x|\theta_k) \dots p_{\xi_N}(x|\theta_k) \right] \tag{2.11}$$

Passo 2: Aproxime a integral:

$$\int p_\xi(x|\theta_k) p(\xi|y) d\xi = \mathrm{E}\left(p_\xi(x|\theta_k)\right) \approx \frac{\sum_{n=1}^{N} p_{\xi_n}(x|\theta_k)}{N} = \hat{H}_\xi \tag{2.12}$$

Passo 3: Obtenha o valor da densidade *a posteriori* não normalizada. Repita os passos 1-3 para todos os $\theta_k's$.

Passo 4: Normalize e calcule os momentos desejados de $p(\theta|x, y)$ através de quadratura (Gaussiana).

2.3 ESTUDOS DE SIMULAÇÃO

Nesta seção apresentaremos um estudo de simulação para avaliar a eficiência da metodologia proposta. Neste estudo foram considerados 2000 indivíduos, onde cada um deles respondeu a 40 itens. A estimação dos parâmetros dos itens foi feita pelo algoritmo proposto por Gonçalves, Dias e Soares (2018).

Os valores reais das proficiências dos indivíduos foram geradas a partir de uma distribuição $N(0, 1)$. A dificuldade, discriminação e acerto casual dos itens foram gerados a partir das distribuições, $U(-3, 3)$, $U(0.8, 2.2)$ e $U(0.0, 0.2)$, respectivamente.

As distribuições *a priori* foram determinadas com o objetivo de facilitar a derivação do algoritmo MCMC e as escolhas foram as seguintes: $\theta_j \sim N(0, 1)$, $a_i \sim N_{(0,\infty)}(1, 3^3)$, $b_i \sim N(0, 4^2)$, $c_i \sim Beta(4, 12)$.

A justificativa teórica para adotar estas *distribuições a priori* são:

- Na prática, no contexto de avaliação educacional os valores de a_i

em geral, devem ser positivos, (considerando que os itens foram bem construídos), sugerindo dessa forma que a distribuição de a_i pode ser modelada por uma distribuição unimodal com assimetria positiva (MISLEVY, 1986).

- O valore do parâmetro b_i é medido na mesma escala da habilidade, podendo assumir qualquer valor na escala da proficiência.

- O parâmetro c_i é definido por uma probabilidade e portanto, seu valor deve pertencer ao intervalo [0;1] (SWAMINATHAN; GIFFORD, 1986).

Foi gerada uma amostra da distribuição *a posteriori* dos parâmetros dos itens de tamanho 50000 para cada parâmetro com um *burn-in* de 20000, dessa forma, o estudo é realizado com uma amostra de Monte Carlo de tamanho 30000.

Para fins de validação da metodologia proposta consideramos indivíduos que estavam na amostra de calibração e tiveram a sua habilidade estimada conjuntamente com a estimação dos parâmetros dos itens via MCMC.

Escolhemos 14 indivíduos da amostra de 2000 respondentes com diferentes níveis de proficiência para aplicar a metodologia proposta. Considerando que a escala de habilidade é definida por uma $N(0,1)$ temos os indivíduos escolhidos desde a menor proficiência encontrada no banco de dados -2,746 até a maior 2,408.

A metodologia proposta foi implementada no *software* R (Core Team, 2019), com a amostra completa de tamanho 30000 da cadeia dos parâmetros dos itens. Essa estimação foi feita através da quadratura Gaussiana com um *grid* de 2000 pontos variando de -6 a 6.

Tabela 1 – Comparação da média e desvio padrão *a posteriori* obtidos via MCMC e pela metodologia proposta

Respondente	Estimativa MCMC		Metodologia Proposta		Diferença em valor absoluto	
	Proficiência	Desvio Padrão	Proficiência	Desvio Padrão	Proficiência	Desvio Padrão
1	-2,7455	0,4860	-2,7340	0,4757	0,0116	0,0103
2	-2,2562	0,4874	-2,2434	0,4624	0,0128	0,0250
3	-1,7764	0,3069	-1,7813	0,3199	0,0049	0,0130
4	-1,6283	0,2923	-1,6266	0,2934	0,0018	0,0011
5	-1,3405	0,2924	-1,3400	0,2955	0,0005	0,0031
6	-0,6505	0,2518	-0,6518	0,2519	0,0014	0,0001
7	-0,3312	0,2110	-0,3287	0,2088	0,0025	0,0023
8	0,1176	0,2010	0,1175	0,2014	0,0000	0,0004
9	0,3451	0,2052	0,3379	0,2066	0,0071	0,0014
10	0,8895	0,1991	0,8877	0,2010	0,0018	0,0019
11	1,2404	0,2362	1,2395	0,2367	0,0009	0,0005
12	1,6744	0,3171	1,6679	0,3144	0,0065	0,0027
13	2,1652	0,4392	2,1657	0,4476	0,0005	0,0084
14	2,4080	0,5476	2,3788	0,5120	0,0292	0,0356

Na Tabela 1 temos a estimativa *a posteriori* das proficiências calculadas pela média e o desvio padrão relativos aos dos 14 indivíduos utilizando a metodologia proposta com os 40 itens respondidos. Note que ambos os algoritmos aproximam a mesma distribuição *a posteriori*. Sendo que um utiliza a estimação conjunta dos parâmetros dos itens e das habilidades e o outro utiliza a incerteza existente na calibração dos itens para estimar a habilidade dos indivíduos posteriormente. Dada a robustez e eficiência do MCMC, a metodologia proposta é dita eficiente se as estimativas estiverem próximas àquelas obtidas via MCMC, o que pode ser observado na Tabela 1 e Figura 3.

Visto que o resultado foi satisfatório, o próximo passo foi encontrar o tamanho mínimo da amostra de MC dos parâmetros dos itens para a qual o nosso modelo ainda estima de forma eficiente e satisfatória a proficiência dos respondentes. Determinou-se então, diferentes tamanhos de amostra dos parâmetros dos itens retirando observações a

Figura 3 – Gráfico de comparação entre a densidade das proficiências obtida pelo MCMC e a densidade obtida pela metodologia proposta.

partir dos *lags* 30, 40, 60 e 100 e obteve-se então, amostras de tamanhos 1000, 750, 500 e 300 respectivamente. Feito isso, estimou-se novamente a proficiência dos indivíduos utilizando a metodologia proposta.

Tabela 2 – Estimativas das proficiências obtidas para diferentes tamanhos de amostra dos parâmetros dos itens.

Resp	Quantidade de interesse	Estimativa MCMC	Metodologia Proposta (EAP)				
		N= 30000	N= 30000	Lag = 30 N = 1000	Lag = 40 N = 750	Lag = 60 N = 500	Lag = 100 N = 300
1	Proficiência	-2,7455	-2,7340	-2,7329	-2,7349	-2,7321	-2,7331
	Desvio Padrão	(0,4860)	(0,4757)	(0,4757)	(0,4757)	(0,4761)	(0,4754)
2	Proficiência	-2,2562	-2,2434	-2,2442	-2,2450	-2,2465	-2,2485
	Desvio Padrão	(0,4874)	(0,4624)	(0,4629)	(0,4628)	(0,4637)	(0,4641)
3	Proficiência	-1,7764	-1,7813	-1,7816	-1,7819	-1,7825	-1,7833
	Desvio Padrão	(0,3069)	(0,3199)	(0,3201)	(0,3199)	(0,3201)	(0,3204)
4	Proficiência	-1,6283	-1,6266	-1,6268	-1,6274	-1,6280	-1,6270
	Desvio Padrão	(0,2923)	(0,2934)	(0,2937)	(0,2935)	(0,2939)	(0,2937)
5	Proficiência	-1,3405	-1,3400	-1,3400	-1,3411	-1,3410	-1,3409
	Desvio Padrão	(0,2924)	(0,2955)	(0,2956)	(0,2959)	(0,2957)	(0,2957)
6	Proficiência	-0,6505	-0,6518	-0,6520	-0,6525	-0,6512	-0,6502
	Desvio Padrão	(0,2518)	(0,2519)	(0,2519)	(0,2521)	(0,2513)	(0,2521)
7	Proficiência	-0,3312	-0,3287	-0,3287	-0,3287	-0,3288	-0,3286
	Desvio Padrão	(0,2110)	(0,2088)	(0,2087)	(0,2087)	(0,2086)	(0,2090)
8	Proficiência	0,1176	0,1175	0,1176	0,1177	0,1175	0,1182
	Desvio Padrão	(0,2010)	(0,2014)	(0,2012)	(0,2012)	(0,2010)	(0,2014)
9	Proficiência	0,3451	0,3379	0,3377	0,3375	0,3380	0,3404
	Desvio Padrão	(0,2052)	(0,2066)	(0,2065)	(0,2068)	(0,2065)	(0,2066)
10	Proficiência	0,8895	0,8877	0,8880	0,8869	0,8885	0,8896
	Desvio Padrão	(0,1991)	(0,2010)	(0,2010)	(0,2010)	(0,2011)	(0,2013)
11	Proficiência	1,2404	1,2395	1,2399	1,2394	1,2397	1,2386
	Desvio Padrão	(0,2362)	(0,2367)	(0,2366)	(0,2368)	(0,2367)	(0,2364)
12	Proficiência	1,6744	1,6679	1,6677	1,6676	1,6687	1,6661
	Desvio Padrão	(0,3171)	(0,3144)	(0,3141)	(0,3144)	(0,3146)	(0,3135)
13	Proficiência	2,1652	2,1657	2,1660	2,1658	2,1664	2,1664
	Desvio Padrão	(0,4392)	(0,4476)	(0,4475)	(0,4477)	(0,4477)	(0,4473)
14	Proficiência	2,4080	2,3788	2,3788	2,3784	2,3788	2,3791
	Desvio Padrão	(0,5476)	(0,5120)	(0,5118)	(0,5120)	(0,5120)	(0,5117)

Percebe-se a partir das estimativas apresentadas na Tabela 2 e dos gráficos apresentados na Figura 4 que podemos utilizar um tamanho mínimo de amostra *a posteriori* dos parâmetros dos itens igual a 300 para a metodologia proposta, e ainda obter uma estimação satisfatória da proficiência dos indivíduos.

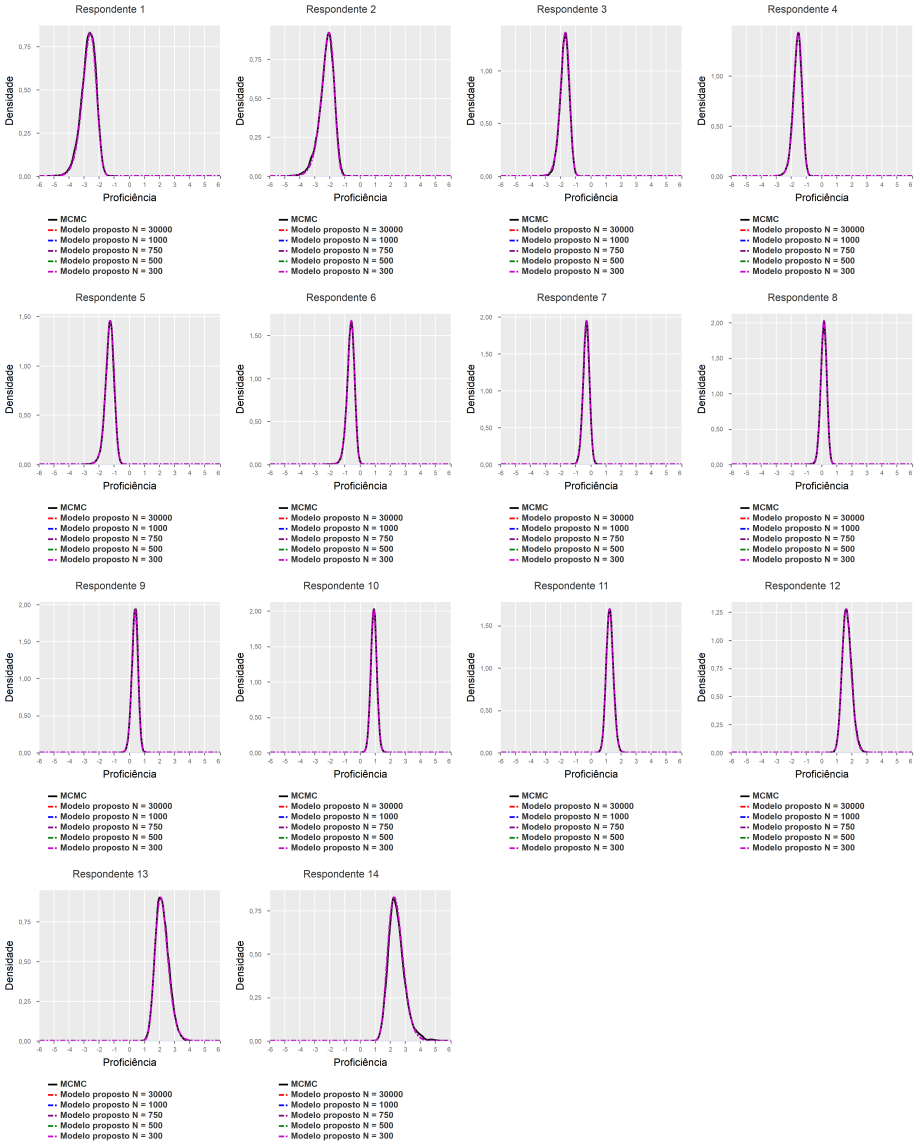

Figura 4 – Densidade da proficiência obtida via MCMC e a obtida pela metodologia proposta para diferentes tamanhos de cadeia dos parâmetros dos itens.

Na Tabela 3 vemos a relação de decréscimo no tempo de rodagem do algoritmo a medida que o tamanho da amostra diminui, ou seja, quanto menor a amostra de MC mais rápido o algoritmo finaliza a estimação.

Tabela 3 – Tempo total gasto (em segundos) pelo algoritmo desenvolvido utilizando a metologia proposta para diferentes tamanhos de amostra dos parâmetros dos itens

Resp	N = 30000	Lag = 30 N = 1000	Lag = 40 N = 750	Lag = 60 N = 500	Lag = 100 N = 300
1	694,80	23,00	17,53	11,93	7,23
2	691,09	23,36	17,20	11,46	7,00
3	690,82	23,10	17,51	11,63	6,96
4	688,45	22,86	17,45	11,83	7,05
5	683,86	23,18	17,01	11,53	7,09
6	662,49	21,92	16,69	11,20	6,87
7	664,42	22,06	16,56	11,47	6,94
8	647,66	21,62	16,63	10,86	6,72
9	657,24	22,08	16,56	10,82	6,79
10	627,75	21,00	15,87	10,41	6,47
11	602,99	20,32	15,00	10,22	6,41
12	602,51	19,71	15,14	10,26	6,30
13	600,81	20,03	15,20	10,27	6,29
14	602,42	19,83	15,20	10,30	6,30

Levando em consideração que queremos uma amostra de MC dos parâmetros que mantenha a simulação robusta e procurando obter um *grid* fino o bastante de modo a melhorar a precisão da estimativa da proficiência, escolhemos a amostra final dos parâmetros dos itens de tamanho 1000 e o *grid* de 2000 pontos variando na escala de -6 a 6 para a quadratura Gaussiana. Dessa forma, gerou-se a proficiência dos 2000 indivíduos que compõem o estudo de simulação e comparou-se com o valor das proficiências estimadas via MCMC.

Na Figura 5 temos a dispersão entre os valores da média *a poste-*

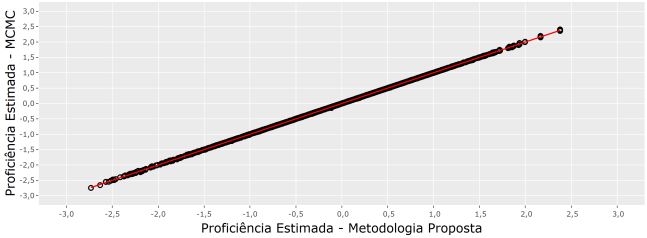

Figura 5 – Gráfico de dispersão entre a proficiência obtida via MCMC e a proficiência obtida pela metodologia proposta.

riori obtida via MCMC e pela metodologia proposta, a linha vermelha representa a reta na qual o valor do ponto no eixo x é igual ao valor do ponto no eixo y. Vemos que as estimativas da proficiência são muito semelhantes, além disso são altamente correlacionadas, sendo o valor da correlação igual a 1,0000.

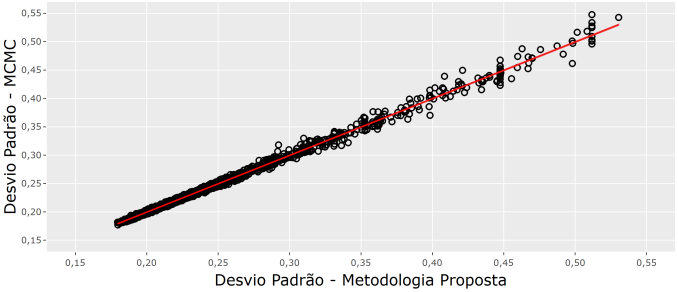

Figura 6 – Gráfico de dispersão entre o desvio padrão obtido via MCMC e o desvio padrão obtido pela metodologia proposta.

Na Figura 6 temos a dispersão entre os valores do desvio padrão *a posteriori* obtido via MCMC e pela metodologia proposta, a linha vermelha representa a reta na qual o valor do ponto no eixo x é igual ao valor do ponto no eixo y. Vemos que apesar de alguns pontos serem um pouco mais dispersos as estimativas do desvio padrão são muito

semelhantes, além disso são altamente correlacionadas, sendo o valor da correlação igual a 0,9979.

Tabela 4 – Estatísticas descritivas dos valores discrepantes *(outliers)* dos desvios padrão a posteriori.

Outliers - Desvio padrão	Total	Mínimo	Máximo	Média	Desvio Padrão
MCMC	100	0,3768	0,5476	0,4414	0,0400
Metodologia Proposta	104	0,3778	0,5305	0,4390	0,0395

Na Tabela 4 observa-se as estatísticas descritivas dos *outliers* dos desvios padrão estimados via MCMC e pela metodologia proposta. Para determinar se um ponto é *outlier*, utilizou-se o coeficiente para determinar os limites (*whisker*) do boxplot igual a 2,5. Vemos que as duas abordagens são muito semelhantes inclusive nos valores discrepantes. Vale ressaltar que para os valores estimados da proficiência dos indivíduos nenhuma das duas abordagens apresentou *outliers*.

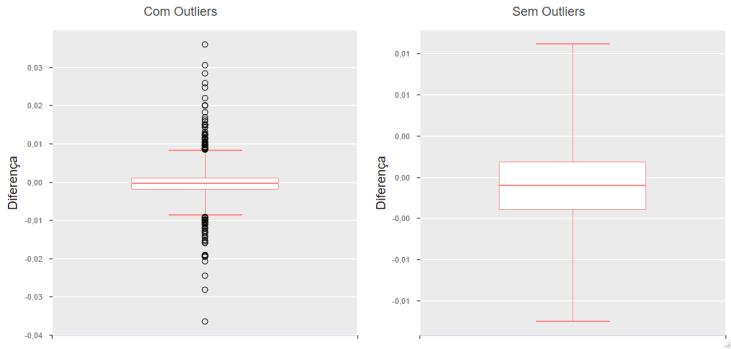

Figura 7 – Boxplot das diferenças entre a estimativa da proficiência obtida via MCMC e a estimativa da proficiência obtida pela metodologia proposta para os 2000 indivíduos do estudo de simulação.

Verifica-se pela Figura 7 e pela Tabela 5 que os valores da diferença entre a estimativa da proficiência obtida via MCMC e a estimativa obtida pela metodologia proposta concentram-se em torno de zero.

Tabela 5 – Estatísticas descritivas dos valores discrepantes *(outliers)* da diferença das proficiências estimadas da Figura 7.

Outliers - Desvio padrão	Total	Mínimo	Máximo	Média	Desvio Padrão
Diferença positiva	34	0,0131	0,0311	0,0205	0,0053
Diferença negativa	27	-0,0408	-0,0151	-0,0205	0,0065

Observa-se também que existem 61 valores discrepantes, nesses casos ora a metodologia superestima, ora subestima o valor da proficiência, sendo que essa diferença em valor absoluto não ultrapassa o valor de 0,041 aproximadamente.

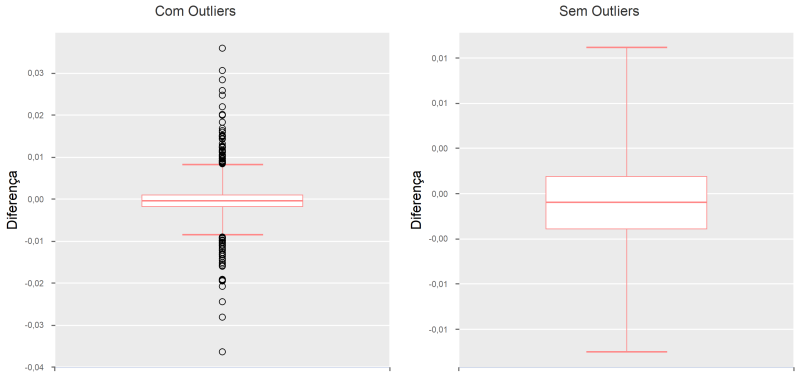

Figura 8 – Boxplot das diferenças entre o desvio padrão a posteriori obtido via MCMC e pela metodologia proposta para os 2000 indivíduos do estudo de simulação.

Tabela 6 – Estatísticas descritivas dos valores discrepantes *(outliers)* da diferença do desvio padrão *a posteriori* da Figura 8.

Outliers - Desvio padrão	Total	Mínimo	Máximo	Média	Desvio Padrão
Diferença positiva	50	0,0076	0,0358	0,0135	0,0064
Diferença negativa	46	-0,0366	-0,0082	-0,0135	0,0058

Verifica-se pela Figura 8 e pela Tabela 6 que os valores da diferença entre o desvio padrão *a posteriori* da proficiência obtida via MCMC e o desvio padrão *a posteriori* obtida pela metodologia pro-

posta concentram-se em torno de zero. Observa-se também que existem 96 valores discrepantes, nesses casos ora a metodologia superestima, ora subestima o valor do desvio padrão a posteriori, sendo que essa diferença em valor absoluto não ultrapassa o valor de 0,0366 aproximadamente.

Através desse estudo de simulação é possível notar que a metodologia proposta é satisfatória e eficiente para estimar a proficiência dos indivíduos e a precisão associada, uma vez que as estimativas obtidas são muito semelhantes as estimativas obtidas via MCMC, na qual os parâmetros dos itens e a proficiência dos indivíduos foram estimadas conjuntamente.

3 TESTE ADAPTATIVO CONSIDERANDO A METODOLOGIA PROPOSTA

O Teste Adaptativo Informatizado (TAI) é comumente baseado na TRI onde os itens são administrados e apresentados pelo computador. Este teste procura apresentar apenas itens que proporcionam uma estimação eficiente da proficiência do indivíduo que o realiza. Isso significa que os examinandos não responderão a questões que não acrescentam informações significativas sobre sua proficiência, fazendo com o que teste termine com um número de questões menor que os testes tradicionais e que seja mais rápido, eficiente e preciso.

Os testes adaptativos permitem que a proficiência do respondente seja estimada de forma iterativa, isso quer dizer que a cada item respondido a estimativa da proficiência é atualizada. Para que um teste adaptativo seja utilizado, é necessário definir três componentes essenciais: o método para estimar a proficiência do indivíduo e a precisão associada, o procedimento para selecionar o próximo item e o critério de parada para finalizar o teste.

Neste trabalho propomos e implementamos um teste adaptativo onde a cada passo do algoritmo, a incerteza existente no processo de calibração dos parâmetros dos itens é considerada na estimação iterativa da habilidade dos indivíduos, ou seja, a cada item respondido, a distribuição *a posteriori* da proficiência é atualizada. Propomos também um critério para escolha dos itens a serem administrados e um critério de parada para o teste adaptativo que levam em consideração a incerteza em torno dos parâmetros dos itens e da proficiência.

3.1 CRITÉRIO DE ESCOLHA DO PRÓXIMO ITEM UTILIZANDO A METODOLOGIA PROPOSTA

Uma vez que um dos componentes essenciais dos testes adaptativos consiste no procedimento de seleção dos itens ao longo do teste, propomos um novo critério de escolha iterativa do próximo item levando em consideração a incerteza existente no processo da estimação dos parâmetros. O critério proposto neste trabalho, da mesma forma que o método de Máxima Informação proposto na literatura e comumente utilizado nas aplicações de testes adaptativos, é baseado na Informação de Fisher (IF) definida na Seção 1.4.1.

Os métodos existentes baseados na Informação de Fisher calculam a função de IF para todos os itens disponíveis no banco de dado baseados no valor da estimativa pontual atual da proficiência. Desta forma, seleciona-se o item que retorna o valor máximo entre essas informações calculadas. O critério de escolha proposto levará em consideração não somente a incerteza sobre os parâmetros dos itens quanto aquela sobre a proficiência.

3.1.0.1 Máxima Informação Esperada *a Posteriori* (MIEAP)

Para a nossa modelagem, o critério de seleção do próximo item denominado por Máxima Informação Esperada *a Posteriori* será dado por:

$$E_{\theta,\xi}[I_{F,I,\xi}(\theta)] \approx E_{\theta}\left[\frac{\sum_{\mathrm{n}=1}^{N} I_{F,I,\xi_n}(\theta)}{N}\right], \tag{3.1}$$

onde, a esperança é tomada com relação à distribuição *a posteriori* de θ e ξ, e $p(\theta,\xi|x,y)$ é aproximada por $p(\theta|x,y)p(\xi|y)$. Para um dado item

i, a Informação de Fisher (verificar a conta no Apêndice A) é dada pela equação apresentada em (1.6).

3.1.0.2 Itens iniciais do Teste Adaptativo

Na fase inicial do teste adaptativo não temos nenhuma informação sobre a proficiência do indivíduo fazendo com que a utilização de um método baseado na verossimilhança possa não ser a melhor opção. Uma alternativa apresentada na literatura para contornar esse problema é substituir a medida de Informação de Fisher pela Informação de Kullback-Leibler (KL) sugerida por Chang e Ying (1996). Não entraremos em detalhe sobre este método, uma vez que não será utilizado neste trabalho, para informações mais detalhadas ver Chang e Ying, (1996).

Para contornar o problema da utilização da IF no início do TAI, realizou-se então um estudo para verificar o número mínimo de itens escolhidos de forma determinística a serem respondidos no início do teste para se obter uma estimativa inicial da proficiência e após esses itens serem administrados, utilizar o método proposto de escolha do próximo item.

A análise para seleção determinística de itens foi feita para 4 critérios diferentes considerando 5 itens fixos. Os critérios estabelecidos foram:

1. 5 itens iniciais previamente fixados com parâmetro de dificuldade do item variando ao longo da escala de proficiência independente do acerto ou erro do respondente, ou seja, aplicamos inicialmente 5 itens iguais a todos os respondentes.

2. 5 itens iniciais com o valor médio do parâmetro de dificuldade do item variando com base nos percentis da distribuição Normal (0,05; 0,1625; 0,275; 0,3875; 0,5; 0,6125; 0,7250; 0,8375; 0,95). O item inicial escolhido é aquele que possui o valor médio do parâmetro de dificuldade mais próximo do percentil 0,5, dessa forma os próximos itens são selecionados de acordo com o acerto (escolhe-se o item com dificuldade média mais próxima do percentil mais próximo à direita) ou erro (escolhe-se o item com dificuldade média mais próxima do percentil mais próximo à esquerda).

3. 5 itens iniciais com valor médio do parâmetro de dificuldade do item aumentando ou diminuindo em 0,5 conforme o acerto ou o erro do respondente.

4. 2 itens são escolhidos com base no percentil da distribuição Normal e 3 itens são escolhidos de acordo com o critério de Máxima Informação convencional plugando o valor da proficiência estimada no passo anterior.

Entre os métodos apresentados acima, o do item 2 foi o que apresentou uma melhor performance e foi escolhido como critério inicial de escolha de itens. O método inicial então funciona da seguinte forma: O item inicial administrado é sempre o item com parâmetro b (de dificuldade) em média mais próximo de zero. Daí em diante, você utiliza o item com parâmetro b em média mais próximo do percentil correspondente à direita (se acerta) ou esquerda (se erra), sem repetir itens, até que os cinco itens iniciais sejam administrados, considerando os seguintes valores:

- Quantil 0,05 $\approx$ - 1,645

- Quantil 0,1625 $\approx$ - 0,984
- Quantil 0,275 $\approx$ - 0,598
- Quantil 0,3875 $\approx$ - 0,286
- Quantil 0,5 = 0
- Quantil 0,6125 $\approx$ 0,286
- Quantil 0,7250 $\approx$ 0,598
- Quantil 0,8375 $\approx$ 0,984
- Quantil 0,95 $\approx$ 1,645

Considere b_i o i-ésimo item administrado, $i = 1, ..., 5$. Alguns casos são apresentados a seguir para exemplificar o algoritmo.

Caso 1: O indivíduo que acerta todos os itens iniciais.

$b_1 \approx 0$ -> Acerta

$b_2 \approx 0,286$ -> Acerta

$b_3 \approx 0,598$ -> Acerta

$b_4 \approx 0,984$ -> Acerta

$b_5 \approx 1,645$ -> Estimativa inicial da proficiência encontrada e começamos com o critério de seleção proposto.

Caso 2: O indivíduo que erra todos os itens iniciais.

$b_1 \approx 0$ -> Erra

$b_2 \approx -0,286$ -> Erra

$b_3 \approx -0,598$ -> Erra

$b_4 \approx -0,984$ -> Erra

$b_5 \approx -1,645$ -> Estimativa inicial da proficiência encontrada e começamos com o critério de seleção proposto.

Caso 3: Caso mais comum, os respondentes acertam e erram os itens selecionados.

$b_1 \approx 0$ -> Acerta

$b_2 \approx 0,286$ -> Acerta

$b_3 \approx 0,598$ -> Erra

$b_4 \approx 0,286$ -> Erra

$b_5 \approx 0$ -> Estimativa inicial da proficiência encontrada e começamos com o critério de seleção proposto.

3.1.0.3 Algoritmo de Seleção de Itens

O algoritmo para seleção de itens é dado por:

Passo 1: Administrar os primeiros 5 itens de forma determinística, como apresentado na Seção 3.1.0.2.

Passo 2: Obter a aproximação da distribuição *a posteriori* da proficiência dado os itens já respondidos. Calcule, para cada $i = 1, ..., I$, $n = 1, ..., N$ e para cada valor de θ no *grid* fixado:

$$I_{F,i,\xi_n}(\theta_k) = \frac{\left[(1-c_{i,n})a_{i,n}\phi(a_{i,n}\theta_k - b^*_{i,n})\right]^2}{[c_{i,n}+(1-c_{i,n})\Phi(a_{i,n}\theta_k - b^*_{i,n})][1-c_{i,n}-(1-c_{i,n})\Phi(a_{i,n}\theta_k - b^*_{i,n})]}$$

retornando, para cada θ_k; a matriz:

$$\begin{bmatrix} I_{F,1,\xi_1}(\theta_k) & I_{F,2,\xi_1}(\theta_k) & \dots & I_{F,\mathrm{I},\xi_1}(\theta_k) \\ I_{F,1,\xi_2}(\theta_k) & I_{F,2,\xi_2}(\theta_k) & \dots & I_{F,I,\xi_2}(\theta_k) \\ \vdots & \vdots & \ddots & \vdots \\ I_{F,1,\xi_N}(\theta_k) & I_{F,2,\xi_N}(\theta_k) & \dots & I_{F,I,\xi_N}(\theta_k) \end{bmatrix}$$

Passo 3: Calcular a esperança da Informação de Fisher com res-

peito a ξ, para cada θ_k:

$$E_{\xi_i}[I_{F,I,\xi_n}(\theta_k)] \approx \frac{\sum_{\mathrm{n=1}}^{N} I_{F,I,\xi_n}(\theta_k)}{N} = g_i(\theta_k),$$

retornando a matriz:

$$\begin{bmatrix} g_1(\theta_1) & g_1(\theta_2) & \dots & g_1(\theta_K) \\ g_2(\theta_1) & g_2(\theta_2) & \dots & g_2(\theta_K) \\ \vdots & \vdots & \ddots & \vdots \\ g_I(\theta_1) & g_I(\theta_2) & \dots & g_I(\theta_K) \end{bmatrix}$$

Passo 4: Aproximar via quadratura (Gaussiana) a esperança de $g_i(\theta)$, onde $p(\theta|x,y)$ é a distribuição atual de θ calculada para os itens já respondidos. Desta forma, obtemos a estimativa de $IEAP_i$ para cada item i:

$$\begin{bmatrix} IEAP_1 & IEAP_2 & \dots & IEAP_I \end{bmatrix}$$

Portanto o próximo item selecionado por esse método será aquele com maior valor para o $IEAP_i$

Passo 5: Apresente o item ao respondente, retire-o do vetor de itens a serem exibidos na próxima iteração e retorne ao passo 2.

Os procedimentos de seleção e aplicação dos itens no teste são feitos repetidamente, até que algum critério de parada seja satisfeito ou que todos os itens sejam selecionados.

3.2 CRITÉRIO DE PARADA DO TESTE ADAPTATIVO

Uma vez que um dos objetivos do teste adaptativo é fazer com os indivíduos não respondam mais itens do que o necessário, onde frequentemente esses itens acrescentam pouca informação sobre as proficiências estimadas, é necessário a construção de um critério de parada do teste adaptativo que indique o momento ideal para que a administração de itens ao respondente seja finalizada.

A análise do critério de parada foi feita para 2 opções diferentes. As escolhas estabelecidas inicialmente foram:

Critério 1: O algoritmo para quando a derivada da curva aproximada de decaimento do desvio padrão *a posteriori* da proficiência é menor que 0,001 em dois itens seguidos.

Critério 2: O algoritmo para quando a razão entre a derivada da curva aproximada de decaimento do desvio padrão *a posteriori* da proficiência após o i-ésimo item respondido e a derivada após o primeiro item respondido for menor que 0,01.

Na Figura 9 temos uma visualização de como o critério de parada funciona. A imagem apresenta um exemplo da primeira derivada da curva de decaimento do desvio padrão que utilizamos como critério de parada do teste. A linha preta na vertical indica o número de itens que o indivíduo respondeu até o critério 1 de parada proposto ter sido satisfeito e a linha azul na vertical indica o número de itens que o indivíduo respondeu até o critério 2 de parada proposto ter sido satisfeito. Nota-se que o critério 1 finaliza depois do critério 2, apesar da curva já ter se estabilizado e não apresenta variações significativas.

Em testes adaptativos temos que quanto mais itens são respondidos, menor é o desvio padrão associado a estimativa da proficiência,

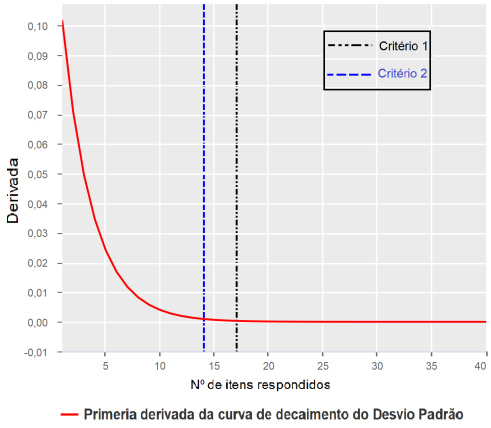

Figura 9 – Exemplo da derivada da curva de decaimento aproximada do desvio padrão.

uma vez que temos mais informação para estimar a habilidade do indivíduo. Portanto, a cada novo item respondido, o desvio padrão estimado sofre decaimento. A ideia de utilizar esse decaimento é criar um critério de parada que termine o teste assim que a variação do desvio padrão *a posteriori* de um item para o outro não seja mais significativa. Para isso utilizamos a primeira derivada da curva uma vez que ela representa a taxa de variação instantânea do desvio padrão em relação a quantidade de itens respondidos.

3.2.1 Algoritmo de estimação da curva

A curva aproximada de decaimento do desvio padrão *a posteriori* da proficiência é estimada da seguinte forma:

Passo 1: Depois de administrados os 6 primeiros itens ao respondente, levando em consideração que os 5 primeiros itens são escolhidos da forma determinística (ver Seção 3.1.0.2). Calcule:

$$d(i) = \delta + \beta \exp\{\alpha i\}, \alpha < 0, \tag{3.2}$$

onde i é o número de itens respondidos até o momento no teste adaptativo. A função $d(i)$ que ajusta a curva de decaimento do desvio padrão é calculada então para a sequência de valores de 1 a i. Nesse primeiro passo, teremos a curva estimada para os 6 pontos.

Passo 2: Estime os parâmetros (δ, β, α) que ajustam a curva de decaimento do desvio padrão baseados no número de itens respondidos.

Passo 3: Encontre a primeira derivada de $d(i)$, dada por:

$$\frac{\partial d}{\partial i} = \alpha\beta \exp\{\alpha i\}, \tag{3.3}$$

Definidas as duas opções de parada do teste, uma simulação foi feita e rodou-se o teste adaptativo para os 2000 indivíduos da amostra considerando essas opções. Fez-se uma comparação para verificar a proporção de vezes em que cada um deles para antes. A comparação foi feita da seguinte forma: Comparamos para cada indivíduo, o número de itens respondidos e critério que obteve o menor número recebe o valor 1. Caso os dois critérios tenham o mesmo número de itens respondidos os dois recebem o valor 1. Soma-se o valor atribuído a cada um dos critérios e divide pelo número de respondentes da amostra que é igual a 2000. Portanto o critério que tiver o maior valor representa o aquele que parou antes o maior número de vezes.

Comparando os resultados para determinar qual o melhor critério e com base na Figura 9 apresentada, o critério 2 foi escolhido, visto que além de finalizar a administração de itens quando a derivada da

curva começa a se estabilizar e de em geral, encerrar o teste adaptativo antes do critério 1, as estimativas das proficiências obtidas em geral são satisfatórias e muito próximas das estimativas obtidas com todos os itens.

A ideia do critério 2 é finalizar o teste quando o decaimento da curva do desvio padrão *a posteriori* está se estabilizando, isto é, a partir do momento em que os itens administrados ao respondente não farão mais diferença significativa na precisão da estimativa de sua proficiência. Ou seja, o teste é encerrado no i-ésimo item respondido se:

$$cp = \frac{\alpha\beta \exp\{\alpha i\}}{\alpha\beta \exp\{\alpha 1\}} = \frac{\exp\{\alpha i\}}{\exp\{\alpha 1\}} = \exp\{\alpha(i-1)\} < 0,01, \alpha < 0. \quad (3.4)$$

3.3 ESTUDO DE SIMULAÇÃO PARA AVALIAÇÃO DO TESTE ADAPTATIVO

O algoritmo adaptativo foi implementado no *software* R utilizando o método de estimação da proficiência, escolha do próximo item e critério de parada do teste propostos nessa dissertação.

Como mencionado na Seção 2.5 escolhemos a amostra final dos parâmetros dos itens de tamanho 1000 e um *grid* de 2000 pontos variando na escala de -6 a 6 para a quadratura Gaussiana.

Para estimar os parâmetros da curva de decaimento do desvio padrão a *a posteriori* da proficiência apresentada em (3.2) utilizou-se o método de Mínimos Quadrados através da função *hjkb* do pacote *dfoptim* dos autores Varadhan et al. (2018) em virtude de que a função *optim* do pacote básico *stats* dos autores R Core Team (2013) usualmente utilizada, não convergia para todas as observações.

Utilizou-se como valor inicial dos parâmetros α e β um ajuste básico de uma regressão linear simples e para o parâmetro δ utilizou - se o valor 0,1, uma vez que este pacote faz a otimização por métodos iterativos. Restringimos também o espaço dos parâmetros da seguinte forma: $\alpha \in (-\infty, 0)$, $\beta \in (0, \infty)$ e $\delta \in (0, \infty)$.

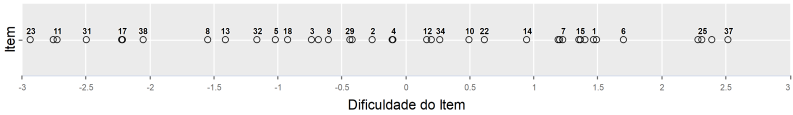

Figura 10 – Valores da média do parâmetro de dificuldade dos 40 itens do estudo de simulação.

Na Figura 10 vemos o gráfico com os valores das médias dos parâmetros de dificuldade dos itens utilizados para estimação das proficiências na escala $N(0, 1)$, percebe-se por esse gráfico que os itens tem dificuldades bem distribuídas em praticamente todos os intervalos da escala.

Na Figura 11 vemos o gráfico de dispersão entre o valor estimado das proficiências e o desvio padrão a posteriori obtidos pela metodologia proposta aplicando o teste adaptativo para os 2000 indivíduos do estudo de simulação. Assim como os modelos presentes na literatura, vemos que a metodologia proposta apresenta maior valor de desvio padrão *a posteriori* para proficiências nos extremos da escala, uma vez que temos menos itens com parâmetro de dificuldade nesses extremos.

Na Figura 12 vemos o boxplot das estimativas das proficiências e dos desvios padrão obtidos via MCMC, pela metodologia proposta com 40 itens respondidos e através do teste adaptativo aplicado a todos os indivíduos do banco de dados simulado. Percebe-se que a distribuição da estimativa da proficiência é muito semelhante, inclusive quando uti-

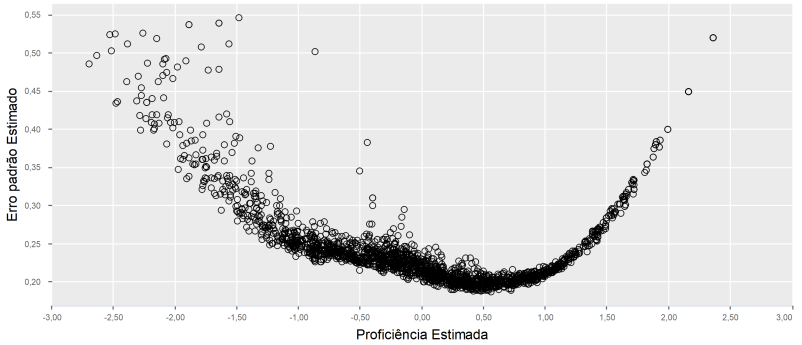

Figura 11 – Gráfico de dispersão - Proficiência x desvio padrão - estimados pela metodologia proposta de teste adaptativo.

lizamos o teste adaptativo e considerando o critério de seleção de itens e de parada propostos. Concluímos então que os indivíduos não necessariamente necessitam responder a todos os 40 itens, uma vez que para certos indivíduos a metodologia proposta de teste adaptativo estima de maneira satisfatória a proficiência e o desvio padrão a posteriori desses indivíduos.

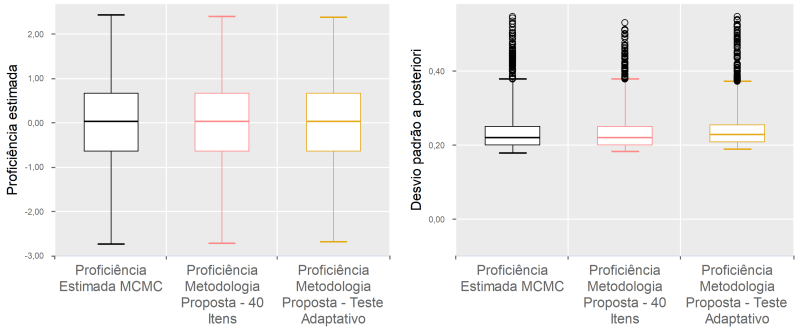

Figura 12 – Boxplot das proficiências e desvios padrão obtidos via MCMC, via metodologia proposta com 40 itens e através do teste adaptativo para os 2000 indivíduos do estudo de simulação.

Na Figura 13 temos a dispersão entre os valores da média *a pos-*

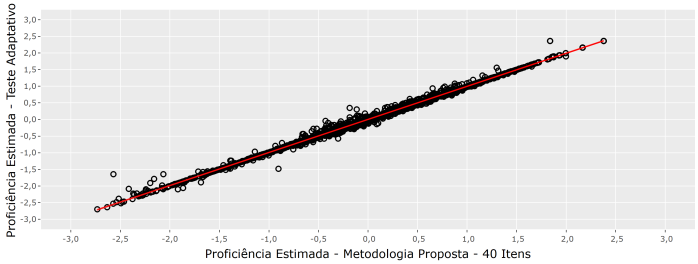

Figura 13 – Gráfico de dispersão entre a proficiência obtida com 40 itens respondidos e a proficiência obtida pelo teste adaptativo utilizando a metodologia proposta.

teriori obtida pela metodologia proposta com 40 itens respondidos e através do teste adaptativo. A linha vermelha representa a reta na qual o valor do ponto no eixo x é igual ao valor do ponto no eixo y. Vemos que as estimativas da proficiência são muito semelhantes, além disso são altamente correlacionadas, sendo o valor da correlação igual a 0,9978.

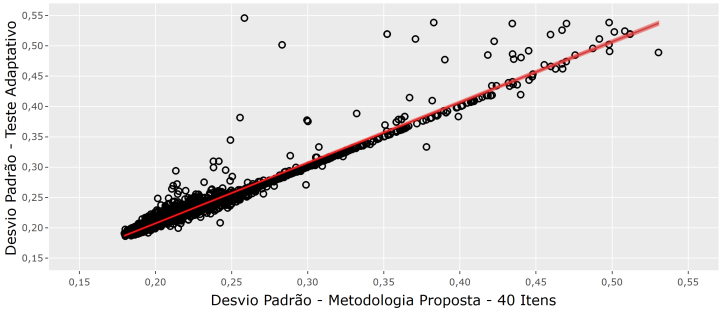

Figura 14 – Gráfico de dispersão entre o desvio padrão obtido com 40 itens respondidos e o desvio padrão obtido pelo teste adaptativo utilizando a metodologia proposta.

Na Figura 14 temos a dispersão entre o desvio padrão *a posteriori* obtido pela metodologia proposta com 40 itens respondidos e através do

teste adaptativo. Vemos que apesar do gráfico apresentar alguns pontos mais distantes da reta, em geral as estimativas do desvio padrão são próximas. Além disso são altamente correlacionadas, sendo o valor da correlação igual a 0,9733.

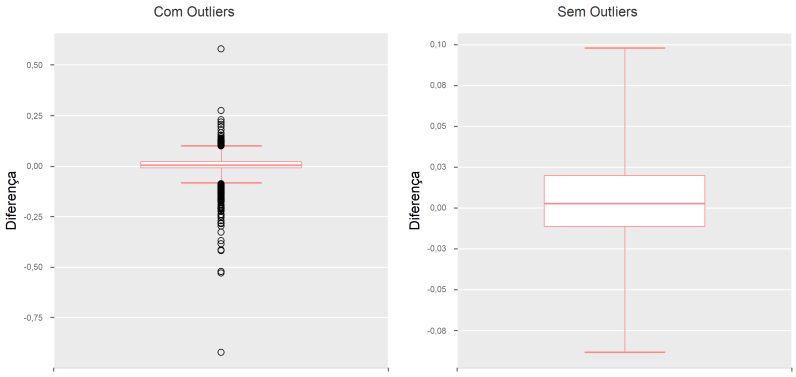

Figura 15 – Boxplot das diferenças entre a proficiência estimada com 40 itens e a proficiência estimada pelo teste adaptativo através da metodologia proposta para os 2000 indivíduos do estudo de simulação.

Tabela 7 – Estatísticas descritivas dos valores discrepantes *(outliers)* da diferença das proficiências estimadas da Figura 15.

	Total	Mínimo	Máximo	Média	Desvio Padrão
Diferença positiva	24	0,1163	0,5791	0,1704	0,0965
Diferença negativa	64	-0,9248	-0,1363	-0,2292	0,1306

Vemos pela Figura 15 e pela Tabela 7 que os valores da diferença entre a estimativa da proficiência estimada com 40 itens e a estimativa da proficiência estimada pelo teste adaptativo concentram-se em torno de -0,1 e 0,1. Entre os 2000 respondentes, vemos que existem 88 valores discrepantes, nesses casos ora a metodologia superestima, ora subestima o valor da proficiência, sendo que a maior diferença em valor absoluto é de 0,9248 para um dos indivíduos que respondeu a apenas 8 itens.

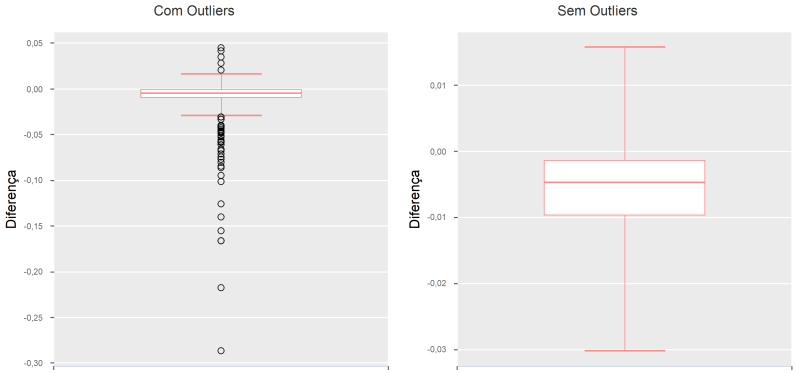

Figura 16 – Boxplot das diferenças entre o desvio padrão a posteriori com 40 itens e o desvio padrão a posteriori estimado pelo teste adaptativo através da metodologia proposta para os 2000 do estudo de simulação.

Tabela 8 – Estatísticas descritivas dos valores discrepantes *(outliers)* da diferença do desvio padrão a posteriori da Figura 16.

	Total	Mínimo	Máximo	Média	Desvio Padrão
Diferença positiva	9	0,0147	0,0447	0,0255	0,0120
Diferença negativa	43	-0,2873	-0,0311	-0,0763	0,0537

Observa-se pela Figura 16 e pela Tabela 8 que os valores da diferença entre o desvio padrão *a posteriori* concentram-se em torno de 0. Vemos também que existem 52 valores discrepantes, e que para esses casos a maioria das diferenças tem valor negativo, indicando que o desvio padrão a posteriori obtido com os 40 itens respondidos são menores que o desvio padrão a posteriori obtido pelo teste adaptativo, o que é de se esperar, uma vez que quanto mais itens são respondidos, maior é a informação que temos sobre a proficiência do indivíduo e portanto menor é o desvio padrão *a posteriori*. Verifica-se que existem alguns pontos em que o desvio padrão estimado através do teste adaptativo foi menor que o desvio padrão estimado com 40 itens.

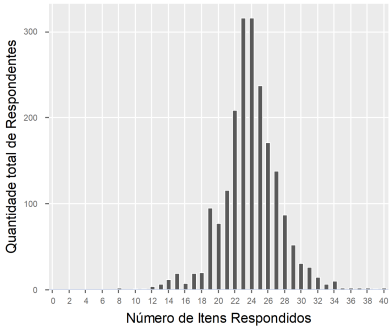

Figura 17 – Histograma do número de itens respondidos até o critério de parada proposto ser satisfeito.

A Figura 17 apresenta a distribuição do número de itens respondidos no teste adaptativo. Nela vemos que é necessário na maioria dos casos, entre 22 a 26 itens para que a proficiência seja estimada de forma satisfatória. Temos apenas 11 *outliers* e as informações desses respondentes referente ao teste com 40 itens e ao teste adaptativo utilizando a metodologia proposta estão dispostas na Tabela 9. Percebe-se que apenas 4 indivíduos respondem menos de 12 itens no teste e estes indivíduos responderam corretamente apenas 2 ou 3 itens. Já para os indivíduos que respondem mais de 36 itens em geral eles acertam entre 12 e 15 itens.

Tabela 9 – Informações dos respondentes que tiveram um comportamento atípico com relação ao número de itens respondidos no teste adaptativo.

Resp	40 Itens				Teste Adaptativo			
	Nº de itens respondidos	Nº acertos	Proficiência	Desvio Padrão	Nº de itens respondidos	Nº acertos	Proficiência	Desvio Padrão
481	40	6	-2,0658	0,3828	8	2	-1,6453	0,5384
1567	40	6	-2,5701	0,4980	8	2	-1,6453	0,5384
1618	40	14	-0,9039	0,2584	10	3	-1,4829	0,5458
1970	40	7	-2,1595	0,4225	11	3	-1,7871	0,5074
1470	40	12	-1,5698	0,3292	37	12	-1,5698	0,3292
1633	40	14	-1,2869	0,3057	37	13	-1,2869	0,3056
1038	40	13	-1,1430	0,2785	38	13	-1,1429	0,2785
1339	40	38	1,6156	0,2896	38	36	1,6156	0,2896
504	40	15	-0,8610	0,2466	39	15	-0,8610	0,2467
1740	40	15	-1,0526	0,2538	40	15	-1,0526	0,2538
1387	40	37	1,3375	0,2390	40	37	1,3375	0,2390

Tabela 10 – Sequência de respostas do teste adaptativo dos indivíduos apresentados na Tabela 9.

Respondente	Respostas
481	0,0,1,1,0,0,0,0
1567	0,0,1,1,0,0,0,0
1618	0,0,1,1,0,0,1,0,0,0
1970	0,1,1,0,0,0,1,0,0,0,0
1470	0,1,0,1,0,0,0,0,0,0,1,0,1,1,0,0,1,1,1,1,0,1,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0
1633	0,1,0,0,1,0,0,0,1,1,0,1,0,1,1,1,0,1,0,0,1,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0
1038	0,0,1,0,0,1,0,0,1,0,1,1,1,0,1,1,0,1,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
1339	1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
504	0,1,1,0,0,0,0,1,0,1,0,1,1,1,1,0,1,1,1,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
1740	0,1,1,0,0,0,0,0,0,1,1,1,1,1,0,1,0,0,1,1,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1
1387	1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1

Na Tabela 10 temos a sequência de respostas fornecidas ao teste adaptativo dadas pelos respondentes considerados como *outliers* na Tabela 9. Vemos que na maioria dos casos os indivíduos não apresentam um comportamento coerente de resposta, uma vez que acertam e erram os itens ao longo do teste. No caso dos indivíduos 1339 e 1387

eles apresentam a resposta coerente até um certo ponto e depois erram alguns itens selecionados, fazendo com que o teste aplique mais itens a esses respondentes.

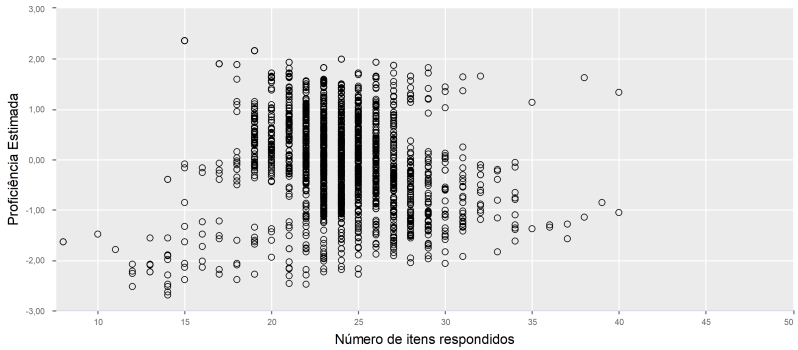

Figura 18 – Gráfico de dispersão - Proficiência estimada pelo teste adaptativo x Número de itens respondidos.

Na Figura 18 vemos o gráfico de dispersão entre o valor estimado das proficiências através do teste adaptativo e o número de itens respondidos de cada um dos 2000 indivíduos do estudo de simulação. Percebe-se que o gráfico possui um certo padrão, no qual os indivíduos de proficiência no inferior da escala possuem uma maior variação no número de itens respondidos; já para os demais indivíduos em geral eles respondem entre 20 a 30 itens no teste adaptativo. Nota-se também que apenas dois indivíduos responderam a todos os 40 itens considerando o teste adaptativo.

Na Figura 19 vemos o gráfico de dispersão entre o desvio padrão estimado das proficiências e o número de itens respondidos por cada um dos 2000 indivíduos do estudo de simulação. Em geral a maioria dos respondentes possui o desvio padrão estimado entre 0,15 e 0,25. Observa-se também que para certos casos, a medida que o número de

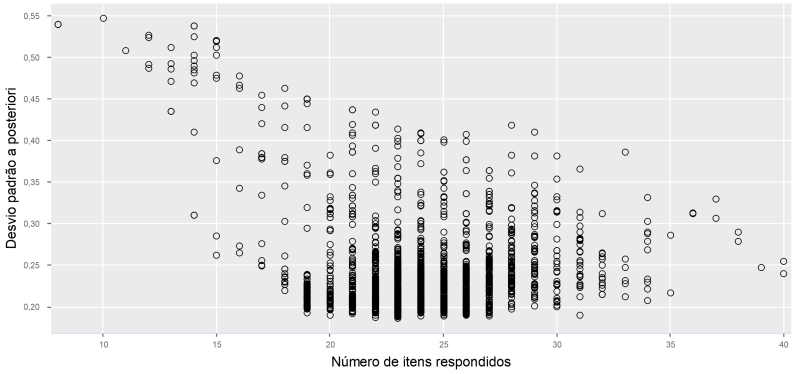

Figura 19 – Gráfico de dispersão - Desvio padrão x Número de itens respondidos.

itens diminui o desvio padrão *a posteriori* aumenta, o que é razoável de se esperar, uma vez que quanto menos itens o indivíduo responde menor é a informação que temos sobre a sua habilidade.

Para uma análise direcionada a alguns indivíduos, utilizamos novamente os respondentes escolhidos e apresentados na Seção 2.5. Na Tabela 11 temos a média e o desvio padrão *a posteriori* desses respondentes quando consideramos o teste não adaptativo utilizando os 40 itens e quando aplicamos a metodologia de teste adaptativo proposta. Vemos novamente por essa tabela, que a administração de todos os itens a todos respondentes não é necessária, visto que alguns dos itens não acrescentam informações significativas a proficiência de determinados respondentes. Vemos também que a administração de um teste adaptativo com o critério de seleção de itens e de parada do teste propostos estimam de forma satisfatória a proficiência desses indivíduos, e que o teste adaptativo reduz em alguns casos o número de itens que deve ser administrado em até 50%.

Tabela 11 – Estimativas das proficiências e desvios padrão obtidas pela metodologia proposta considerando os 40 itens respondidos e considerando o teste adaptativo.

	40 Itens				Teste Adaptativo			
	Nº de itens respondidos	Nº de acertos	Proficiência	Desvio padrão	Nº de itens respondidos	Nº de acertos	Proficiência	Desvio padrão
1	40	5	-2,7340	0,4757	14	0	-2,7006	0,4847
2	40	8	-2,2434	0,4624	16	4	-2,1399	0,4619
3	40	7	-1,7813	0,3199	24	6	-1,7808	0,3205
4	40	11	-1,6266	0,2934	27	8	1,6602	0,3165
5	40	14	-1,3400	0,2955	29	11	-1,3399	0,2954
6	40	15	-0,6518	0,2519	20	10	-0,4361	0,2527
7	40	20	-0,3287	0,2088	29	17	-0,3101	0,2136
8	40	25	0,1175	0,2014	28	17	0,1125	0,2045
9	40	27	0,3379	0,2066	22	14	0,3211	0,2209
10	40	32	0,8877	0,2010	24	17	0,8753	0,2050
11	40	34	1,2395	0,2367	24	19	1,2393	0,2399
12	40	37	1,6679	0,3144	20	17	1,6602	0,3165
13	40	39	2,1657	0,4476	19	18	2,1620	0,4491
14	40	40	2,3788	0,5120	15	15	2,3612	0,5193

Na Figura 20 temos uma visualização de como o critério de parada funciona. A primeira imagem mostra os desvios padrão estimados a medida em que os itens vão sendo aplicados no teste adaptativo e a curva de decaimento ajustada desses desvios através da Equação 3.2. A segunda imagem apresenta a primeira derivada dessa curva de decaimento que utilizamos como critério de parada do teste. As linhas azuis na vertical indicam o número de itens que o indivíduo respondeu até o critério de parada proposto ter sido satisfeito, e as linhas na horizontal indicam o desvio padrão final estimado e a derivada final da curva quando o teste é encerrado.

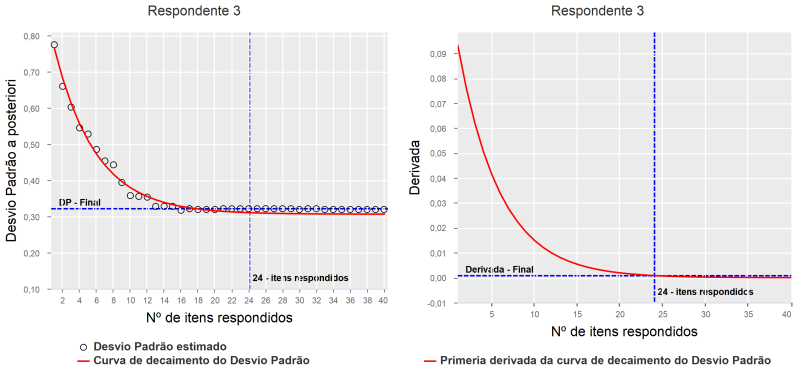

Figura 20 – Curva de decaimento aproximada do desvio padrão e derivada da curva para o respondente 3 apresentado na Tabela 11.

As curvas presentes nos gráficos foram ajustadas no vigésimo quarto item respondido por esse indivíduo, uma vez que este foi o número de itens administrados ao indivíduo até que o critério de parada fosse satisfeito no teste adaptativo. Nota-se que o critério de parada finaliza a administração de itens quando a derivada da curva começa a se estabilizar e já não apresenta variações significativas.

Na Figura 21 temos o gráfico de evolução da dificuldade média dos itens escolhidos e das proficiências estimadas a cada item administrado ao respondente 3 apresentado na Tabela 11. A linha preta indica até onde os itens determinísticos apresentados na Seção 3.1.0.2 foram administrados ao respondente e a linha vermelha indica o número de itens respondidos até o critério de parada proposto ser satisfeito. As observações em vermelho indicam que o indivíduo errou o item e as observações em verde indicam que o indivíduo acertou o item. Percebe-se pelo gráfico da proficiência estimada que o teste adaptativo foi encerrado de forma satisfatória, ou seja, quando a proficiência do indivíduo já não sofria alterações significativas.

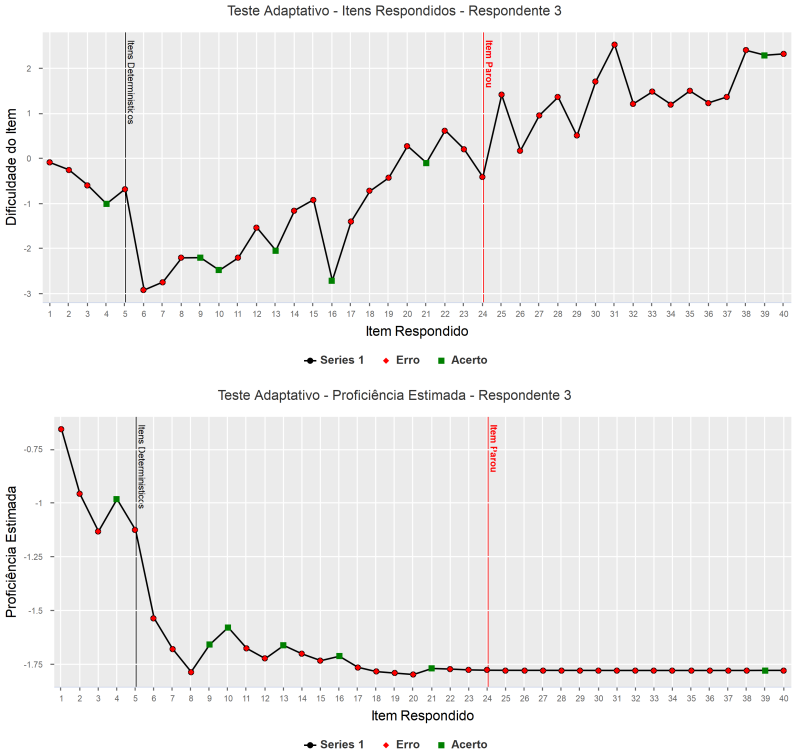

Figura 21 – Gráfico de evolução da dificuldade média dos itens escolhidos e das proficiências estimadas a cada passo do teste adaptativo administrado ao respondente 3 apresentado na Tabela 11.

Uma das grandes preocupações nos testes informatizados é o tempo que o teste gasta entre a resposta e a exibição do próximo item ao respondente. Fizemos uma simulação para verificar o tempo gasto pelo algoritmo entre a resposta de um indivíduo a uma determinada questão e a apresentação do próximo item. Considerando então o tamanho da amostra dos parâmetros dos itens igual a 1000 e um *grid* com 2000 pontos para a quadratura Gaussiana, o tempo médio gasto entre a apresentação de um item e outro considerando resposta imediata do indivíduo

foi de 0.62 segundos, utilizando um Sistema Operacional Windows 10 de 64 bits e um processador Intel(R) Core(TM) i5-6300HQ. Obteve-se então a indicação de que além de estimar de forma satisfatória a proficiência dos indivíduos temos um algoritmo computacionalmente eficiente que permite que a metodologia proposta e o algoritmo implementado sejam utilizados em aplicações reais de testes informatizados.

4 APLICAÇÃO DA METODOLOGIA PROPOSTA NOS DADOS DO ENEM 2017

Neste capítulo a metodologia proposta será aplicada ao conjunto de dados do Exame Nacional do Ensino Médio (Enem) do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) autarquia vinculada ao Ministério da Educação do Brasil (MEC).

O Enem é uma prova elaborada pelo INEP para avaliar a qualidade do ensino no país, ou seja, para verificar o domínio de competências e habilidades dos estudantes que estão concluindo o ensino médio. Na maioria dos casos, a nota obtida é usada para o ingresso em diversas universidades do país, dessa forma o Enem é considerado o maior exame vestibular do Brasil.

O Enem analisa o domínio dos respondentes em quatro áreas do conhecimento, Ciências Humanas, Ciências da Natureza, Linguagens, Códigos e suas Tecnologias e Matemática. Para evitar fraude, a prova é impressa em 4 versões diferentes, identificadas geralmente pelas cores amarela, branca, rosa e azul. O que difere uma prova da outra é apenas a mudança na ordem das questões e alternativas.

Os dados foram obtidos através do portal oficial do INEP que disponibiliza os microdados Enem. Os dados analisados se referem a prova amarela de Matemática, aplicada em todo o Brasil no ano de 2017. Alguns filtros foram feitos para assegurar a consistência dos dados e garantir uma boa estimação dos parâmetros dos itens. Esses filtros foram:

- Os alunos que estavam presentes na realização de todas as quatro áreas de conhecimento da prova. Este filtro é importante, em

razão de que buscamos analisar os indivíduos que usarão a nota obtida para o ingresso em uma universidade, desconsiderando os alunos que realizam apenas a prova de matemática por diversas razões.

- Alunos concluintes do ensino médio no ano de 2017.
- Apenas a primeira aplicação da prova foi considerada, uma vez que 2017 realizou-se uma segunda aplicação da prova. Entre os motivos para a segunda aplicação estão a interrupção do fornecimento de luz que afetou 3.574 participantes de nove locais do Brasil.

Realizados estes filtros no banco de dados, retirou-se uma amostra aleatória simples e por fim, foram analisadas provas de 5000 alunos onde cada um deles possui um vetor de respostas referente aos 45 itens da prova de matemática. As especificações para rodagem da metodologia e algoritmos propostos são as mesmas apresentadas na Seção 3.3. A estimação dos parâmetros dos itens foi feita novamente pelo algoritmo proposto por Gonçalves, Dias e Soares (2018). Foi então gerada uma amostra da distribuição *a posteriori* dos parâmetros dos itens de tamanho 30000 para cada parâmetro com um *burn-in* de 15000, dessa forma, o estudo tem uma amostra de Monte Carlo de tamanho 15000. Um *lag* de tamanho 15 foi utilizado a fim de obter uma amostra de MC de tamanho 1000.

4.1 RESULTADOS

Na Figura 22 vemos o gráfico com os valores das médias dos parâmetros de dificuldade dos itens utilizados para estimação das profi-

ciências na escala $N(0,1)$ para a amostra de respondentes do Enem. Percebe-se por esse gráfico que os parâmetros da prova de matemática não cobrem toda a escala, visto que os parâmetros dos itens tem dificuldades somente acima do valor -1. A área de Matemática e suas tecnologias é tida como mais difícil entre aquelas avaliadas no Enem, uma vez que quanto maior a dificuldade do item, maior é a proficiência exigida do indivíduo para responder corretamente ao item.

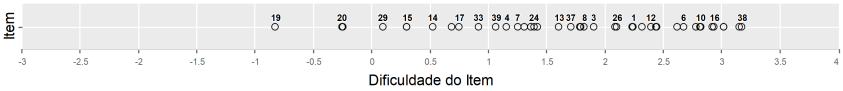

Figura 22 – Valores da média do parâmetro de dificuldade dos 45 da prova Enem 2017 de Matemática.

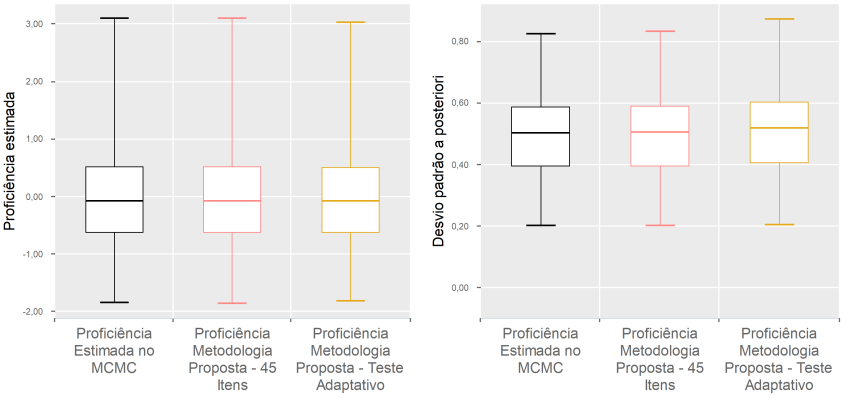

Figura 23 – Boxplot das proficiências e desvios padrão obtidos via MCMC e pela metodologia proposta com 45 itens e através do teste adaptativo para os 5000 indivíduos do Enem.

Na Figura 23 vemos o boxplot das estimativas das proficiências e dos desvios padrão obtidos via MCMC e pela metodologia proposta com 45 itens respondidos e através do teste adaptativo aplicado a todos os indivíduos da amostra retirada dos microdados do Enem. Percebe-

se que a distribuição da estimativa da proficiência é muito semelhante, inclusive quando utilizamos o teste adaptativo considerando o critério de seleção de itens e de parada do teste propostos. Concluímos então que os respondentes não necessariamente necessitam responder a todos os 45 itens, uma vez que o teste adaptativo utilizando a metodologia proposta estima de maneira satisfatória a proficiência e o desvio padrão a posteriori desses indivíduos.

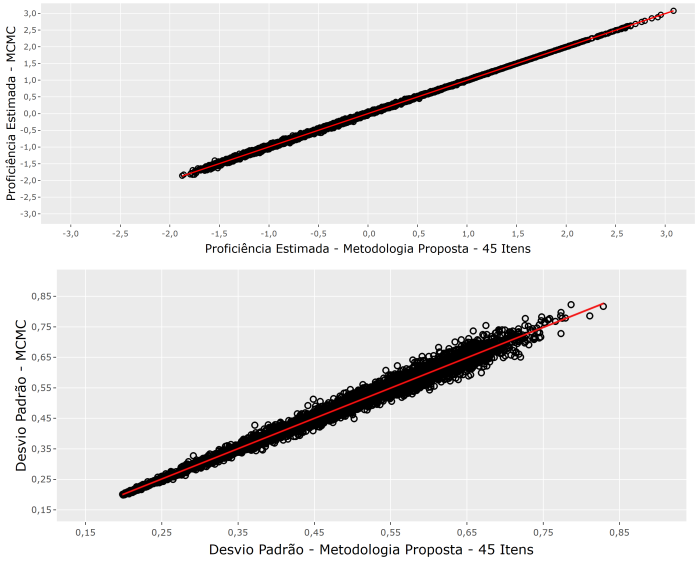

Figura 24 – Gráficos de dispersão entre as proficiências e entre os desvios padrão obtidos via MCMC e pela metodologia proposta com 45 itens respondidos.

Na Figura 24 temos a dispersão entre os valores da média *posteriori* e a dispersão entre os desvios padrão obtidos via MCMC e pela metodologia proposta com 45 itens respondidos. A linha vermelha representa a reta na qual o valor do ponto no eixo x é igual ao valor do ponto no eixo y. As estimativas são altamente correlacionadas, sendo

o valor da correlação igual a 0,9996 para as proficiências e 0,9930 para os desvios padrão. Nota-se pelo gráfico que os valores estimados da proficiência e dos desvios padrão (ainda que para esse caso os valores estejam mais dispersos) se aproximam bastante dos valores obtidos via MCMC utilizando os 45 itens considerando a metodologia proposta.

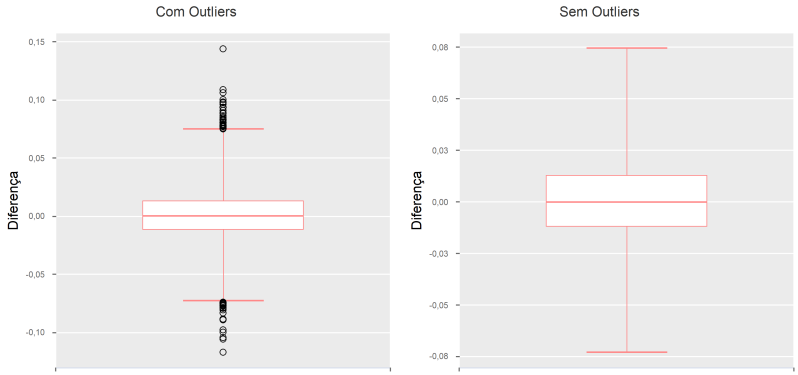

Figura 25 – Boxplot das diferenças entre a proficiência estimada pelo MCMC e a proficiência estimada com 45 itens utilizando a metodologia proposta para os 5000 respondentes do Enem 2017.

Tabela 12 – Estatísticas descritivas dos valores discrepantes *(outliers)* da diferença das proficiências estimadas da Figura 25.

	Total	Mínimo	Máximo	Média	Desvio Padrão
Diferença positiva	24	0,0921	0,0137	0,0798	0,1438
Diferença negativa	30	-0,0814	0,0126	-0,1176	-0,0695

Observa-se pela Figura 25 e pela Tabela 12 que os valores da diferença entre a proficiência estimada pelo MCMC e a proficiência obtida com 45 itens, concentram-se em torno de 0. Vemos também que existem 54 valores discrepantes entre os 5000 respondentes, nesses casos ora a metodologia superestima, ora subestima o valor da proficiência, sendo que a maior diferença em valor absoluto é de 0,1438.

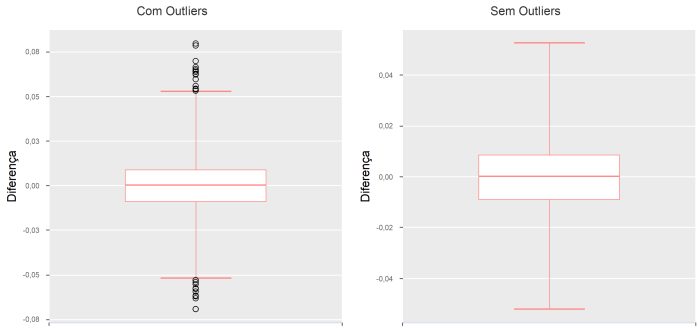

Figura 26 – Boxplot das diferenças entre o desvio padrão a posteriori estimado pelo MCMC e o desvio padrão a posteriori obtidos pela metodologia proposta utilizando os 45 itens para os 5000 respondentes do Enem 2017.

Tabela 13 – Estatísticas descritivas dos valores discrepantes *(outliers)* da diferença do desvio padrão a posteriori da Figura 26.

	Total	Mínimo	Máximo	Média	Desvio Padrão
Diferença positiva	23	0,0493	0,0796	0,0590	0,0089
Diferença negativa	17	-0,0695	-0,0503	-0,0573	0,0062

Observa-se pela Figura 26 e pela Tabela 13 que os valores da diferença entre o desvio padrão *a posteriori* concentram-se em torno de 0. Vemos também que existem 40 valores discrepantes, sendo que a maior diferença em valor absoluto é de 0,0796.

Na Figura 27 temos a dispersão entre os valores da média *posteriori* e a dispersão entre os desvios padrão obtidos pela metodologia proposta com 45 itens respondidos e através do teste adaptativo. A linha vermelha representa a reta na qual o valor do ponto no eixo x é igual ao valor do ponto no eixo y. As estimativas são altamente correlacionadas, sendo o valor da correlação igual a 0,9906 para as proficiências e 0,9627 para os desvios. Nota-se pelo gráfico que, apesar de alguns valores serem mais dispersos, em geral os valores estimados da

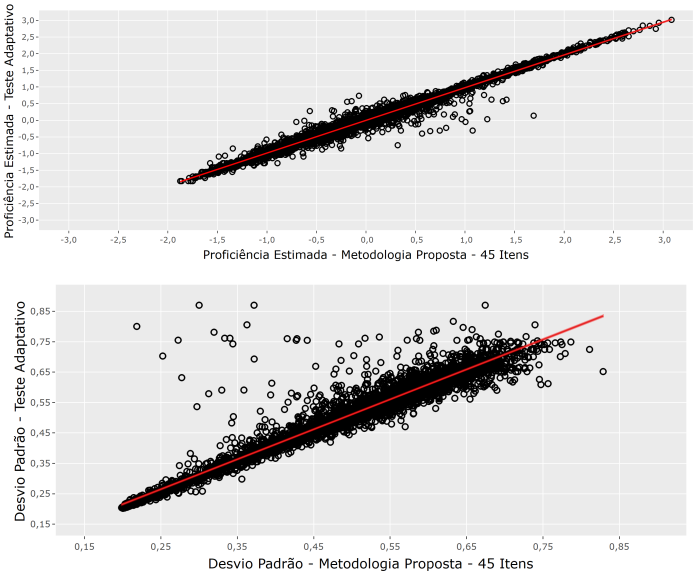

Figura 27 – Gráficos de dispersão entre as proficiências e entre os desvios padrão obtidos com 45 itens respondidos e a proficiência obtida pelo teste adaptativo utilizando a metodologia proposta.

proficiência, utilizando o teste adaptativo, se aproximam bastante dos valores obtidos com o teste considerando os 45 itens baseado na metodologia proposta. Já nos casos dos desvios padrão, temos que para alguns indivíduos o valor do desvio padrão é muito maior quando utilizamos o teste adaptativo e em geral isso acontece quando o indivíduo por alguma razão responde a poucos itens antes do teste ser finalizado.

Tabela 14 – Estatísticas descritivas dos valores discrepantes *(outliers)* da diferença das proficiências estimadas da Figura 15.

	Total	Mínimo	Máximo	Média	Desvio Padrão
Diferença positiva	93	0,2864	1,5546	0,5046	0,2595
Diferença negativa	77	-0,8431	-0,2398	-0,3598	0,1350

Observa-se pela Figura 28 e pela Tabela 14 que os valores da dife-

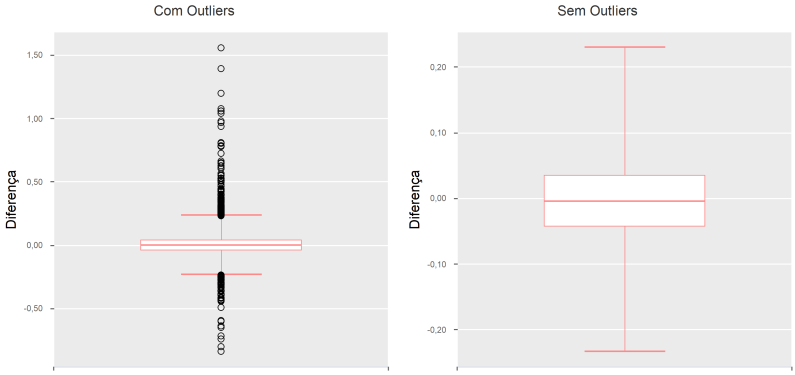

Figura 28 – Boxplot das diferenças entre a proficiência estimada com 45 itens e a proficiência estimada pelo teste adaptativo através da metodologia proposta para os 5000 respondentes do Enem 2017.

rença entre a proficiência obtida com 45 itens e a proficiência estimada pelo teste adaptativo, desconsiderando os *outliers*, concentram-se entre -0,25 e 0,25 aproximadamente. Vemos também que existem 173 valores discrepantes entre os 5000 respondentes, nesses casos ora a metodologia superestima, ora subestima o valor da proficiência, sendo que a maior diferença em valor absoluto é de 1,5546 para um dos indivíduos que respondeu apenas a 9 itens.

A Tabela 15 apresenta as estatísticas descritivas das diferenças entre a proficiência estimada com 45 itens e a proficiência estimada pelo teste adaptativo através da metodologia proposta para todos os indivíduos onde essa diferença ultrapassa 0,3 em valor absoluto. As estatísticas estão agrupadas pelo número de itens respondidos por esses indivíduos. Nota-se que o maior número de indivíduos responde a apenas 8 ou 9 itens com número médio de acertos igual a 3 e 5, respectivamente. Vemos pela tabela que esses indivíduos representam aproximadamente 3% dos indivíduos do estudo, uma vez que temos

uma amostra de 5000 respondentes.

Tabela 15 – Estatísticas descritivas das diferenças maiores que 0,3 em valor absoluto apresentadas na Figura 15.

Nº de itens respondidos	Total de indivíduos	Média de acertos	Diferença média	Diferença mínima	Diferença máxima
7	5	3	0,6210	0,3020	1,0500
8	16	3	0,7550	0,3510	1,3900
9	19	5	0,5590	0,3160	1,5500
10	10	5	0,4920	0,3450	0,9740
11	3	4	0,5700	0,3030	1,0700
12	7	5	0,3830	0,3030	0,5210
13	10	6	0,4210	0,3080	0,5630
14	13	6	0,4920	0,3060	0,9660
15	11	7	0,4740	0,3160	1,0400
16	8	7	0,4380	0,3040	0,7750
17	8	8	0,3820	0,3020	0,6240
18	8	8	0,3930	0,3010	0,5200
19	5	9	0,4160	0,3060	0,5470
20	4	9	0,3750	0,3100	0,4490
21	5	10	0,3490	0,3090	0,3750
23	3	10	0,3120	0,3100	0,3140
24	2	10	0,3220	0,3040	0,3390
28	1	11	0,3850	0,3850	0,3850

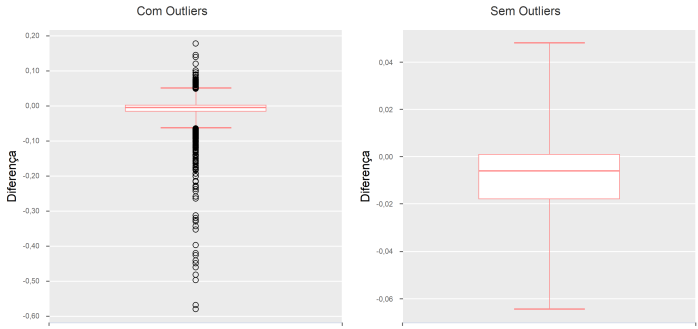

Figura 29 – Boxplot das diferenças entre o desvio padrão a posteriori estimado com 45 itens e o desvio padrão a posteriori estimado pelo teste adaptativo através da Metodologia proposta para os 5000 respondentes do Enem 2017.

Tabela 16 – Estatísticas descritivas dos valores discrepantes *(outliers)* da diferença do desvio padrão a posteriori da Figura 29.

	Total	Mínimo	Máximo	Média	Desvio Padrão
Diferença positiva	40	0,0616	0,1768	0,0809	0,0248
Diferença negativa	154	-0,5819	-0,0735	-0,1497	0,1065

Observa-se pela Figura 29 e pela Tabela 16 que os valores da diferença entre o desvio padrão *a posteriori* concentram-se em torno de 0 desconsiderando os *outliers*. Vemos também que existem 194 valores discrepantes, e que para esses casos a maioria das diferenças tem valor negativo, indicando que o desvio padrão a posteriori obtido com os 45 itens respondidos são menores que o desvio padrão a posteriori obtido pelo teste adaptativo para 106 indivíduos, sendo que a maior diferença em valor absoluto é de 0,5819. Verifica-se também um resultado inusitado em relação aos desvios padrão estimados, uma vez que a diferença positiva apresentada na Tabela 16 indica que o desvio padrão a posteriori obtido com os 45 itens respondidos são maiores que o desvio padrão a posteriori obtido pelo teste adaptativo. Uma possível

explicação para isso, esta no fato de que ao apresentar um teste com questões determinadas previamente, alguns indivíduos erram itens nos quais eles possuem alta probabilidade de acerto e sendo que alguns desses itens não acrescentam informação significativa a proficiência desses respondentes, fazendo com que a estimativa do desvio padrão aumente.

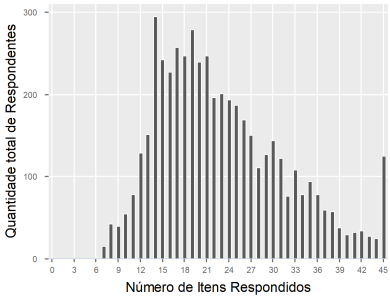

Figura 30 – Histograma do número de itens respondidos até o critério de parada proposto ser satisfeito.

Na Figura 30 percebe-se que ao fornecer um teste adaptativo aos respondentes, 49% dos indivíduos responderam entre 14 a 23 itens considerando o critério de parada proposto. Nota-se que apesar da distribuição do número de itens respondidos não apresentar nenhum *outlier*, temos 96 indivíduos que respondem menos de 10 itens, o que afeta em alguns casos a proficiência desses indivíduos levando a uma grande diferença na proficiência estimada como apresentado na Tabela 15.

A Figura 31 apresenta o gráfico de dispersão entre o valor estimado das proficiências e o número de itens respondidos por cada um dos 5000 respondentes do Enem. Percebe-se que o gráfico possui um certo padrão, quanto menor a proficiência dos indivíduos menor o número de itens respondidos, vemos também que em geral os indivíduos que respondem entre 30 a 45 itens são os indivíduos com proficiên-

cia no intervalo (0,5; 1,5). Indivíduos com alta proficiência em geral respondem entre 30 a 40 itens.

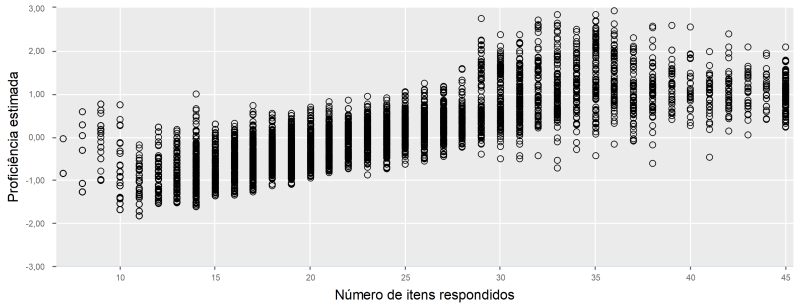

Figura 31 – Gráfico de dispersão - Proficiência estimada pelo teste adaptativo x Número de itens respondidos.

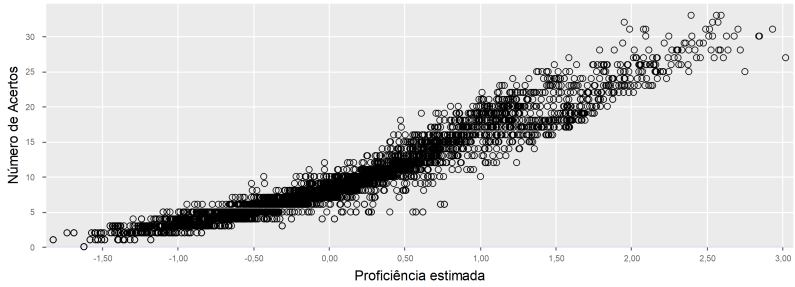

Figura 32 – Gráfico de dispersão - Proficiência estimada pelo teste adaptativo x Número acertos.

Na Figura 32 vemos o gráfico de dispersão entre o valor estimado das proficiências e o número de acertos considerando os itens administrados no teste adaptativo para cada um dos 5000 respondentes do Enem. Vemos um padrão claro, onde quanto maior o número de acertos do indivíduo, maior a sua proficiência estimada.

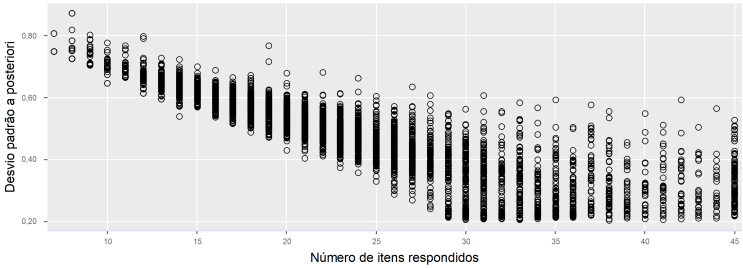

Figura 33 – Gráfico de dispersão - Desvio Padrão estimado pelo teste adaptativo x Número de itens respondidos.

Na Figura 33 vemos o gráfico de dispersão entre o desvio padrão estimado das proficiências e o número de itens respondidos por cada um dos 5000 respondentes da prova Enem. Percebe-se um certo padrão, no qual a medida que o número de itens diminui o desvio padrão *posteriori* aumenta.

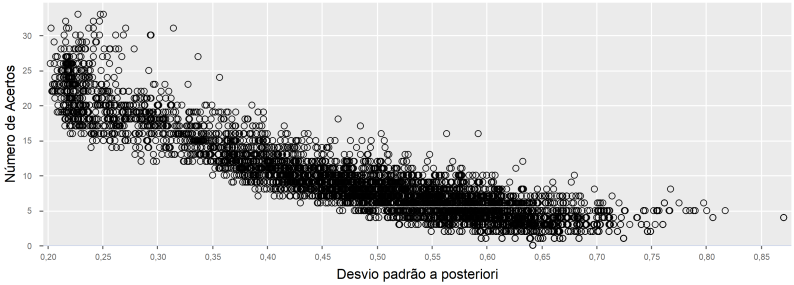

Figura 34 – Gráfico de dispersão - Proficiência estimada pelo teste adaptativo x Número acertos.

Na Figura 34 vemos o gráfico de dispersão entre o valor estimado das proficiências e o número de acertos considerando os itens administrados no teste adaptativo para cada um dos 5000 respondentes do Enem. Vemos um padrão claro, onde quanto menor o número de acertos do indivíduo, menor é o desvio padrão estimado.

É possível notar através desse estudo com dados reais que em geral a metodologia proposta, tanto considerando todos os itens do teste quanto considerando o teste adaptativo, é satisfatória e eficiente para estimar a proficiência dos indivíduos e a precisão associada. Esta conclusão é obtida uma vez que as estimativas são muito semelhantes àquelas via MCMC, na qual os parâmetros dos itens e a proficiência dos indivíduos foram estimadas conjuntamente.

5 CONCLUSÃO E TRABALHOS FUTUROS

Com o objetivo de se considerar a incerteza contida na calibração dos parâmetros dos itens na estimação das proficiências via TRI de alunos que não participam da amostra de calibração, discutimos uma nova metodologia para se obter a distribuição *a posteriori* das proficiências sem considerar os parâmetros dos itens conhecidos, no qual as únicas aproximações utilizadas são Monte Carlo e quadratura unidimensional.

Através de estudos de simulação, mostrou-se que esta abordagem chega em resultados próximos à metodologia MCMC de estimação conjunta dos parâmetros dos itens e das proficiências dos indivíduos. O que indica que temos uma estimação satisfatória e eficiente. Uma vez que as metodologias tradicionais utilizadas em testes adaptativos também consideram os parâmetros dos itens conhecidos após a calibração, a metodologia proposta de estimação das proficiências foi estendida para o contexto de testes adaptativos, no qual as proficiências são estimadas de forma iterativa a cada resposta do indivíduo. Além disso propomos um novo método para escolha adaptativa de itens e um novo critério de parada do teste.

Nos estudos de simulação, mostrou-se que os testes adaptativos em geral são muito úteis, uma vez que a proficiência do indivíduo pode ser estimada de forma satisfatória e eficiente, com menos itens respondidos considerando um critério de escolha de itens e de parada do teste adequados e intuitivos. Em alguns casos vimos que alguns indivíduos responderam a pouquíssimos itens, o que levou a uma grande diferença entre a proficiência estimada com todos os itens do teste e proficiência

estimada através do teste adaptativo, porém esses indivíduos correspondem a uma pequena parcela de respondentes, apenas 3% no caso da aplicação nos dados do Enem. É necessário investigar mais a fundo porque critério de parada proposto parou tão cedo para esses indivíduos específicos. Pensando em alguma solução, o aumento do número mínimo de itens que os indivíduos obrigatoriamente precisam responder talvez ajudasse a resolver esse problema.

Em geral as proficiências dos indivíduos estimadas através da administração de todos os itens do teste e as proficiências estimadas pelo teste adaptativo são muito semelhantes, uma vez que escolhemos de forma eficiente quais itens seriam administrados a cada indivíduo e o momento ideal de parada do teste.

Como trabalhos futuros, é importante a aplicação da metodologia proposta para diferentes áreas do conhecimento analisadas no Enem (Ciências Humanas, Ciências da Natureza e Linguagens e suas Tecnologias) e diferentes anos de aplicação das provas Enem. Outro desenvolvimento interessante, motivado pelo trabalho de Pena, Costa e Oliveira (2018), seria criar um teste adaptativo eficiente utilizando a modelagem proposta que permita ao respondente a escolha do item a ser respondido estabelecidas algumas restrições, uma vez que vários estudos forneceram evidências de que a permissão da escolha do item respondido pelos indivíduos tem um impacto positivo em termos de desenvolvimento educacional (BRIGHAM, 1979; BALDWIN; MAGJUKA; LOHER, 1991; CORDOVA; LEPPER, 1996). Esses estudos indicaram que permitir que os alunos escolham quais perguntas responder aumenta a motivação e o engajamento no processo de aprendizagem e a confiança na hora de realização do teste (JENNINGS et al., 1999).

REFERÊNCIAS BIBLIOGRÁFICAS

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

BAKER, F. B. *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, 2001. <https://eric.ed.gov/?id=ED458219>.

BAKER, F. B.; KIM, S.-H. *Item response theory: Parameter estimation techniques*. New York: CRC Press, 2004.

BALDWIN, T. T.; MAGJUKA, R. J.; LOHER, B. T. The perils of participation: Effects of choice of training on trainee motivation and learning. *Personnel psychology*, Wiley Online Library, v. 44, n. 1, p. 51–65, 1991.

BARBETTA, P.; ANDRADE, D.; BORGATTO, A. Análise de provas do enem segundo modelos de tri multidimensionais. *CONBRATRI II, Salvador–BA*, 2011.

BÉGUIN, A. A.; GLAS, C. A. Mcmc estimation and some model-fit analysis of multidimensional irt models. *Psychometrika*, Springer, v. 66, n. 4, p. 541–561, 2001.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, Addison-Wesley, 1968.

BOCK, R. D.; LIEBERMAN, M. Fitting a response model forn dichotomously scored items. *Psychometrika*, Springer, v. 35, n. 2, p. 179–197, 1970.

BRIGHAM, T. Some effects of choice on academic performance. *Choice and perceived control*, Lawrence Erlbaum Hillsdale, NJ, p. 131–141, 1979.

CARLO, C. M. Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB*, v. 581, 2004. <https://www3.ime.usp.br/ jstern/miscellanea/LabSimulacao/Walsh04.pdf>.

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.

CHILDS, R. A.; OPPLER, S. H. Implications of test dimensionality for unidimensional irt scoring: An investigation of a high-stakes testing program. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 60, n. 6, p. 939–955, 2000.

CORDOVA, D. I.; LEPPER, M. R. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of educational psychology*, American Psychological Association, v. 88, n. 4, p. 715, 1996.

COSTA, D. R. Métodos estatísticos em testes adaptativos informatizados. *Mestrado em Estatística, – Departamento de Métodos Estatísticos, Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro*, v. 9, 2009.

GAMERMAN, D.; LOPES, H. F. *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*. [S.l.]: Chapman and Hall/CRC, 2006.

GONÇALVES, F. B.; DIAS, B. da C. C.; SOARES, T. M. Bayesian item response model: a generalized approach for the abilities' distribution using mixtures. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 88, n. 5, p. 967–981, 2018.

HABERMAN, S. J. Identifiability of parameters in item response models with unconstrained ability distributions. *ETS Research Report Series*, v. 2005, p. i–22, 12 2005.

JENNINGS, M. et al. The test-takers' choice: an investigation of the effect of topic on language-test performance. *Language Testing*, Sage Publications Sage CA: Thousand Oaks, CA, v. 16, n. 4, p. 426–456, 1999.

JÚNIOR, F. d. J. M. Sistemática para a implantação de testes adaptativos informatizados baseados na teoria da resposta ao item. *Tese de Doutorado, Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Engenharia de Produção, Florianópolis, SC*, 2011.

KIM, J. K.; NICEWANDER, W. A. Ability estimation for conventional tests. *Psychometrika*, Springer, v. 58, n. 4, p. 587–599, 1993.

LAWLEY, D. N. Xxiii.—on problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, Royal Society of Edinburgh Scotland Foundation, v. 61, n. 3, p. 273–287, 1943.

LINACRE, J. Computer-adaptive testing cat: A methodology whose time has come. *MESA Psychometric Laboratory University of Chicago, MESA Memorandum*, n. 69, 2000. <https://www.rasch.org/memo69.pdf>.

LINDEN, W. J. van der. Bayesian item selection criteria for adaptive testing. *Psychometrika*, Springer, v. 63, n. 2, p. 201–216, 1998.

LINDEN, W. J. van der. *Handbook of item response theory, volume one: models*. Monterey, California: Chapman and Hall/CRC - Statistics in the Social and Behavioral Sciences Series, 2016.

LINDLEY, D. V. Approximate bayesian methods. *Trabajos de estadística y de investigación operativa*, Springer, v. 31, n. 1, p. 223–245, 1980.

LORD, F. A theory of test scores. *Psychometric monographs*, n. 7, 1952. <https://www.psychometricsociety.org/sites/default/files/pdf/MN07.pdf>.

LORD, F. M. Robbins-monro procedures for tailored testing. *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 31, n. 1, p. 3–31, 1971.

LORD, F. M. *Applications of item response theory to practical testing problems*. New York: Routledge, Taylor & Francis Group, 2012.

LORD, F. M.; NOVICK, M. R. *Statistical theories of mental test scores*. [S.l.]: IAP, 2008.

MIGON, H. S.; GAMERMAN, D.; LOUZADA, F. *Statistical inference: an integrated approach*. New York: Chapman and Hall/CRC, 2014.

MISLEVY, R. J. Bayes modal estimation in item response models. *Psychometrika*, Springer, v. 51, n. 2, p. 177–195, 1986.

MISLEVY, R. J.; STOCKING, M. L. A consumer's guide to logist and bilog. *Applied psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 13, n. 1, p. 57–75, 1989.

MOSIER, C. I. Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, American Psychological Association, v. 47, n. 4, p. 355, 1940.

OWEN, R. J. A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 70, n. 350, p. 351–356, 1975.

PATZ, R. J.; JUNKER, B. W. A straightforward approach to markov chain monte carlo methods for item response models. *Journal of educational and behavioral Statistics*, Sage Publications, v. 24, n. 2, p. 146–178, 1999.

PENA, C. S.; COSTA, M. A.; OLIVEIRA, R. P. B. A new item response theory model to adjust data allowing examinee choice. *PloS one*, Public Library of Science, v. 13, n. 2, p. e0191600, 2018.

PRIMI, R. et al. The use of the bi-factor model to test the uni-dimensionality of a battery of reasoning tests. *Psicothema*, v. 25, n. 1, p. 115–122, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. ISBN 3-900051-07-0. <http://www.R-project.org/>.

RASCH, G. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche, Oxford, England, 1960.

RICHARDSON, M. W. The relation between the difficulty and the differential validity of a test. *Psychometrika*, Springer, v. 1, n. 2, p. 33–49, 1936.

ROBERT, C.; CASELLA, G. *Monte Carlo statistical methods*. New York: Springer Science & Business Media, 2013.

RUDNER, L. M. An on-line, interactive, computer adaptive testing tutorial. *ERIC Clearinghouse on Assessment and Evaluation*, 1998. <http://EdRes.org/scripts/cat>.

SAMEJIMA, F. A comment on birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, Springer, v. 38, n. 2, p. 221–233, 1973.

SWAMINATHAN, H.; GIFFORD, J. A. Bayesian estimation in the three-parameter logistic model. *Psychometrika*, Springer, v. 51, n. 4, p. 589–601, 1986.

THURSTONE, L. L. Attitudes can be measured. *American Journal of Sociology*, University of Chicago Press, v. 33, n. 4, p. 529–554, 1928.

TSUTAKAWA, R. K.; JOHNSON, J. C. The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, Springer, v. 55, n. 2, p. 371–390, 1990.

TUCKER, L. R. Maximum validity of a test with equivalent items. *Psychometrika*, Springer, v. 11, n. 1, p. 1–13, 1946.

VALLE, R. d. C. The construction and interpretation of knowledge scales - general considerations and a view of what has been done in saresp. 27th Annual IAEA Conference, Rio de Janeiro, 2001.

VARADHAN, R. et al. *dfoptim: Derivative-Free Optimization*. [S.l.], 2018. R package version 2018.2-1. <https://CRAN.R-project.org/package=dfoptim>.

WAINER, H. et al. *Computerized adaptive testing: A primer*. New York: Routledge, Taylor & Francis Group, 2000.

WIBERG, M. An optimal design approach to criterion-referenced computerized testing. *Journal of educational and behavioral statistics*, SAGE Publications Sage CA: Los Angeles, CA, v. 28, n. 2, p. 97–110, 2003.

APÊNDICE A

Neste apêndice estão os cálculos para se obter a Informação Observada e Esperada de Fisher para o modelo Logístico de três Parâmetros referentes às seções 1.4.1 e 2.4.1.

Informação de Fisher Observada e Esperada

A função de verossimilhança associada à resposta do i-ésimo item é dada por:

$$L(\theta; y_i) = P_i(\theta)^{y_i}[1 - P_i(\theta)]^{1-y_i}$$

A função de informação do item é a segunda derivada do log da verossimilhança. Como esse procedimento representa a curvatura da função de verossimilhança observada em θ , esse método permite avaliar a magnitude do erro associado à habilidade estimada em relação aos parâmetros do i-ésimo item. Para melhor entendimento, serão descritos os cálculos das funções de informação esperada e de Fisher. O logaritmo da Verossimilhança é dado por:

$$\ln L(\theta; y_i) = y_i ln P_i(\theta) + (1 - y_i) ln[1 - P_i(\theta)$$

A medida de Informação Observada do i-ésimo item consiste em:

$$\begin{aligned} J_{y_i}(\theta) &= -\frac{\partial^2}{\partial\theta^2} \ln L(\theta; y_i) \\ &= \frac{y_i P''_i(\theta)}{P_i(\theta)} + \frac{y_i [P'_i(\theta)]^2}{P_i^2(\theta)} - \frac{P''_i(\theta)[y_i - 1]}{1 - P_i(\theta)} - \frac{P'_i(\theta)]^2 [y_i - 1]}{[1 - P_i^2(\theta)]^2} \end{aligned}$$

Já a medida de Informação Esperada (também conhecida como Informação de Fisher) do i-ésimo item é dada por:

$$I_{Y_i}(\theta) = E_{\mathrm{y}_i|(\theta)} = \left[-\frac{\partial^2}{\partial\theta^2} \ln L(\theta; y_i) \right]$$

Como $Y_i \sim Bernoulli(P_i)$, a Informação de Fisher do i-ésimo item ($IF_{y_i}(\theta)$) será igual a:

$$\begin{aligned}
&= E_{y_i|\theta} \left[\frac{y_i {P''}_i(\theta)}{P_i(\theta)} + \frac{y_i [{P'}_i(\theta)]^2}{P_i^2(\theta)} - \frac{{P''}_i(\theta)[y_i - 1]}{1 - P_i(\theta)} - \frac{[{P'}_i(\theta)]^2 [y_i - 1]}{[1 - P_i(\theta)]^2} \right] \\
&= -\frac{P_i(\theta){P''}_i(\theta)}{P_i(\theta)} + \frac{P_i(\theta)[{P'}_i(\theta)]^2}{P_i^2(\theta)} - \frac{{P''}_i(\theta)[P_i(\theta) - 1]}{1 - P_i(\theta)} - \frac{[{P'}_i(\theta)]^2 [P_i(\theta) - 1]}{[1 - P_i(\theta)]^2} \\
&= -{P''}_i(\theta) + \frac{[{P'}_i(\theta)]^2}{P_i(\theta)} + \frac{{P''}_i(\theta)[1 - P_i(\theta)]}{1 - P_i(\theta)} + \frac{[{P'}_i(\theta)]^2 [1 - P_i(\theta)]}{[1 - P_i(\theta)]^2} \\
&= -{P''}_i(\theta) + \frac{[{P'}_i(\theta)]^2}{P_i(\theta)} + {P''}_i(\theta) + \frac{[{P'}_i(\theta)]^2}{[1 - P_i(\theta)]} \\
&= \frac{[{P'}_i(\theta)]^2}{P_i(\theta)} + \frac{[{P'}_i(\theta)]^2}{[1 - P_i(\theta)]} = \frac{[{P'}_i(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}
\end{aligned}$$

Sob o Modelo da ogiva normal de 3 parâmetros (3PNO) dado pela Equação em (2.1), tem-se que:

$$P_i{}'(\theta) = \frac{\partial}{\partial\theta} [c_i + (1 - c_i)\Phi(a_i\theta_j - b_i^*)] = (1 - c_i)a_i\phi(a_i\theta_j - b_i^*) \quad (5.1)$$

Portanto:

$$\begin{aligned}
IF_{Y_i}(\theta) &= E_{\mathrm{y}_i|(\theta)} = \frac{[P_i{}'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]} \qquad (5.2) \\
&= \frac{[(1 - c_i)a_i\phi(a_i\theta_j - b_i^*)]^2}{[c_i + (1 - c_i)\Phi(a_i\theta_j - b_i^*)][1 - c_i - (1 - c_i)\Phi(a_i\theta_j - b_i^*)]}
\end{aligned}$$

APÊNDICE B

Indepêndencia Condicional de x e y

Neste apêndice está a prova da independência condicional de x e y dado ξ, onde x é o vetor de respostas de um novo indivíduo e y são os dados da calibração referente à Equação 2.6.

Seja θ_x a proficiência de um novo respondente fornecendo as respostas x e θ_y as proficiências dos alunos dos dados de calibração fornecendo as respostas y, então:

$$
\begin{aligned}
p(x,y|\xi) &= \int p(x,y,\theta_x,\theta_y|\xi)d\theta_x d\theta_y \\
&= \int p(x,y|\theta_x,\theta_y,\xi)p(\theta_x,\theta_y|\xi)d\theta_x d\theta_y \\
&= \int p(x|\theta_x,\xi)p(y|\theta_y,\xi)p(\theta_x|\xi)p(\theta_y|\xi)d\theta_x d\theta_y \\
&= \int p(x|\theta_x,\xi)p(\theta_x|\xi)d\theta_x \int p(y|\theta_y,\xi)p(\theta_y|\xi)d\theta_y \\
&= p(x|\xi)p(y|\xi)
\end{aligned}
$$

APÊNDICE C

Neste apêndice serão apresentadas algumas definições que facilitam o entendimento das metodologias propostas neste trabalho.

Definição (Quadratura Gaussiana): A quadratura numérica, consiste em obter um número real que aproxime o valor de uma integral definida:

$$l(f) = \int_a^b f(x)dx. \tag{5.3}$$

Assumindo que f é contínua e, geralmente, assumindo também que f é suave. Assim, a integral pode ser aproximada por uma outra função que assume um número finito de pontos. Dessa forma o problema de obter a integral de uma função contínua será substituído pela obtenção da soma das áreas de um número finito p de retângulos.

$$l(f) = \int_a^b f(x)dx = \lim_{\mathrm{n}\to\infty} \sum_{k=1}^{p} g(x_k)\Delta x \tag{5.4}$$

em que $\Delta x = \frac{b-a}{p}$ e x_k são pontos amostrais no intervalo $[a, b]$ também chamados de nós de integração.

Na quadratura numérica, portanto aproximamos $l(f)$ por uma soma finita, na qual f é amostrada em alguns pontos, a justificativa é o resultado básico de cálculo que diz que as somas parciais convergem para o valor exato da integral.