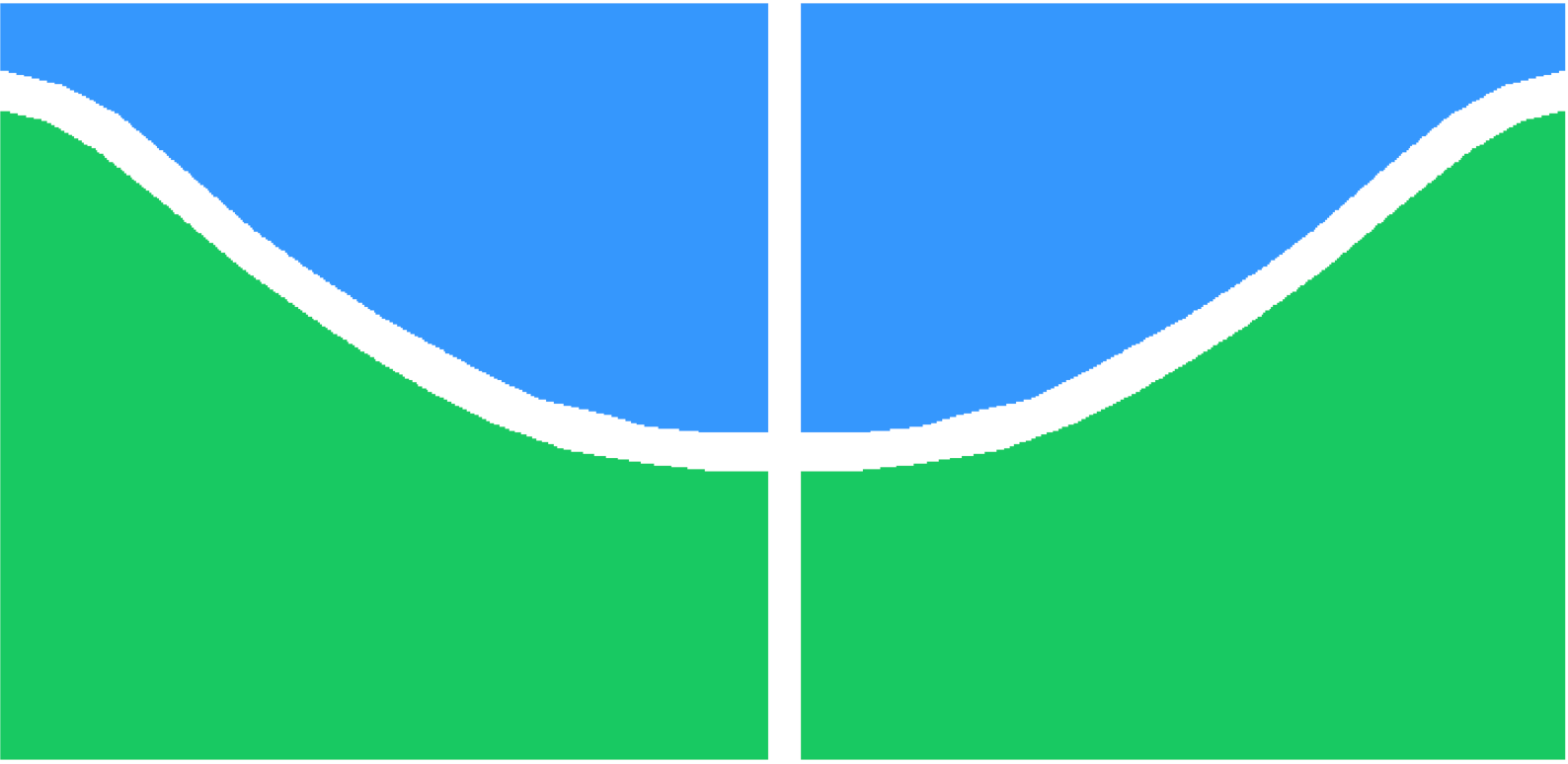


Universidade de Brasília - UnB

Instituto de Ciências Exatas - IE

Departamento de Estatística - EST

Uso do Tempo de Resposta para Melhorar a Convergência do Algoritmo de Testes Adaptativos Informatizados

Autor: Antonio Geraldo Pinto Maia Júnior

Orientador: Prof. Gustavo L. Gilardoni

Brasília, DF

2015

Antonio Geraldo Pinto Maia Júnior

Uso do Tempo de Resposta para Melhorar a Convergência do Algoritmo de Testes Adaptativos Informatizados

Dissertação submetida ao programa de Pós-Graduação em Estatísticada Universidade de Brasília, como requisito parcial para obtenção do Título de Mestre em Estatística.

Universidade de Brasília - UnB

Instituto de Ciências Exatas - IE

Departamento de Estatística - EST

Orientador: Prof. Gustavo L. Gilardoni

Brasília, DF

2015

Este trabalho é dedicado aos futuros estudantes, que terão oportunidade de serem mais bem avaliados com as novas ferramentas e tecnologias que surgirão.

Agradecimentos

À minha mãe, que sempre me incentivou a estudar, para eu vencer na vida através do esforço próprio e méritos pessoais. À minha amada esposa, cuja paciência me foi necessária, para a conclusão deste trabalho. Aos ilustres professores do Departamento de Estatística da UnB, dos quais recebi ensinamento e orientações tão importantes. Ao professor Gustavo Gilardoni, especialmente, pela confiança, pelo incentivo e pela grandeza de, sabiamente, repassar um pouco de seu profundo conhecimento. Aos meus amigos da UnB, pela parceria, pela amizade conquistada, principalmente, pela superação das dificuldades, pela felicidade experimentada e pela vitória conquistada.

"Nós somos aquilo que fazemos repetidas vezes, repetidamente. Excelência, então, não é um modo de agir, mas sim, um hábito."

(Aristóteles)

Resumo

O presente trabalho tem como objetivo central melhorar os Testes Adaptativos Informatizados (Computerized Adaptative Tests, CATs na sigla, em inglês) clássicos, que são aqueles administrados por computador e que ajustam os itens do teste à medida que ele é realizado. Isso é possível, pois, dada a resposta do respondente, estima-se a sua habilidade momentânea, obtendo-se o próximo item a ser administrado, com base em um critério estatístico (Máxima Informação, Máxima Informação Global ou Máxima Informação Esperada).

Para isso, inseriu-se a covariável Tempo de Resposta ao modelo. Pois, acreditou-se que há informação nessa covariável e, portanto, ao se considerá-la, o teste pode ser encurtado, melhorando, assim, a convergência do algoritmo.

Nessa perspectiva, fez-se uma revisão bibliográfica de TRI (sigla de Teoria de Resposta ao Item) e CAT, para se estruturar o novo modelo com a covariável Tempo de Resposta, calculando-se todas as equações que serão utilizadas na aplicação.

Por fim, a aplicação com dados simulados concluiu nosso estudo, pois, ao comparar a convergência do algoritmo de um CAT tradicional em relação ao novo CAT, observou-se que os objetivos do presente trabalho foram cumpridos.

Palavras-chaves: CAT. TRI. Tempo de Resposta.

Abstract

Computerized adaptive tests (CATs) are tests administered by computer which adjust the test items as the test is carried out. This work proposes to improve CATs by taking into account the time that the respondents use to answer the different questions to obtain provisional estimates of their ability in order to choose the next item.

This information is used to modify the classical criteria (maximal information, overall maximum information or maximum information expected). It is believed that the use of this covariate may improve the convergence of the CAT algorithm, thus allowing for shorter tests.

The dissertation presents a review of TRI and CAT and the new model which takes into account the response time time.

An application using simulated data is used to compare the convergence of a traditional CAT algorithm and that of the model using the response time.

Key-words: TRI. CAT. Response Time

Lista de ilustrações

Lista de tabelas

Sumário

Introdução

Tem-se percebido, nos últimos anos, a disseminação em larga escala de computadores. E, naturalmente, o uso desse recurso é fundamental nos mais diversos setores de atividades.

Com a inserção de um ambiente informatizado nas escolas, o desenvolvimento de novas ferramentas de ensino-aprendizagem tornou-se propício. A criação de testes assistidos por computador é um exemplo de iniciativas que estão avançando bastante.

As crescentes pesquisas para a implementação desses testes fizeram surgir os Testes Adaptativos Informatizados, que denominaremos de CAT, no presente trabalho.

Veja a reportagem da Folha de São Paulo, em Janeiro de 2015:

"O novo ministro da Educação está disposto a promover uma verdadeira revolução no Exame Nacional do Ensino Médio. Ele declarou em entrevista à Folha que pretende levar à presidente Dilma Rousseff um projeto que torna o ENEM uma prova online, além da possibilidade de aplicá-la mais de uma vez durante o ano. A proposta tem como objetivo principal acabar com o ENEM da forma que é aplicado hoje, em um único fim de semana para todos os candidatos. Ao digitalizar a prova, o aluno teria uma janela de vários dias para comparecer a um posto credenciado e prestar a prova em um computador, abolindo de vez o exame em papel. Ao tornar o ENEM digital o sistema de ensino teria outro ganho, que é a minimização de fraudes e a objetivação do exame: cada prova seria única, composta por questões escolhidas em um enorme banco de dados do MEC."

O grande objetivo em um CAT é montar uma avaliação adaptativa que não prejudique nenhum respondente e cujo tamanho seja o ideal para estimarmos a habilidade do participante. Nesse sentido, a prova precisa ser personalizada para cada participante e ela precisa ser comparável com todas as outras provas dos demais respondentes.

O presente trabalho objetiva contribuir no aprimoramento desses testes, inserindo a covariável Tempo de Resposta. Em um CAT tradicional, a escolha de um próximo item depende exclusivamente das respostas dos itens anteriores. E a nossa pesquisa pretende demonstrar que há informação também no tempo de resposta do respondente nos itens respondidos corretamente, influenciando a escolha do próximo item, melhorando a convergência do algoritmo.

Objetivos

Objetivo Geral

Criar um modelo estatístico que leve em conta a covariável Tempo de Resposta, calculando a nova função de verossilhança, a informação esperada e observada assim como a medida de Kullback-Leibler.

Objetivos Específicos

- Implementar 2 algoritmos de testes adaptativos informatizados: um sem utilizar a covariável Tempo de Resposta e outro utilizando tal covariável.
- Comparar a convergência desses dois algoritmos (através do número de questões necessárias para a parada do teste), utilizando como critério de parada a precisão do estimador.

Organização do trabalho

O presente trabalho foi dividido em 2 partes. Na primeira, fez-se uma revisão da Teoria de Resposta ao Item (TRI) e de um Teste Adaptativo Informatizado (CAT). Na segunda, propõe-se uma nova modelagem, uma aplicação com dados simulados e o desenvolvimento da programação utilizada na simulação. A primeira parte foi subdividida em 2 capítulos, a segunda parte em 3.

Teoria de Resposta ao Item

A Teoria de Resposta ao Item reúne um conjunto de modelos estatísticos que relacionam um ou mais traços latentes (não observados) de um indivíduo com a probabilidade deste dar uma certa resposta a um item. Como nosso estudo de TRI será voltado para a área educacional, entenderemos o traço latente como a habilidade ou proficiência em alguma área. Por exemplo, matemática, português, física, dentre outras. Para padronizar a linguagem deste trabalho, substituiremos a expressão traço latente por habilidade[1] e representaremo-la por θ.

A probabilidade de um respondente acertar um item é modelada como função da habilidade do respondente e dos parâmetros que expressam certa propriedade dos itens. Respondentes e itens são posicionados na mesma escala, como se fosse em uma mesma régua. Quanto maior a habilidade do candidato, maior a probabilidade de ele acertar o

[1] É proficiência do respondente, ou seja, característica do indivíduo que não pode ser observada diretamente. Esse tipo de variável deve ser inferida a partir da observação de variáveis secundárias que estejam relacionadas a ela.

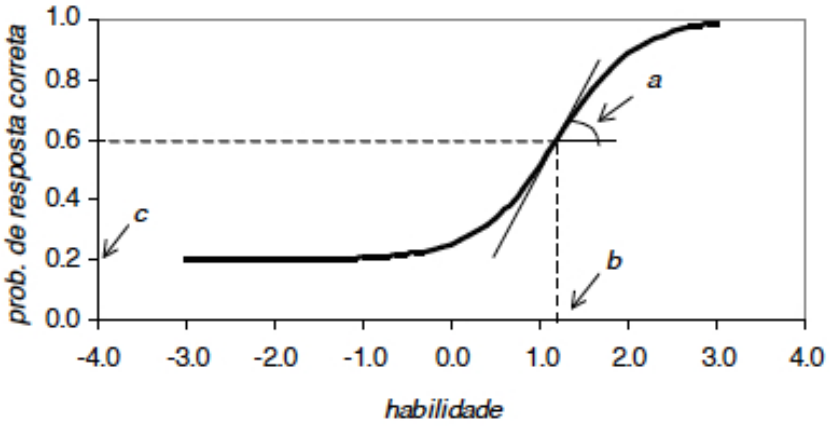


Figura 1: Curva Característica do Item - CCI

item, chamado de modelo acumulativo, na literatura. Um modelo adequado, que contempla todas essas propriedades e que utilizaremos em nosso trabalho é o modelo logístico unidimensional de 3 parâmetros (ML3), também conhecido como modelo de Birnbaum de 3 parâmetros (1968), e ele é expresso por

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

com $i = 1, 2, ..., I$ e $j = 1, 2, ..., N$, onde:

- U_{ji} é uma variável dicotômica que assume os valores 1, quando o indivíduo j responde corretamente o item i, ou 0 quando o indivíduo j não responde corretamente ao item i;
- θ_j representa a habilidade do j-ésimo respondente;
- $P(U_{ji} = 1|\theta_j)$ é a probabilidade de um indivíduo j com habilidade θ_j responder corretamente o item i;
- a_i é o parâmetro de discriminação do item i (observemos o posicionamento de a na figura 1), com valor proporcional à declividade da Curva Característica do Item (CCI) no ponto de inflexão b_i. Assim, itens com $a < 0$ não são esperados com esse modelo, uma vez que indicariam que a probabilidade de responder corretamente o item diminui com o aumento da habilidade. Baixos valores de a_i indicam que o item tem pouco poder de discriminação, uma vez que habilidades bastante diferentes em torno de b_i têm probabilidades bem próximas de acertar o item. Em contrapartida, valores altos de a_i fazem com a CCI do item i seja bem íngrime, fazendo com que o poder de discriminação seja fortíssimo, pois, basicamente, os respondentes são subdivididos em dois grupos: os que possuem habilidade abaixo e acima de b_i;

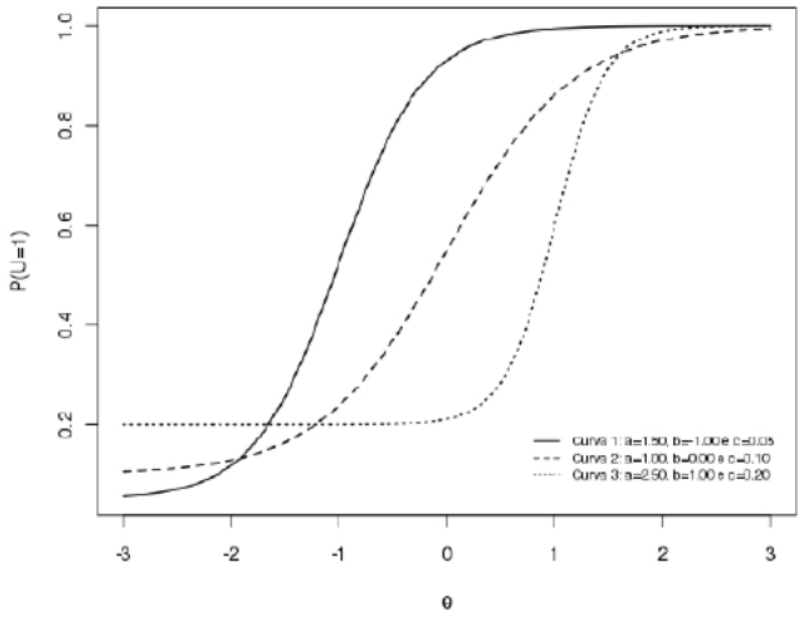


Figura 2: Curva característica de três itens em que (i) a curva 1 apresenta $a = 1,5$, $b = 1$ e $c = 0,05$; (ii) a curva 2 apresenta $a = 1$, $b = 0$ e $c = 0,1$; (iii) e a curva 3 apresenta $a = 2,5$, $b = 1$ e $c = 0,2$

- b_i é o parâmetro de dificuldade do item i, medido na mesma escala da habilidade θ_j (observemos a indicação de b na figura 1, percebamos que está no mesmo eixo de θ e que ele é a abscissa relacionada à mudança de concavidade da CCI). Uma interpretação interessante é que ele representa o ponto na escala da habilidade onde a probabilidade de acertar o item i é $0,5$, desde que c_i, parâmetro que será comentado a seguir, seja igual a zero;

- c_i é o parâmetro do item que representa a probabilidade de indivíduos com baixa habilidade responderem corretamente o item i (muitas vezes referido como a probabilidade de acerto casual, observemos na figura 1, que respondentes com baixíssima habilidade, têm a probabilidade c de acertar o item, e que em um item com 5 alternativas, c será $0,2$). D é um fator de escala, constante e igual a 1. Utiliza-se o valor 1,702 quando desejamos que a função logística forneça resultados semelhantes ao da função Ogiva Normal.

Observemos a figura 2, que possui curvas características de 3 itens, e percebamos a influência dos parâmetros a, b e c nos correspondentes gráficos.

Vários pesquisadores destacam-se no estudo de TRI, mas sem dúvida, a obra de Andrade, Tavares e Valle (2000) merece atenção especial, pelas inúmeras citações em outros artigos, dissertações e teses, pela clareza como os temas são abordados, pelas referências bibliográficas, pelo cuidado com a notação e com a escrita. Aos interessados em estudar TRI, recomenda-se iniciar por essa obra. O trabalho de Embretson (2013) tam-

bém merece destaque, pois é um livro recente que além de ter a teoria necessária para se aprofundar nesse estudo, ainda possui 4 capítulos destinados à aplicação.

No Capítulo 1, o estudo de TRI será mais detalhado.

Teste Adaptativo Informatizado

Segundo Costa (2009), um Teste Adaptativo Informatizado, Computerized Adaptative Test (CAT), em inglês é aquele administrado pelo computador que pretende encontrar um teste ótimo para cada respondente. Para atingir isso, a habilidade do respondente é estimada iterativamente durante a administração do teste.

Como citado por Wainer (2000), "a noção básica de um CAT é imitar o que um sábio examinador faria". Um CAT tem por finalidade administrar itens, de um banco de itens previamente calibrados (esse assunto será aprofundado na seção 1.2.1). No presente trabalho, esses itens são selecionados de acordo com o modelo TRI. Ao contrário dos testes tradicionais (papel-e-caneta), em um CAT, diferentes respondentes podem receber diferentes testes de tamanhos variados.

Diversos CATs estão em pleno funcionamento, tais como o Graduate Record Examination (GRE), o Test of English as a Foreign Language (TOEFL), a Armed Services Vocational Aptitude Test Battery (ASBAV). No Brasil, os DETRANs de SC e SP fazem uso de CAT em algumas avaliações e o MEC dá sinais de que em breve o maior teste aplicado em um único dia no mundo, o ENEM, deverá seguir o modelo de um CAT.

Maiores detalhes sobre CAT serão abordados no capítulo 2 desse trabalho.

Modelo com a Covariável Tempo de Resposta

O terceiro capítulo tem como proposta estruturar um modelo que leve em conta o Tempo de Resposta do item no modelo TRI, calculando-se a nova função de verossimilhança e recalculando-se as medidas de informações de Fisher, Kullback Leibler e a Máxima Informação Esperada para essa nova abordagem. Essa nova modelagem objetiva melhorar a escolha do próximo item em um CAT, utilizando além da resposta dada em itens anteriores, a informação do tempo de resposta que o candidato levou para acertar os itens respondidos até então.

Aplicação com Dados Simulados

No quarto capítulo do presente trabalho, fez-se uma aplicação por meio de dados simulados para comparar a convergência do algoritmo de um CAT tradicional (sem a utilização do tempo) com a de um CAT implementado com a nova modelagem.

Programação e Estrutura dos Algoritmos Utilizados

No Anexo deste trabalho, disponibilizaram-se os algoritmos utilizados bem como a estruturação e comentário dos mesmos para cumprirem-se os objetivos desse estudo.

Parte I

Revisão Teórica de TRI e CAT

1 Teoria de Resposta ao Item

Com base no modelo de TRI proposto na introdução desse trabalho, desenvolveu-se o seguinte estudo, que será sucinto e pretende fazer apenas uma ambientação da teoria de resposta ao item. Para um estudo mais aprofundado, além das referências já citadas, recomenda-se o trabalho de Linden e Hambleton (2013), que reúne um conjunto de artigos científicos recentes de Teoria de Resposta ao Item.

1.1 Função de Informação do Item

Uma medida bastante utilizada em conjunto com a Curva Característica do Item - CCI é a função de informação do item. Ela permite analisar quanto um item contém de informação para a medida de habilidade. Acompanhemos o seguinte raciocínio para a obtenção da função de informação de um item.

A Função de Verossimilhança associada à resposta do i-ésimo item é dada por

$$L(\theta; u_i) = P(U_i = u_i|\theta) = [P_i(\theta)]^{u_i}[1 - P_i(\theta)]^{1-u_i}. \tag{1.1}$$

O Logaritmo da Função de Verossimilhança será dado por

$$l(\theta; u_i) = \log(L(\theta; u_i)) = u_i \log[P_i(\theta)] + (1 - u_i) \log[1 - P_i(\theta)]. \tag{1.2}$$

A medida de informação observada $J_{u_i}(\theta)$ é dada por

$$\begin{aligned} J_{u_i}(\theta) &= -\frac{\partial^2}{\partial\theta^2} l(\theta; u_i) \\ &= -\frac{u_i P_i^{''}(\theta)}{P_i(\theta)} + \frac{u_i[P_i^{'}(\theta)]^2}{P_i^2(\theta)} - \frac{[u_i - 1]P_i^{''}(\theta)}{1 - P_i(\theta)} - \frac{[u_i - 1][P_i^{'}(\theta)]^2}{[1 - P_i(\theta)]^2}\,. \end{aligned} \tag{1.3}$$

A medida de informação esperada ou informação de Fisher do i-ésimo item é dada por

$$I_{U_i}(\theta) = E_{U_i|\theta}\left[-\frac{\partial^2}{\partial\theta^2} l(\theta; U_i)\right].$$

Como $U_i \sim Bernoulli(P_i)$, então $E(U_i) = P_i(\theta)$. Portanto, $I_{U_i}(\theta)$ será dada por

$$\begin{aligned} I_{U_i}(\theta) &= E_{U_i|\theta}\left[-\frac{U_i P_i^{''}(\theta)}{P_i(\theta)} + \frac{U_i[P_i^{'}(\theta)]^2}{P_i^2(\theta)} - \frac{[U_i - 1]P_i^{''}(\theta)}{1 - P_i(\theta)} - \frac{[U_i - 1][P_i^{'}(\theta)]^2}{[1 - P_i(\theta)]^2}\right] \\ &= -\frac{P_i(\theta)P_i^{''}(\theta)}{P_i(\theta)} + \frac{P_i(\theta)[P_i^{'}(\theta)]^2}{P_i^2(\theta)} - \frac{[P_i(\theta) - 1]P_i^{''}(\theta)}{1 - P_i(\theta)} - \frac{[P_i(\theta) - 1][P_i^{'}(\theta)]^2}{[1 - P_i(\theta)]^2} \\ &= \frac{[P_i^{'}(\theta)]^2}{P_i(\theta)} + \frac{[P_i^{'}(\theta)]^2}{[1 - P_i(\theta)]} = \frac{[P_i^{'}(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}\,. \end{aligned} \tag{1.4}$$

Sob o modelo exposto na equação (1), extraímos

$$P_i^{'}(\theta) = \frac{Da_i(1-c_i)e^{-Da_i(\theta-b_i)}}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2}. \tag{1.5}$$

Logo, a Informação de Fisher do item para o ML3 - equação (1) - será expressa por

$$\begin{aligned}
I_{U_i}(\theta) &= \frac{[P_i^{'}(\theta)]^2}{P_i(\theta)[1-P_i(\theta)]} = [P_i^{'}(\theta)]^2 \,.\, [P_i(\theta)]^{-1} \,.\, [1-P_i(\theta)]^{-1} \\
&= \left[\frac{Da_i(1-c_i)e^{-Da_i(\theta-b_i)}}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2}\right]^2 \cdot \left[c_i + (1-c_i)\frac{1}{1+e^{-Da_i(\theta b_i)}}\right]^{-1} \\
&\cdot \left[1-\left(c_i+(1-c_i)\frac{1}{1+e^{-Da_i(\theta-b_i)}}\right)\right]^{-1} \\
&= \frac{D^2a_i^2(1-c_i)^2e^{-2Da_i(\theta-b_i)}}{\left[1+e^{-Da_i(\theta-b_i)}\right]^4} \cdot \left[\frac{1+c_ie^{-Da_i(\theta b_i)}}{1+e^{-Da_i(\theta b_i)}}\right]^{-1} \cdot \left[\frac{e^{-Da_i(\theta-b_i)}(1-c_i)}{1+e^{-Da_i(\theta-b_i)}}\right]^{-1} \\
&= \frac{D^2a_i^2(1-c_i)^2e^{-2Da_i(\theta-b_i)}}{\left[1+e^{-Da_i(\theta-b_i)}\right]^4} \cdot \frac{1+e^{-Da_i(\theta b_i)}}{1+c_ie^{-Da_i(\theta b_i)}} \cdot \frac{1+e^{-Da_i(\theta-b_i)}}{e^{-Da_i(\theta-b_i)}(1-c_i)} \\
&= \frac{D^2a_i^2(1-c_i)e^{-Da_i(\theta-b_i)}}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2} \cdot \frac{1}{1+c_ie^{-Da_i(\theta b_i)}} \\
&= \frac{D^2a_i^2}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2} \cdot \frac{(1-c_i)e^{-Da_i(\theta-b_i)}}{1+c_ie^{-Da_i(\theta b_i)}} \\
&= \frac{D^2a_i^2}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2} \cdot \frac{1-c_i}{e^{Da_i(\theta b_i)}+c_i} \\
&= \frac{D^2a_i^2(1-c_i)}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2\left[c_i+e^{Da_i(\theta b_i)}\right]}.
\end{aligned} \tag{1.6}$$

Observando a figura 3, percebemos que o item discrimina bem o candidato em uma região limitada, em torno da inflexão b e que o resultado da equação (1.6) mostra que a informação depende diretamente de a^2. Observamos nessa figura, que quanto maior a, mais informação em torno de b o item possui. Portanto, a será considerado o parâmetro de qualidade do item. Diminuindo a, perde-se informação do item.

Segundo Andrade, Tavares e Valle (2000), o teste (conjunto dos itens) possui uma infomação, chamada Função de Informação do Teste - $FIT(\theta)$, que é simplesmente a soma das informações de todos os itens que compõem o teste, dada por $FIT(\theta) = \sum_{i=1}^{I} I_{U_i}(\theta)$. Pode-se mostrar que o erro-padrão da estimativa de θ é expresso por $EP(\theta) = \frac{1}{\sqrt{FIT(\theta)}}$.

O modelo proposto (ML3) pressupõe a unidimensionalidade do teste, isto é, a homogeneidade do conjunto de itens que supostamente devem estar medindo um único traço latente (θ). Em outras palavras, deve haver apenas uma habilidade responsável pela realização de todos os itens da prova. Segundo Andrade, Tavares e Valle (2000) parece

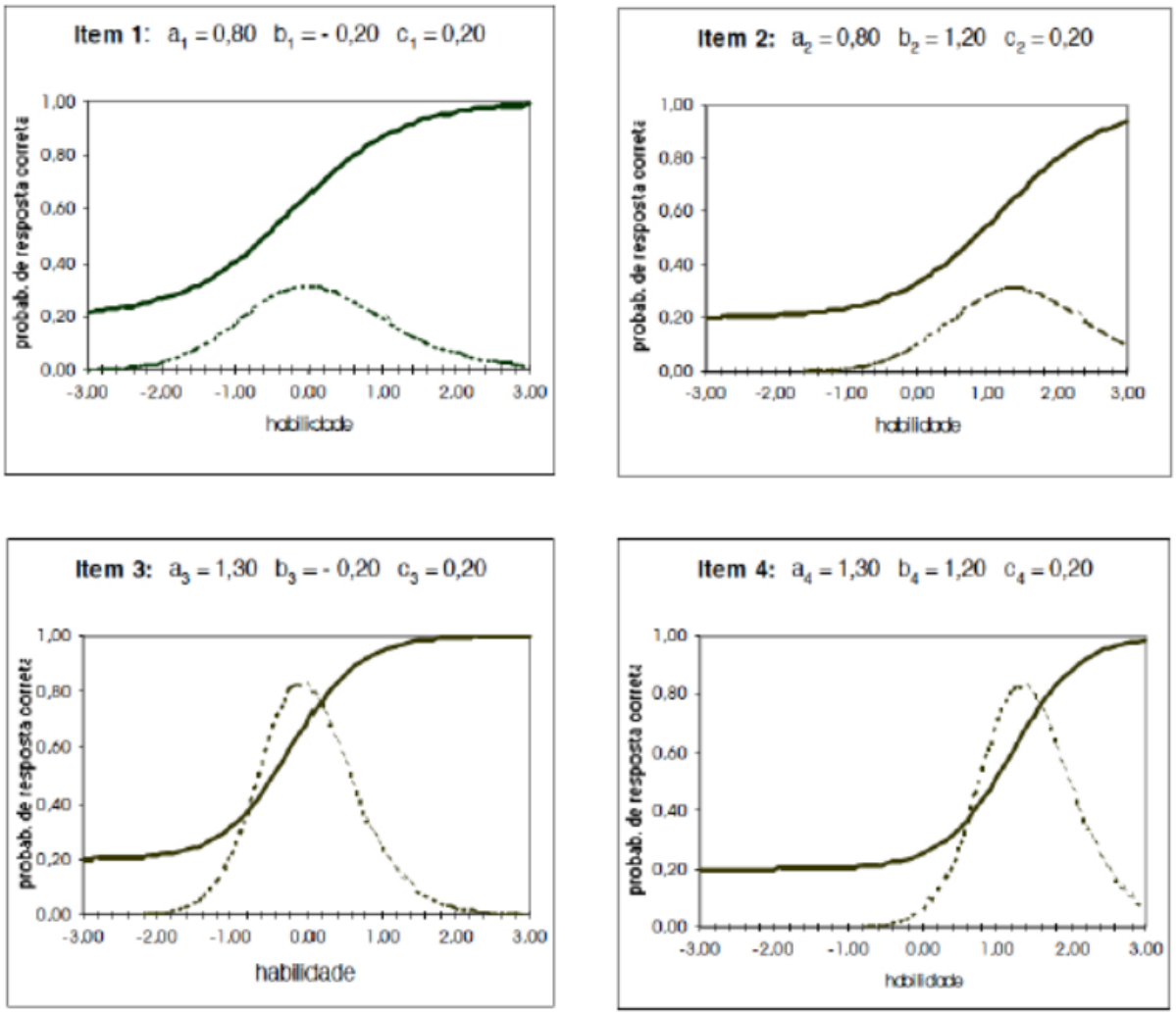


Figura 3: A curva contínua representa a CCI e a tracejada a Curva de Informação de 4 itens

claro que qualquer desempenho humano é sempre multideterminado ou multimotivado, dado que mais de um traço latente entra na execução de qualquer tarefa. Contudo, para satisfazer o postulado da unidimensionalidade, é suficiente admitir que haja uma habilidade dominante (um fator dominante) responsável pelo conjunto de itens. Uma outra suposição do modelo é a chamada independência local (ou independência condicional), a qual assume que, para uma dada habilidade, as respostas aos diferentes itens da prova são independentes. Essa suposição será fundamental para o processo de estimação dos parâmetros do modelo. Segundo Hambleton et al. (2001), a unidimensionalidade implica independência local. Portanto, itens devem ser elaborados de modo a satisfazer a suposição de unidimensionalidade.

1.2 Estimação dos Parâmetros

Essa é uma das etapas mais importantes da TRI e, como vimos no ML3, a probabilidade de acertar um determinado item depende de dois tipos de parâmetros. Um tipo relacionado ao item (a, b e c) e outro tipo relacionado ao respondente (θ). Dependendo da situação, o estatístico pode receber três situações-problema no processo de estimação dos parâmetros: i) se já conhece os parâmetros dos itens, basta estimar as habilidades dos respondentes; ii) se já conhece as habilidades dos respondentes, basta estimar os parâmetros dos itens[1] e iii) estimar os parâmetros dos itens e as habilidades dos indivíduos simultaneamente. Em grandes exames (como o ENEM, por exemplo), conduz-se o processo para a situação i), pois os itens já foram calibrados com os chamados pré-testes. Isso também acontecerá nos Testes Adaptativos Informatizados (CATs), que será estudado no próximo capítulo. Nesse sentido, é fundamental a construção de um banco de itens.

1.2.1 Construção do Banco de Itens

Entendemos que um banco de itens é considerado bem calibrado se as estimativas dos parâmetros dos itens forem adequadas e seus respectivos erros padrões forem baixos. Olea et al. (1999) destaca sete passos para a elaboração de um banco de itens:

1. Definição da estrutura do banco de itens: definem-se os tipos e os formatos de itens de acordo com as diferentes áreas de conteúdo;

2. Desenvolvimento dos itens: elaboração dos itens, onde podem-se aproveitar itens pré-existentes ou construírem-se novos itens, procedendo com a análise de conteúdo clássica, segundo Pasquali (1996) e Pasquali (1998);

3. Coleta de dados: definição do processo de coleta de dados para a calibração dos parâmetros dos itens por meio da TRI;

4. Administração dos itens: todos os itens deverão ser respondidos para a calibração dos parâmetros, mas não necessariamente pelos mesmos indivíduos, ainda mais porque, em geral, o banco de itens é extenso. Essa aplicação poderá ser feita por um teste administrado por computador ou por um teste tradicional "papel e lápis". Segundo Segall (2005), vários estudos encontraram diferenças insignificantes no funcionamento da resposta do item devido ao modo de administração (computador ou teste tradicional "papel e lápis"). Segall (2005) destaca ainda que o modo de coleta de dados por meio do formato tradicional "papel e lápis" é mais rápido e tem um custo menor do que a coleta feita por meio do computador;

[1] Em TRI, o processo de estimação dos parâmetros dos itens é conhecido como calibração

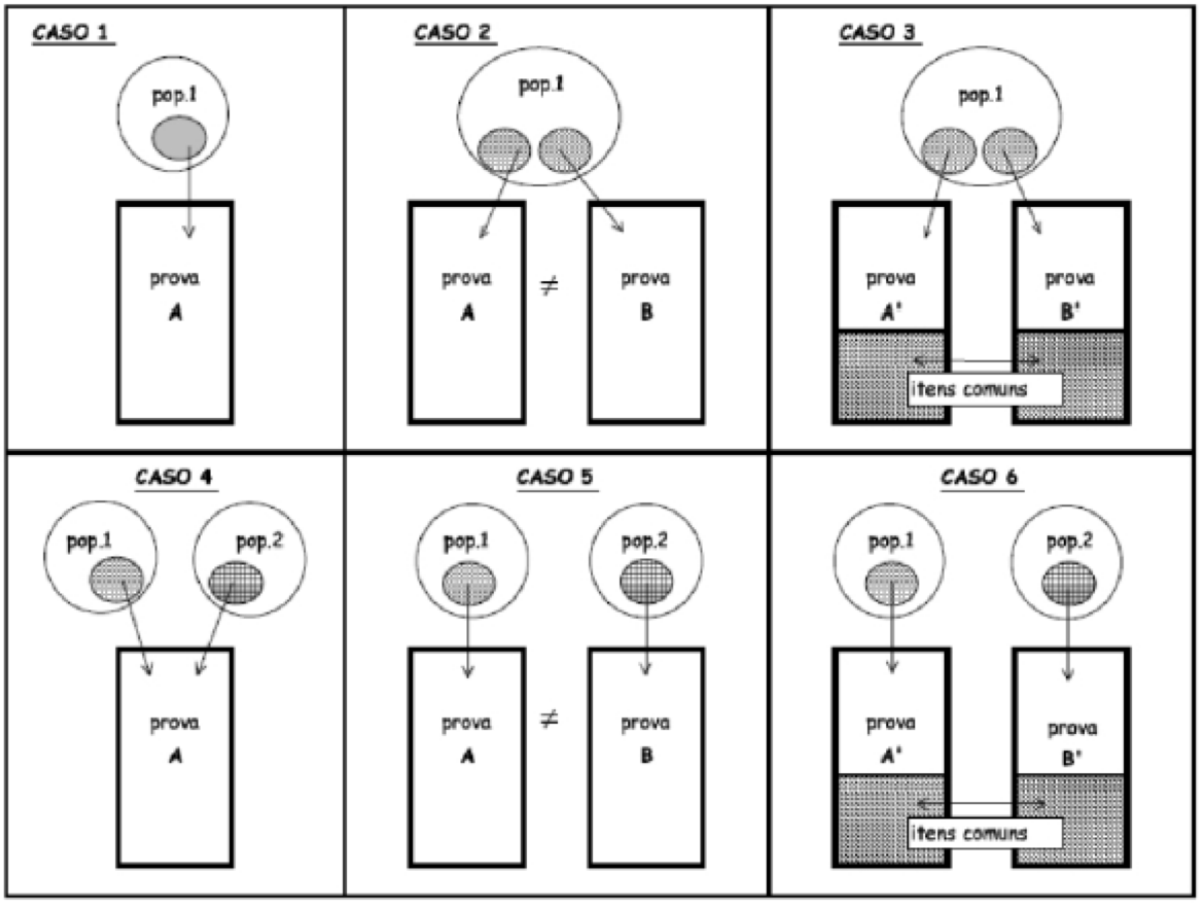


Figura 4: Representação gráfica das seis formas diferentes de aplicações de testes (Fonte: Andrade, Tavares e Valle (2000))

5. Análise dos itens: após a coleta de uma amostra suficiente de respostas, é realizada uma análise preliminar dos itens utilizando-se recursos da TRI;

6. Calibração dos itens: processo de estimação dos parâmetros dos itens por meio da TRI, o qual será melhor detalhado na Seção 1.3;

7. Armazenamento de informação: os parâmetros estimados dos itens pela TRI devem ser armazenados juntamente com os itens no banco de itens.

Para calibrar os itens, é necessário que eles já tenham sido aplicados segundo um teste tradicional. De acordo com Andrade, Tavares e Valle (2000), seis formas diferentes de aplicações de testes podem ser encontradas na prática, as quais são ilustradas na Figura 4 para uma e duas populações (ou grupos):

1. Uma única população fazendo uma única prova;

2. Uma única população, dividida em dois ou mais subgrupos, fazendo duas provas totalmente distintas (nenhum item comum);

3. Uma única população, dividida em dois ou mais subgrupos, fazendo duas provas parcialmente distintas (com alguns itens comuns);

4. Duas ou mais populações, com características diferentes, fazendo uma única prova;

5. Duas ou mais populações, com características diferentes, fazendo duas provas totalmente distintas (nenhum item comum);

6. Duas ou mais populações, com características diferentes, fazendo duas provas parcialmente distintas (com alguns itens comuns).

Maiores detalhes podem ser encontrados no capítulo 4 do trabalho de Andrade, Tavares e Valle (2000). Em geral, os casos 3 e 6 são mais utilizados e recomenda-se pelo menos 20% de itens comuns para obter-se um bom resultado na equalização [2], segundo Navas (1996). O caso 6, segundo Andrade, Tavares e Valle (2000) representa o melhor exemplo do uso e da importância da equalização e sem dúvida, ilustra o maior avanço da TRI sobre a Teoria Clássica dos Testes (TCT).

O tamanho da amostra necessário para calibração depende da quantidade de itens do banco, da quantidade de parâmetros do modelo da TRI a ser utilizado e do padrão de respostas da própria amostra, ou seja, é necessário que todas as categorias de respostas tenham uma quantidade de respostas suficientes para a estimação dos parâmetros dos itens.

Segundo Moreira (2011), devem-se eliminar do banco os itens com propriedades psicométricas inadequadas (item pouco discriminativo, com erro padrão alto ou que não se ajusta adequadamente). Por outro lado, a inclusão de novos itens pode ser feita gradualmente, sendo adicionados a um teste juntamente com os demais itens calibrados, onde eles não seriam utilizados para avaliar o respondente, mas apenas para serem calibrados. A calibração dos itens do banco pode ser atualizada quando se dispuser de mais respostas.

1.2.2 Métodos de Estimação dos Parâmetros dos Itens e das Habilidades

O processo de calibração dos itens é muito importante para o bom desempenho do uso da TRI. Existem três métodos para Estimação dos parâmetros na TRI frequentemente usados na literatura: Método da Máxima Verossimilhança, Métodos Bayesianos e Métodos Bayesianos com MCMC (Markov Chain Monte Carlo).

[2] Equalização é um dos conceitos mais importantes da TRI e um dos grandes objetivos das Avaliações Educacionais. Equalizar significa equiparar, tornar comparável, o que no caso da TRI significa colocar parâmetros de itens vindos de provas distintas ou habilidades de respondentes de diferentes grupos, na mesma métrica, isto é, numa escala comum, tornando os itens e/ou as habilidades comparáveis. Existem dois tipos de equalização: via população e a via itens comuns

Em todos esses métodos, que demonstraremos a seguir, algumas notações e suposições serão necessárias para o desenvolvimento do modelo. Em particular, sejam θ_j a habilidade e U_{ji} a variável aleatória que representa a resposta do indivíduo j ao item i. Sejam $\mathbf{U}_{j\cdot} = (U_{j1}, U_{j2}, \cdots, U_{jI})$ o vetor aleatório de respostas binárias (1 para correta e 0 para incorreta) do respondente j e $\mathbf{U}_{\cdot\cdot} = (\mathbf{U}_{1\cdot}, \mathbf{U}_{2\cdot}, \cdots, \mathbf{U}_{N\cdot})$ o conjunto integral de respostas. De forma similar, representaremos as observações por u_{ji}, $\mathbf{u}_{j\cdot}$ e $\mathbf{u}_{\cdot\cdot}$. Ainda $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)$ representará o vetor de habilidades dos N respondentes e $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \cdots, \boldsymbol{\zeta}_I)$ o conjunto dos parâmetros dos itens, onde $\boldsymbol{\zeta}_i = (a_i, b_i, c_i)$.

Na próxima seção detalharemos os Métodos de Estimação mais utilizados na literatura e nos algoritmos atuais.

1.3 Métodos de Estimação

Nos primeiros estudos de TRI, os parâmetros dos itens e das habilidades eram estimados e maximizados simultaneamente (era o Método da Máxima Verossimilhança Conjunta). Entretanto, por envolver uma quantidade muito grande de parâmetros a serem estimados, existem grandes problemas computacionais na utilização desse método. Com o objetivo de resolver esse problema, foi proposto o Método da Máxima Verossimilhança Marginal (MVM) para a estimação dos parâmetros.

Conforme Andrade, Tavares e Valle (2000), o método da MVM pode apresentar problemas de indeterminação e problemas na estimação do parâmetro de acerto casual, obtendo valores fora do intervalo $[0, 1]$, e da discriminação, obtendo valores negativos. Além disso, esse método não está definido para alguns padrões de resposta (itens respondidos corretamente ou incorretamente por todos os respondentes).

Estimação dos Parâmetros dos Itens

Pela independência entre as respostas de diferentes respondentes e a independência local, podemos escrever a verossimilhança como

$$\begin{aligned} L(\boldsymbol{\zeta}) &= P(\mathbf{U}_{\cdot\cdot} = u_{\cdot\cdot}|\boldsymbol{\theta}, \boldsymbol{\zeta}) \\ &= \prod_{j=1}^{n}\prod_{i=1}^{I} P(U_{ji} = u_{ji}|\boldsymbol{\theta_j}, \boldsymbol{\zeta}_i) \\ &= \prod_{j=1}^{n}\prod_{i=1}^{I} P_{ji}^{u_{ji}}[1 - P_{ji}]^{1-u_{ji}}, \end{aligned} \tag{1.7}$$

onde $P_{ji} = P(U_{ji} = 1|\boldsymbol{\theta}_j, \boldsymbol{\zeta}_i)$. Logo, o Logaritmo da Verossimilhança será dado por

$$l(\boldsymbol{\zeta}) = \sum_{j=1}^{n}\sum_{i=1}^{I} u_{ji}\log P_{ji} + (1-u_{ji})\log(1-P_{ji}). \tag{1.8}$$

Os estimadores de Máxima Verossimilhança de $\boldsymbol{\zeta}_i$, $i = 1, \cdots, I$ serão obtidos a partir das equações

$$\frac{\partial l(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_i} = \mathbf{0},\ i = 1, \cdots, I. \tag{1.9}$$

Com essa equação e fazendo $\frac{\partial l(\boldsymbol{\zeta})}{\partial a_i} = 0, \frac{\partial l(\boldsymbol{\zeta})}{\partial b_i} = 0$ e $\frac{\partial l(\boldsymbol{\zeta})}{\partial c_i} = 0$, obtém-se:

$$D(1-c_i)\sum_{j=1}^{n}(u_{ji} - P_{ji})(\theta_j - b_i)W_{ji} = 0, \tag{1.10}$$

$$-Da_i(1-c_i)\sum_{j=1}^{n}(u_{ji} - P_{ji})W_{ji} = 0 \tag{1.11}$$

e

$$\sum_{j=1}^{n}(u_{ji} - P_{ji})\frac{W_{ji}}{P^*_{ji}} = 0, \tag{1.12}$$

onde $W_{ji} = \frac{P^*_{ji}[1-P^*_{ji}]}{P_{ji}[1-P_{ji}]}$ e $P^*_{ji} = \left(1 + e^{-Da_i(\theta_j - b_i)}\right)^{-1}$.

Como essas equações não apresentam soluções explícitas para a_i, b_i e c_i, utiliza-se um método iterativo para obterem-se as estimativas desejadas. Andrade, Tavares e Valle (2000) descrevem o desenvolvimento para a aplicação dos processos iterativos de Newton-Raphson e "Scoring" de Fisher.

Estimação das Habilidades

Para a estimação das habilidades considera-se $l(\theta) = \sum_{j=1}^{n}\sum_{i=1}^{I} u_{ji}\log P_{ji} + (1 - u_{ji})\log(1 - P_{ji})$ e fazendo-se $\frac{\partial l(\theta)}{\partial \theta_j} = 0$, $j = 1, \cdots, n$, obtém-se

$$D\sum_{i=1}^{I} a_i(1-c_1)(u_{ji} - P_{ji})W_{ji} = 0 \tag{1.13}$$

Novamente, esta equação não apresenta solução explícita para θ_j e, por isso, precisamos de algum método iterativo para obter as estimativas desejadas. Andrade, Tavares e Valle (2000) descrevem o desenvolvimento para a aplicação dos processos iterativos de Newton-Raphson e "Scoring" de Fisher.

1.3.1 Método da Máxima Verossimilhança Marginal

O método da MVM propõe fazer a estimação em duas etapas: na primeira, estimam-se os parâmetros dos itens assumindo-se uma certa distribuição para as habilidades (consideremos uma densidade $g(\theta|\boldsymbol{\eta})$ para θ. Ao supor que $\theta \sim N(\mu, \sigma^2)$, temos $\boldsymbol{\eta} = (\mu, \sigma^2)$, por

exemplo). Agora, utiliza-se um artifício relativamente simples para eliminar as habilidades na verossimilhança: basta marginalizar a verossimilhança, integrando-a com respeito à distribuição da habilidade; e em seguida, estimam-se as habilidades assumindo-se os parâmetros dos itens conhecidos (esse ponto já foi resolvido anteriormente).

Para chegarmos às equações da primeira etapa, vamos considerar a seguinte abordagem de Andrade, Tavares e Valle (2000): quando o número de respondentes é grande com relação ao número de itens, existem vantagens computacionais em trabalhar com o número de ocorrências dos diferentes padrões de resposta. Neste sentido, daqui em diante vamos trabalhar considerando este raciocínio. O índice j não mais representará um indivíduo, mas sim um padrão de resposta. Seja r_j o número de ocorrências distintas do padrão de resposta j, e ainda $s \leq min(n, S)$ o número de padrões de resposta com $r_j > 0$. Segue disso que $\sum_{j=1}^{s} r_j = n$. Pela independência entre as respostas dos diferentes indivíduos, os dados seguem uma distribuição Multinomial, isto é,

$$L(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \frac{n!}{\prod_{j=1}^{s} r_j!} \prod_{j=1}^{s} [P(u_j|\zeta, \boldsymbol{\eta})]^{r_j}. \tag{1.14}$$

O logaritmo da verossilhança será

$$l(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \log\left(\frac{n!}{\prod_{j=1}^{s} r_j!}\right) + \sum_{j=1}^{s} r_j \log P(u_j|\boldsymbol{\zeta}, \boldsymbol{\eta}). \tag{1.15}$$

As equações de estimação para os parâmetros dos itens serão obtidas a partir de

$$\frac{\partial l(\boldsymbol{\zeta}, \boldsymbol{\eta})}{\partial \boldsymbol{\zeta}_i} = \mathbf{0},\ i = 1, \cdots, I. \tag{1.16}$$

Com essa equação e fazendo $\frac{\partial l(\boldsymbol{\zeta}, \boldsymbol{\eta})}{\partial a_i} = 0, \frac{\partial l(\boldsymbol{\zeta}, \boldsymbol{\eta})}{\partial b_i} = 0$ e $\frac{\partial l(\boldsymbol{\zeta}, \boldsymbol{\eta})}{\partial c_i} = 0$, obtém-se:

$$D(1 - c_i) \sum_{j=1}^{s} r_j \int_{\mathbb{R}} [(u_{ji} - P_i)(\theta - b_i)W_i] g_j^*(\theta) d\theta = 0, \tag{1.17}$$

$$-Da_i(1 - c_i) \sum_{j=1}^{s} r_j \int_{\mathbb{R}} [(u_{ji} - P_i)W_i] g_j^*(\theta) d\theta = 0 \tag{1.18}$$

e

$$\sum_{j=1}^{s} r_j \int_{\mathbb{R}} \left[(u_{ji} - P_i)\frac{W_i}{P_i^*}\right] g_j^*(\theta) d\theta = 0. \tag{1.19}$$

E para evitar que todos os parâmetros dos itens sejam estimados simultaneamente utiliza-se o algoritmo EM (um processo iterativo para determinação de estimativas de máxima verossimilhança) que permite que os itens possam ter seus parâmetros estimados em separado, facilitando em muito o aspecto computacional do processo de estimação (Andrade, Tavares e Valle (2000), página 64). Para isso, algumas alterações nas expressões anteriores

- equações (1.17), (1.18) e (1.19) - são necessárias. Observêmo-nas

$$\begin{aligned}\frac{\partial l(\boldsymbol{\zeta},\boldsymbol{\eta})}{\partial a_i} &= D(1-c_i)\sum_{j=1}^{s} r_j \int_{\mathbb{R}}[(u_{ji}-P_i)(\theta-b_i)W_i]g_j^*(\theta)d\theta \\ &= D(1-c_i)\sum_{j=1}^{s} r_j \int_{\mathbb{R}}(\theta-b_i)[(u_{ji}g_j^*(\theta)-P_i g_j^*(\theta)W_i]d\theta \\ &= D(1-c_i)\int_{\mathbb{R}}(\theta-b_i)\left[\sum_{j=1}^{s} r_j u_{ji}g_j^*(\theta) - P_i\sum_{j=1}^{s} r_j g_j^*(\theta)\right]W_i d\theta \\ &= D(1-c_i)\int_{\mathbb{R}}(\theta-b_i)\left[r_i(\theta)-P_i f_i(\theta)\right]W_i d\theta, \end{aligned} \tag{1.20}$$

onde $r_i(\theta)=\sum_{j=1}^{s} r_j u_{ji}g_j^*(\theta)$, $f_i(\theta)=\sum_{j=1}^{s} r_j g_j^*$.

Analogamente das equações (1.18) e (1.19), extraem-se:

$$\frac{\partial l(\boldsymbol{\zeta},\boldsymbol{\eta})}{\partial b_i} = -Da_i(1-c_i)\int_{\mathbb{R}}[r_i(\theta)-P_i f_i(\theta)]W_i d\theta \tag{1.21}$$

e

$$\frac{\partial l(\boldsymbol{\zeta},\boldsymbol{\eta})}{\partial c_i} = \int_{\mathbb{R}}[r_i(\theta)-P_i f_i(\theta)]\frac{W_i}{P_i^*}d\theta. \tag{1.22}$$

1.3.2 Métodos Bayesianos

Mais recentemente, os Métodos Bayesianos foram propostos para, entre outras coisas, resolver dois problemas das estimações por Máxima Verossimilhança: (1) estimação dos parâmetros dos itens respondidos corretamente ou incorretamente por todos os respondentes, (2) estimação das proficiências dos respondentes que acertaram ou erraram todos os itens da prova.

Nos métodos de Máxima Verossimilhança também há a possibilidade de que as estimativas dos parâmetros dos itens fiquem fora do intervalo esperado, por exemplo, valores negativos para a discriminação ou valores estimados para o acerto casual fora do intervalo $[0,1]$. A utilização *de prioris* adequadas nos métodos bayesianos é uma solução para esses problemas.

A estimação bayesiana consiste em estabelecer distribuições *a priori* para os parâmetros, construir uma nova função denominada distribuição *a posteriori* e estimar os parâmetros de interesse com base em alguma característica dessa distribuição. Os métodos bayesianos mais utilizados para estimar os parâmetros são o da Média *a posteriori* (EAP), que utiliza a média da distribuição *a posteriori*; e o da Moda *a posteriori* (MAP), que utiliza a moda da distribuição *a posteriori*.

Conforme Andrade, Tavares e Valle (2000), para tornar o tratamento mais geral, considera-se que a distribuição da habilidade é função de um vetor de parâmetros $\boldsymbol{\eta}$, com densidade $g(\theta|\boldsymbol{\eta})$, e que a distribuição de $\boldsymbol{\zeta}_i$, $i=1,\cdots,I$ é a função de um vetor de

parâmetros $\boldsymbol{\tau}$, com densidade $f(\boldsymbol{\zeta}|\boldsymbol{\tau})$. Definem-se, ainda, distribuições *a priori* para os parâmetros $\boldsymbol{\tau}$ e $\boldsymbol{\eta}$: $f(\boldsymbol{\tau})$ e $g(\boldsymbol{\eta})$.

Considerando a função de verossimilhança

$$L(u_{..}|\theta, \boldsymbol{\eta})$$

e a distribuição *a priori*

$$\begin{aligned} f(\theta, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\tau}) &= f(\boldsymbol{\zeta}|\boldsymbol{\tau})g(\theta|\boldsymbol{\eta})f(\boldsymbol{\tau})g(\boldsymbol{\eta}) \\ &= \left[\prod_{i=1}^{I} f(\boldsymbol{\zeta}_i|\boldsymbol{\tau})\right]\left[\prod_{j=1}^{n} g(\theta_j|\boldsymbol{\eta})\right] f(\boldsymbol{\tau})g(\boldsymbol{\eta}), \end{aligned} \tag{1.23}$$

a distribuição *a posteriori* será proporcional a

$$f(\theta, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\tau}|\boldsymbol{u}_{..}) \propto L(u_{..}|\theta, \boldsymbol{\eta})f(\boldsymbol{\zeta}|\boldsymbol{\tau})g(\theta|\boldsymbol{\eta})f(\boldsymbol{\tau})g(\boldsymbol{\eta}). \tag{1.24}$$

Estimação dos Parâmetros dos Itens

Para se fazer inferências com relação aos parâmetros dos itens, marginaliza-se a distribuição *a posteriori*, integrando-a com respeito a θ e $\boldsymbol{\tau}$

$$\begin{aligned} f^*(\boldsymbol{\zeta}, \boldsymbol{\eta}|\boldsymbol{u}_{..}) &\propto \int\int L(\boldsymbol{u}_{..}|\theta, \boldsymbol{\eta})f(\boldsymbol{\zeta}|\boldsymbol{\tau})g(\theta|\boldsymbol{\eta})f(\boldsymbol{\tau})g(\boldsymbol{\eta})d\theta d\boldsymbol{\tau} \\ &\propto g(\boldsymbol{\eta})\left[\int f(\boldsymbol{\zeta}|\boldsymbol{\tau})f(\boldsymbol{\tau})d\boldsymbol{\tau}\right]\left[\int L(\boldsymbol{u}_{..}|\theta, \boldsymbol{\eta})g(\theta|\boldsymbol{\eta})d\theta\right] \\ &\propto g(\boldsymbol{\eta})f(\boldsymbol{\zeta})L(\boldsymbol{u}_{..}|\boldsymbol{\zeta}, \boldsymbol{\eta}) \end{aligned} \tag{1.25}$$

Para o estimador de $\boldsymbol{\zeta}$, podemos escolher alguma característica de $f^*(\boldsymbol{\zeta}, \boldsymbol{\eta}|\boldsymbol{u}_{..})$, por exemplo, a moda ou a média. Segue-se, pois, com o desenvolvimento da moda *a posteriori* - MAP

$$\log f^*(\boldsymbol{\zeta}, \boldsymbol{\eta}|\boldsymbol{u}_{..}) = C + \log g(\boldsymbol{\eta}) + \log f(\boldsymbol{\zeta}) + \log L(\boldsymbol{u}_{..}|\boldsymbol{\zeta}, \boldsymbol{\eta}) \tag{1.26}$$

$$\frac{\partial \log f^*(\boldsymbol{\zeta}, \boldsymbol{\eta}|\boldsymbol{u}_{..})}{\partial \boldsymbol{\zeta}_i} = \frac{\partial \log f(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_i} + \frac{\partial l(\boldsymbol{\zeta}, \boldsymbol{\eta})}{\partial \boldsymbol{\zeta}_i} = 0 \tag{1.27}$$

Comparando esta última equação com a (1.16), observa-se que a abordagem bayesiana adiciona uma nova parcela - a primeira parcela da equação (1.27), $\frac{\partial \log f(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_i}$, relativa à distribuição *a priori* associada aos parâmetros dos itens. Já a segunda parcela da equação (1.27): $\frac{\partial l(\boldsymbol{\zeta}, \boldsymbol{\eta})}{\partial \boldsymbol{\zeta}_i}$ foi desenvolvida pelas equações (1.17), (1.18) e (1.19).

Assumindo independência *a priori* e levando em conta todas as limitações dos parâmetros dos itens, escolhemos as seguintes prioris para o nosso trabalho:

- a_i segue uma distribuição Log-Normal com o parâmetro $\tau = (\mu_a, \sigma_a^2)$ (pois a_i precisa ser positivo):

$$f(a_i|\mu_a, \sigma_a^2) = \frac{1}{\sqrt{2\pi}a_i\sigma_a} e^{\left[-\frac{1}{2\sigma_a^2}(\log a_i - \mu_a)^2\right]}. \tag{1.28}$$

$$\frac{\partial \log f(a_i|\mu_a, \sigma_a^2)}{\partial a_i} = -\frac{1}{a_i}\left[1 + \frac{\log a_i - \mu_a}{\sigma_a^2}\right]. \tag{1.29}$$

- b_i segue uma distribuição Normal com o parâmetro $\tau = (\mu_b, \sigma_b^2)$ (pois b_i tem que ter a mesma escala das habilidades)

$$f(b_i|\mu_b, \sigma_b^2) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{\left[-\frac{1}{2\sigma_b^2}(b_i - \mu_b)^2\right]}. \tag{1.30}$$

$$\frac{\partial \log f(b_i|\mu_b, \sigma_b^2)}{\partial b_i} = -\frac{b_i - \mu_b}{\sigma_b^2}. \tag{1.31}$$

- c_i segue uma distribuição Beta com o parâmetro $\tau = (\alpha - 1, \beta - 1)$ (pois c_i deve estar no intervalo $[0, 1]$)

$$f(c_i|\alpha, \beta) = \frac{\Gamma(\alpha + \beta - 2)}{\Gamma(\alpha - 1)\Gamma(\beta - 1)} c_i^{\alpha-2}(1 - c_i)^{\beta-2}, \tag{1.32}$$

onde $\Gamma(\cdot)$ é a função Gama.

$$\frac{\partial \log f(c_i|\alpha, \beta)}{\partial c_i} = \frac{\alpha - 2}{c_i} - \frac{\beta - 2}{1 - c_i}. \tag{1.33}$$

Com as parcelas obtidas com as equações (1.29), (1.31) e (1.33), completamos as equações de estimação para as componentes de $\boldsymbol{\zeta_i}$, utilizando os resultados de (1.20), (1.21) e (1.22)

$$D(1 - c_i)\int_{\mathbb{R}} (\theta - b_i)\left[r_i(\theta) - P_i f_i(\theta)\right] W_i d\theta - \frac{1}{a_i}\left[1 + \frac{\log a_i - \mu_a}{\sigma_a^2}\right] = 0, \tag{1.34}$$

$$-Da_i(1 - c_i)\int_{\mathbb{R}} [r_i(\theta) - P_i f_i(\theta)] W_i d\theta - \frac{b_i - \mu_b}{\sigma_b^2} = 0 \tag{1.35}$$

e

$$\int_{\mathbb{R}} [r_i(\theta) - P_i f_i(\theta)] \frac{W_i}{P_i^*} d\theta + \frac{\alpha - 2}{c_i} - \frac{\beta - 2}{1 - c_i} = 0. \tag{1.36}$$

Estimação das Habilidades

De maneira análoga ao método de MVM, a estimação bayesiana das habilidades é feita em uma segunda etapa, considerando os parâmetros dos itens fixos.

Vamos supor que a distribuição *a priori* para θ_j é Normal, com vetor de parâmetros $\boldsymbol{\eta} = (\mu, \sigma^2)$. Sabemos, ainda, que a verossimilhança é dada por $L(\boldsymbol{u}_{j\cdot}|\theta_j, \boldsymbol{\zeta})$ e, portanto, a distribuição *a posteriori* para a habilidade do respondente j pode ser escrita como

$$\begin{aligned} g_j^*(\theta_j) = g(\theta_j|\boldsymbol{u}_{j\cdot}, \boldsymbol{\zeta}, \boldsymbol{\eta}) &\propto L(\boldsymbol{u}_{j\cdot}|\theta_j, \boldsymbol{\zeta}) g(\theta_j|\boldsymbol{\eta}) \\ &\propto \prod_{i=1}^{I} P(u_{ji}|\theta_j, \boldsymbol{\zeta}_i) g(\theta_j|\mu, \sigma^2) \\ &\propto \prod_{i=1}^{I} P_{ji}^{u_{ji}} [1 - P_{ji}]^{1-u_{ji}} \frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{1}{2\sigma^2}(\theta_j - \mu)^2\right]}. \end{aligned} \tag{1.37}$$

- Estimação pela moda *a posteriori* - MAP.

  Por facilidade algébrica, trabalharemos com o logaritmo da posteriori de θ_j

$$\begin{aligned} \log g_j^*(\theta_j) &= C + \log L(u_{j\cdot}|\theta_j, \boldsymbol{\zeta}) + \log g(\theta_j|\boldsymbol{\eta}) \\ &= C + \sum_{i=1}^{I} \log P(u_{ji}|\theta_j, \boldsymbol{\zeta}_i) - \log \sigma - \frac{1}{2\sigma^2}(\theta_j - \mu)^2. \end{aligned} \tag{1.38}$$

  Derivando a equação (1.38) com respeito a θ_j e igualando-a a 0, obtemos a equação de estimação para θ_j observando o resultado da equação (1.13)

$$\begin{aligned} \frac{\partial \log g_j^*(\theta_j)}{\partial \theta_j} &= \frac{\partial \log L(u_{j\cdot}|\theta_j, \boldsymbol{\zeta})}{\partial \theta_j} + \frac{\partial \log g(\theta_j|\boldsymbol{\eta})}{\partial \theta_j} \\ &= \sum_{i=1}^{I} \frac{\partial \log P(u_{ji}|\theta_j, \boldsymbol{\zeta}_i)}{\partial \theta_j} - \frac{\theta_j - \mu}{\sigma^2} \\ &= D \sum_{i=1}^{I} a_i (1 - c_1)(u_{ji} - P_{ji}) W_{ji} - \frac{\theta_j - \mu}{\sigma^2} = 0. \end{aligned} \tag{1.39}$$

  Como esse resultado não tem solução explícita, utiliza-se um método iterativo, tal como o método "Scoring" de Fisher.

- Estimação pela média *a posteriori* - EAP.

$$\theta_j^{bayes} = E[\theta_j|u_{j\cdot}, \boldsymbol{\zeta}, \boldsymbol{\eta}] = \frac{\int_{\mathbb{R}} \theta_j L(u_{j\cdot}|\theta_j, \boldsymbol{\zeta}) g(\theta_j|\boldsymbol{\eta}) d\theta_j}{\int_{\mathbb{R}} L(u_{j\cdot}|\theta_j, \boldsymbol{\zeta}) g(\theta_j|\boldsymbol{\eta}) d\theta_j}. \tag{1.40}$$

  Alguns autores, como Andrade, Tavares e Valle (2000) e Mislevy e Stocking (1989), por exemplo, recomendam o método EAP, pois não há necessidade de métodos iterativos para a estimação.

Como as equações de estimação possuem integrais que não apresentam soluções analíticas, algum meio deve ser encontrado para a solução (aproximação) numérica delas. Embora existam muitos métodos de aproximações de integrais, na TRI têm sido frequente, segundo Andrade, Tavares e Valle (2000), a aplicação do método Hermite-Gauss, usualmente denominado método de quadratura. Dessa forma, o problema de obter a integral de uma função contínua é substituído pela obtenção da soma das áreas de um número finito de ratângulos.

Uma outra alternativa utilizada em TRI para efetuar tais aproximações é a utilização de métodos Bayesianos com MCMC, onde realiza-se um conjunto de simulações de amostras aleatórias da distribuição *a posteriori*, baseada na construção de uma cadeia de Markov cuja distribuição estacionária é a distribuição de interesse, conforme o trabalho de Bazan (2005) explicita. A pesquisa de Azevedo (2008) destaca que os métodos MCMC permitem obter, de forma empírica, a estrutura de distribuições a posteriori conjuntas e marginais que são complicadas ou impossíveis de serem obtidas de maneira explícita.

No nosso trabalho, utilizaremos o método de quadratura proposto por Gray (2001), que apesar de ser um método clássico, é considerado por muitos estudiosos o "estado da arte" para se obter estimadores em TRI. Para tanto, basta considerar a seguinte aproximação numérica do estimador EAP de θ_j

$$\begin{aligned}\theta_j^{bayes} &= \frac{\int_{\mathbb{R}} \theta_j L(\theta_j|u_1,\cdots,u_{k-1})g(\theta_j)d\theta_j}{\int_{\mathbb{R}} L(\theta_j|u_1,\cdots,u_{k-1})g(\theta_j)d\theta_j} \\ &\approx \frac{\sum_{t=1}^{q} \overline{\theta_t} L(\overline{\theta_t}|u_1,\cdots,u_{k-1})A_t}{\sum_{t=1}^{q} L(\overline{\theta_t}|u_1,\cdots,u_{k-1})A_t},\end{aligned} \tag{1.41}$$

em que $\overline{\theta_t}$ representa os pontos de quadratura e A_t, o peso associado a $\overline{\theta_t}$. Para mais detalhes, vide Gray (2001).

A variância *a posteriori* associada ao método EAP é dada por

$$\begin{aligned}Var[\theta_j|u_1,\cdots,u_{k-1}] &= \frac{\int_{\mathbb{R}}[\theta_j-\theta_j^{bayes}]^2 L(\theta_j|u_1,\cdots,u_{k-1})g(\theta_j)d\theta_j}{\int_{\mathbb{R}} L(\theta_j|u_1,\cdots,u_{k-1})g(\theta_j)d\theta_j} \\ &\approx \frac{\sum_{t=1}^{q}[\overline{\theta_t}-\theta_j^{bayes}]^2 L(\overline{\theta_t}|u_1,\cdots,u_{k-1})A_t}{\sum_{t=1}^{q} L(\overline{\theta_t}|u_1,\cdots,u_{k-1})A_t}.\end{aligned} \tag{1.42}$$

2 Teste Adapatativo Informatizado - CAT

2.1 Visão Geral de um CAT

Quando se realizam exames avaliativos com muitos respondentes, o examinador deve se responder a seguinte questão: *Como avaliar a habilidade de milhares de candidatos, sem perder a comparabilidade de seus resultados?*

Se a resposta for *Utilizando uma mesma prova*, o examinador estará utilizando o modelo clássico de avaliação e necessitará de um teste grande (com muitos itens), desgastando o candidato, tornando o teste pouco atrativo. Por exemplo, o Exame Nacional do Ensino Médio - ENEM, utiliza dois dias de provas com 180 questões ao todo. Provas de concursos públicos não se afastam muito desse modelo, pois os candidatos se submetem a provas únicas e são muito longas.

Se a resposta for *Utilizando provas diferentes*, o examinador fará uso de um CAT, que mesmo com itens diferentes em diversos testes submetidos a diversos candidatos, pode comparar as diferentes habilidades dos respondentes (e com alta precisão). Nesse caso, os testes são bem menores (mais rápidos) que os testes clássicos e podem ser muito eficientes.

Para a segunda resposta (realização de um CAT), estabelece-se um primeiro problema: *Como montar um teste ideal para um candidato?* Para um candidato com alta habilidade não perder tempo com itens fáceis, seria conveniente que ele responda um teste com itens mais difíceis. Analogamente, um respondente com baixa habilidade precisa ser submetido a um teste com itens mais fáceis. No fundo, um teste eficiente precisa fornecer ao candidato itens com nível de dificuldade condizente com sua habilidade.

Precisamos, portanto, montar uma avaliação adaptativa que não prejudique nenhum respondente e cujo tamanho seja o ideal para estimarmos a habilidade do participante. Temos que ter atenção com o número de itens no teste. Por um lado, forçamos para que o teste seja o menor possível para que ele seja atrativo, por outro, um número insuficiente de itens em cada um dos níveis coloca a avaliação em risco. Nesse sentido, a prova precisa ser personalizada para cada participante e ela precisa ser comparável com todas as outras provas dos demais respondentes.

Para avançarmos com a construção de um CAT, vale a pena estabelecermos a seguinte reflexão:

Se um aluno do terceiro ano acertou 8 questões de uma prova de 10 questões e um outro, do segundo ano, acertou 6 das 10 questões de uma outra prova. Podemos afirmar que o primeiro apresenta uma habilidade maior do que o segundo?

Não. São provas diferentes e para compará-las, não podemos nos basear apenas no número de acertos. Não é uma medida apropriada. Afinal estamos estudando duas populações distintas (terceiro ano e segundo ano) que foram submetidas a duas avaliações distintas e a comparação entre as habilidades dos alunos dessas duas populações não é recomendada com a metodologia clássica. Mas se utilizarmos a metodologia estudada no capítulo anterior, a Teoria da Resposta ao Item (TRI), em que todos os itens já estariam calibrados e o banco de itens devidamente equalizado, os itens poderiam ser colocados numa mesma régua, numa mesma escala (por exemplo, em ordem crescente de dificuldade - b_i) e assim, a informação do teste será maior, pois perceberemos se o candidato está acertando itens mais difíceis (alto valor de b_i) ou se ele está acertando itens mais fáceis (baixo valor de b_i). Desse modo conseguiríamos classificar e comparar esses dois participantes.

Nessa perspectiva, temos que ter um banco de itens rico, robusto, com muitos itens e com um alto poder de discriminação ($a_{i's}$ superiores a $0,8$, por exemplo). Ou seja, o banco de itens precisa ter qualidade e para isso é necessário fazer pré-testes, descartando itens com baixa qualidade. Por isso os itens precisam ser calibrados.

Percebe-se, portanto, que o objetivo de um CAT é apresentar itens ao indivíduo que sejam adequados ao seu nível de habilidade. A consequência disso é uma estimação mais precisa da proficiência com menos itens aplicados e em menos tempo do que nos testes convencionais do tipo "papel e lápis" onde todos os indivíduos devem responder todas as questões de um mesmo teste.

Observemos a Figura 5, que apresenta um exemplo típico de um CAT para um teste com itens dicotômicos do tipo acerta/erra. Para isso, precisamos estruturar um algoritmo para construir um CAT.

2.2 Construção de um CAT

A prova não é definida *a priori*. Ela é construída à medida que o indivíduo vai respondendo às questões. Precisamos apresentar a prova mais apropriada para cada respondente (a prova é adaptada a cada indivíduo). Para implementarmos um CAT necessitamos:

- **Banco de itens calibrados (na mesma régua)**
  Utiliza-se a TRI, fazendo pré-testes para que o banco seja rico em itens com qualidade e que seja suficientemente grande para contemplar itens com diversas proficiências. Não é uma amostra aleatória e sim intencional.

- **Seleção do primeiro item ou dos primeiros itens do CAT**
  Por exemplo, iniciaremos os testes com item de dificuldade mediana (ou alguns itens,

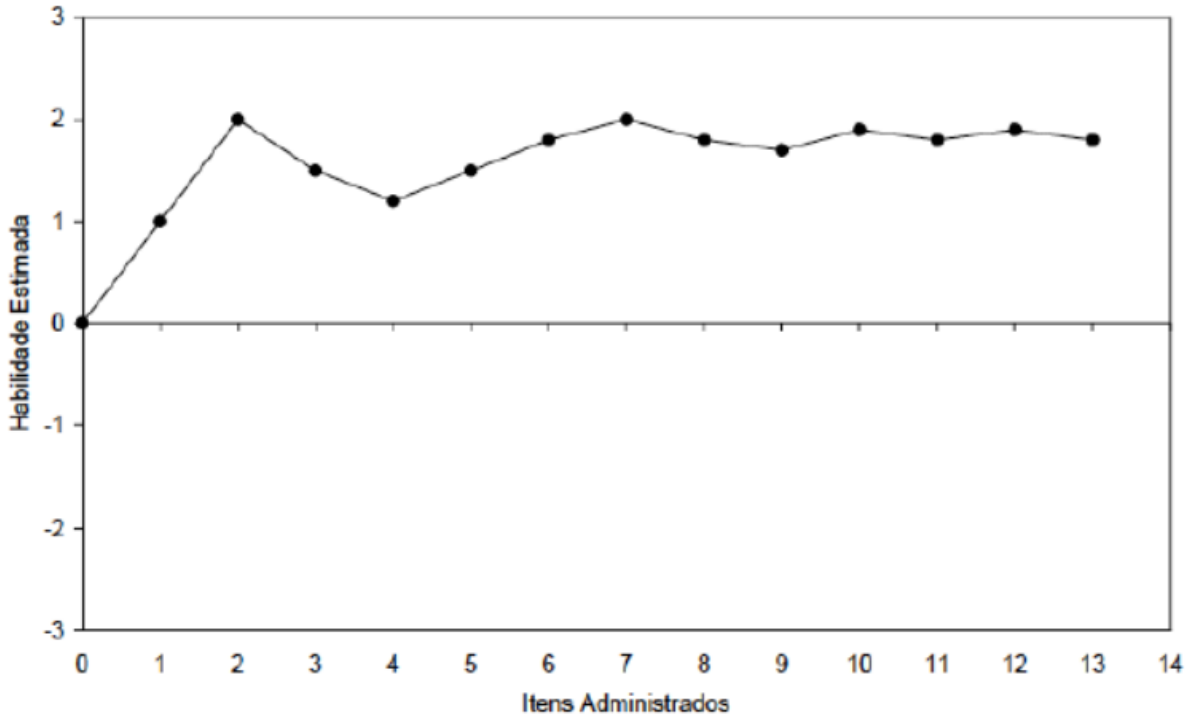


Figura 5: Exemplo de um CAT em que o examinando inicia o teste com uma habilidade mediana, considerando a escala (0, 1). O primeiro item é administrado, o examinando acerta e sua habilidade estimada aumenta. O segundo item é administrado, o examinando acerta e sua habilidade estimada aumenta. O terceiro é administrado, o examinando erra e sua habilidade estimada diminui. O teste continua seguindo essa lógica até que seja encontrado um ponto de equilíbrio, onde o examinando domina o conhecimento que está abaixo desse ponto, mas não domina o conhecimento que está acima. É nesse ponto de equilíbrio que a sua habilidade deverá estar situada.

por exemplo 5, em torno da dificuldade mediana). Nos testes com ponto de corte, podem-se selecionar os primeiros itens com dificuldade próxima ao ponto de corte.

- **Algoritmo de seleção dos próximos itens**

  Um dos componentes mais importantes do CAT consiste nos procedimentos de seleção dos itens ao longo do teste. De acordo com Lord (1980), um examinando é avaliado mais eficientemente quando os itens dos testes não são muito difíceis nem muito fáceis para este candidato. Contudo, os métodos de seleção adaptativa não só avaliam o nível de dificuldade dos itens, mas procuram encontrar uma Medida de Informação (que é uma combinação dos parâmetros dos itens e da estimativa da habilidade) em busca de uma melhor escolha dos itens para a estimação das proficiências. Existem três critérios muito utilizados na literatura e nos algoritmos de seleção dos próximos itens e que serão apresentados na seção 2.3.

- **Método de Estimação da Habilidade**

Toda vez que um item é selecionado e aplicado num teste, a habilidade do examinando é reestimada juntamente com o seu erro padrão. Os principais métodos utilizados na estimação da habilidade foram mencionados na seção 1.3. Entretanto, existem diversas adaptações, alterações ou combinações desses métodos no contexto de um CAT, além da criação de novos métodos.

Por exemplo, Abad et al. (2004) utilizaram a seguinte estratégia para estimar a habilidade: se ocorre um padrão inicial de resposta constante (até o quinto item), utiliza-se a média entre a última habilidade estimada e 2 (se acerta) ou -2 (se erra). Após o quinto item aplica-se o procedimento de Herrando (1989) se o padrão se mantém constante, caso contrário, utiliza-se o método da máxima verossimilhança. É comum utilizar um método no início do teste, quando o erro padrão da estimativa da habilidade ainda é grande e pode ocorrer um padrão de resposta constante, e outro método durante o teste, quando o erro padrão é menor.

No contexto de um CAT, a literatura afirma que o Método MV (Máxima Verossimilhança) apresenta, em relação aos Métodos Bayesianos, maior erro padrão (especialmente para valores extremos da habilidade, tanto para cima, como para baixo), menor viés, menor fidelidade (correlações entre valores estimados e parâmetros), menor eficiência (precisa de mais itens para alcançar a mesma precisão), e maior tempo para os cálculos computacionais. Há autores que consideram mais adequado utilizar o método MV, pelo fato de a estimativa da habilidade não ser afetada por qualquer outra coisa que não seja o desempenho no teste atual. Mas essa é uma opinião minoritária.

Segundo Segall (2005), em um CAT, as estimativas bayesianas tendem a ter a vantagem de erros-padrão condicionais menores, mas possuem a desvantagem de ter viés da estimativa da habilidade condicional maior, especialmente para os níveis extremos de θ. Assim, a escolha do método de estimação deve levar em conta tanto a variância pequena (das estimativas bayesianas) quanto o viés pequeno (das estimativas por MV). Os procedimentos Bayesianos oferecem um menor erro quadrático médio (que é uma função de ambos variância e viés condicionais) do que o Método MV. Isto sugere que as estimativas Bayesianas podem fornecer uma classificação mais precisa da ordenação dos examinandos ao longo da escala do traço latente. Estudiosos que estão preocupados com os efeitos do viés ou que não têm informações sobre a distribuição da habilidade tendem a utilizar a abordagem MV. Por outro lado, estudiosos cujo principal objetivo é minimizar o erro-padrão médio ou a variância condicional tendem a utilizar abordagens Bayesianas.

- **Critério de Parada do Teste**

Uma importante característica de Testes Adaptativos Informatizados é que o critério que finaliza o teste pode depender dos objetivos do teste. Alguns testes são usados para seleção ou classificação, por exemplo, para classificar o indivíduo em uma escala do conhecimento ou para selecionar quais estudantes serão admitidos na universidade ou em um processo seletivo para um trabalho. Outros testes são usados para pesquisas médicas, por exemplo. Para o nosso trabalho, vamos considerar o objetivo de classificação.

Para esse fim, a habilidade de um examinando é comparada com algum valor de corte. A literatura indica que, para implementação no CAT, tanto a estimativa da habilidade como o erro-padrão da medida associado devem ser usados. No caso da estimação das habilidades pelo método EAP, PSD é o erro-padrão associado à medida. Um indivíduo pode ser classificado como sendo acima do valor de corte (expresso na escala do traço latente, θ) se a estimativa da habilidade e seu intervalo de 95% de confiança (calculada como sendo mais ou menos duas vezes o erro-padrão da medida) estão acima ou abaixo do escore de corte. Após a decisão sobre o ponto de corte, o teste pode ser finalizado quando esta condição for satisfeita. O resultado de cada teste será um conjunto de classificações feito por um grupo de examinados que tem pelo menos uma taxa de 5% de erro. A taxa de erro pode ser controlada pela mudança do tamanho do intervalo de confiança do erro-padrão da medida em torno da estimativa da habilidade.

Alguns algoritmos em CAT são finalizados pelo administrador quando atingirem um número fixo de itens ou por imposição de um tempo limite. Ambos os casos são usados por conveniência do administrador do teste o que não é considerada uma boa prática. No nosso caso (em que o CAT é utilizado para classificação), a qualidade do teste pode prejudicar a estimativa de alguns examinandos. Para obter o máximo de benefícios de um CAT, nem o tempo limite nem o tamanho do teste deveriam ser impostos como critérios de parada.

- **Controle na Exposição do Item**

Muitos programas operacionais de testes adaptativos encontram necessariamente uma base para seleção de itens não somente nos procedimentos estatísticos mas também impondo restrições ao procedimento de seleção de itens. Essas restrições visam controlar certos atributos como balanceamento do conteúdo ou frequência de exposição do item.

A imposição de restrições torna-se necessária para melhor aproveitamento das estruturas presentes nos bancos de itens. De fato, a idéia principal na implementação

de algoritmos é poder realizar um Teste Adaptativo Informatizado com as mesmas especificações (e a mesma validade) de um teste comum de "papel e lápis" e ainda fornecer um menor número de itens. O número de restrições no procedimento de seleção de itens para se alcançar esse ideal pode chegar a centenas facilmente. Cabe, portanto, a análise cuidadosa dos objetivos a serem atingidos ao se implementar um CAT.

A restrição em relação à frequência de exposição do item é muito importante em CAT, pois ao se usar o critério de Máxima Informação, por exemplo, os itens de maior parâmetro a tendem a ser administrados diversas vezes no CAT, o que pode levar muitos examinandos a memorizá-los, adicionando assim um erro na estimativa da habilidade e, consequentemente, prejudicando a validade do teste.

Georgiadou et al. (2007) cita diversas estratégias para controle da exposição de itens com pesquisas realizadas entre 1983 e 2005. Destacaremos uma delas, o Procedimento Probabilístico, em que, a exposição de itens pode ser controlada sobre a abordagem da seleção condicional dos itens. O procedimento condicional para seleção de itens foi originalmente proposto por Hetter e Sympson em 1997 e ainda continua sendo um dos métodos mais utilizados na prática. O procedimento Sympson-Hetter (SH) calcula parâmetros de exposição do item para controlar probabilisticamente a frequência com a qual o item é selecionado.

Para reduzir a quantidade de itens superexpostos e satisfazer aos requisitos de segurança operacionais de um CAT, Hetter e Sympson (1997) desenvolveram um algoritmo que pode ser visto no trabalho de Costa (2009).

- **Balanceamento do Conteúdo**

A restrição sobre o balanceamento de conteúdo permite a divisão do banco de itens em várias seções, sendo que cada uma delas representará um conteúdo (também conhecido, na Pedagogia, como habilidade, competência, descritor) que se deseja avaliar no CAT. Dessa forma, o teste adaptativo conterá uma boa variedade de itens de diferentes competências da mesma forma que no teste "papel e lápis".

Em muitas situações, o delineamento em CAT tenta levar em consideração algumas restrições adicionais para a seleção de itens, tal como o balanceamento pelo conteúdo. Imaginemos o seguinte exemplo: um estudo piloto em CAT foi realizado para análise das habilidades dos estudantes do Ensino Fundamental em Matemática. Dessa maneira, foram considerados quatro descritores para avaliar essa área do conhecimento (essa etapa de ser feita em conjunto com um profissional da área de Pedagogia). Para assegurar que cada teste adaptativo mensure todos os quatro descritores, alguns mecanismos são necessários.

Um método proposto por Kingsbury e Zara (1989) leva em consideração o balanceamento do conteúdo. Este algoritmo é uma modificação do procedimento de seleção do item pela Máxima Informação levando também em conta a categoria do conteúdo de cada item no processo de seleção. Uma vez que o item é selecionado pela Máxima Informação para o corrente examinando, se o item selecionado representa um descritor da área do conhecimento que ainda não foi representado no teste, o item é administrado. Caso contrário, o item que oferece a próxima maior informação é avaliado em relação aos descritores estabelecidos e o processo é repetido até que os itens de uma matriz de descritores estabelecidos sejam identificados.

2.3 Critérios para o Algoritmo de Seleção dos Próximos Itens

2.3.1 Critério de Máxima Informação (MI)

Lord (1980) propôs o critério de Máxima Informação (MI) para o CAT que se tornou um dos mais utilizados procedimentos para seleção dos itens. Basicamente, esse método consiste em selecionar o próximo item no CAT com base na medida de Informação de Fisher avaliada na proficiência corrente. Conforme os cálculos apresentados na seção 1.1, equação (1.6).

Segundo Costa (2009), a Informação de Fisher é naturalmente relacionada à estimação da habilidade pela MV e é inversamente proporcional ao erro-padrão do estimador MV. Maximizar a $I_{U_i}(\theta)$ significa intuitivamente selecionar um item de dificuldade que corresponda exatamente ao nível de habilidade do examinando. Em relação ao CAT, a $I_{U_i}(\theta)$ serve como referência para seleção de itens quando existe conhecimento suficiente sobre a localização da habilidade. Nas aplicações atuais, esse critério tem sido o mais utilizado porque, entre outras vantagens, permite estabelecer previamente tabelas calculadas de informações, chamadas *infotable*.

Itens com maior discriminação serão preferencialmente selecionados pelo algoritmo, o que pode causar dois tipos de problemas no início do CAT, quando a quantidade de itens do teste ainda é muito pequena para se avaliar com precisão o valor verdadeiro da habilidade: Primeiro, a aplicação do método da Informação de Fisher pode ser pouco eficiente se a estimativa da habilidade não estiver próxima do valor verdadeiro. Por exemplo, a Figura 6 mostra o que Linden (1998) e Linden e Glas (2010) chamam de paradoxo, onde dois itens estão posicionados no valor atual estimado da habilidade. O critério de MI selecionaria o item mais informativo para a habilidade atual estimada, $\hat{\theta}$, que seria o Item 1, entretanto esse item praticamente não fornece informação onde o verdadeiro valor da habilidade, θ^*, está. No início do CAT, critérios de seleção de itens que não se baseiam na estimativa provisória de θ podem ser mais eficientes do que os critérios de MI. À medida que o teste avança, a estimação da habilidade se torna mais precisa, de

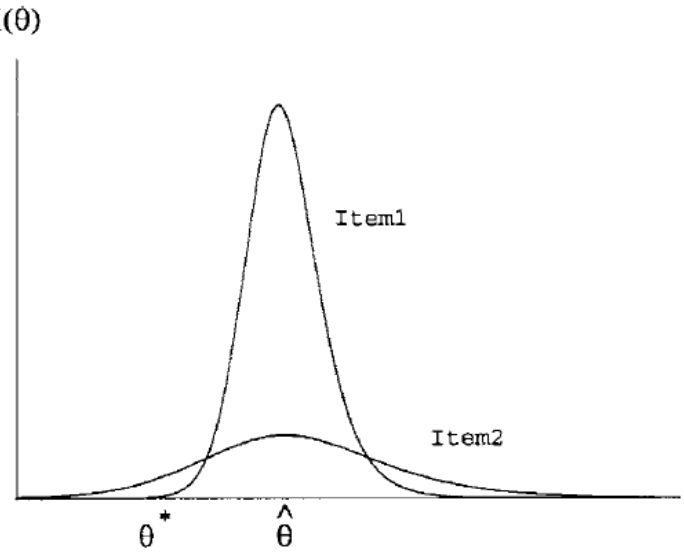


Figura 6: Paradoxo na seleção de itens de um CAT (Fonte: Linden e Glas (2010)

modo que os critérios de seleção que consideram a estimativa provisória de θ serão mais eficientes; Segundo, esses itens deveriam ser utilizados no final do teste, para estimar a habilidade de indivíduos que realmente estejam nesse nível de habilidade.

O critério de MI seleciona como melhor item aquele que produz a menor variância das estimativas. A eficácia dessa estratégia nos CAT's tem sido comprovada através de estudos de simulação, onde se verificou que é possível obter uma boa estimação da habilidade com um número reduzido de itens, em média, 20 itens Olea et al. (1999).

A utilização "pura" desse critério selecionará sempre os mesmos itens para indivíduos que apresentarem as mesmas respostas. Isso causará um problema de superexposição dos itens, principalmente os primeiros, que poderão tornar-se conhecidos. Para eliminar esse problema, outros métodos que podem ser combinados com esse critério foram mencionados na seção 2.2.

2.3.2 Critério de Máxima Informação Global (MIG)

Chang e Ying (1996) sugerem substituir a medida de Informação de Fisher pela Informação de Kullback-Leibler (KL). A motivação para o uso de KL é que a aplicação da Informação de Fisher pode ser pouco eficiente se a estimativa da proficiência não estiver próxima ao valor verdadeiro, especialmente na fase inicial do CAT quando a quantidade de itens do teste ainda é muito pequena para se avaliar com acurácia o valor verdadeiro da proficiência. O maior objetivo do CAT consiste em estimar eficientemente θ com poucos itens. A redução da quantidade de itens no teste adaptativo faz com que a escolha de itens de qualidade na fase inicial do teste seja crucial. Segundo esses autores, a medida de Kullback-Leibler fornece uma Informação Global, ideal para seleção de itens quando a amostra das respostas do examinando ainda é pequena. A medida de informação de KL

com base na função de verossimilhança dada na equação (1.1) pode ser expressa por:

$$K_i(\theta||\theta_0) = E_{\theta_0} \log \left[\frac{[P_i(\theta_0)]^{u_i}[1 - P_i(\theta_0)]^{1-u_i}}{[P_i(\theta)]^{u_i}[1 - P_i(\theta)]^{1-u_i}} \right]$$
$$= P_i(\theta_0) \log \left[\frac{P_i(\theta_0)}{P_i(\theta)} \right] + [1 - P_i(\theta_0)] \log \left[\frac{1 - P_i(\theta_0)}{1 - P_i(\theta)} \right], \quad (2.1)$$

onde θ_0 é o valor verdadeiro da habilidade. K é uma superfície de informação e representa o poder discriminatório de um item nos dois níveis θ e θ_0, resumindo a informação contida no item com respeito a uma amplo intervalo de θ. Se θ_0 varia ao longo da escala, K se torna uma superfície de informação global num espaço tridimensional.

2.3.3 Critério de Máxima Informação Esperada (MIE)

O MIE é um dos procedimentos Bayesianos mais empregados em CAT para seleção de itens. De fato, testes adaptativos parecem ser naturalmente ajustados por uma abordagem Bayesiana empírica ou sequencial. Por exemplo: a distribuição *a posteriori* de θ estimada após $k-1$ itens pode ser prontamente usada para selecionar o k-ésimo item e ser utilizada como distribuição *a priori* para a obtenção da próxima distribuição *a posteriori*. Todos os critérios Bayesianos para seleção de itens no CAT envolvem alguma forma de ponderação baseada na distribuição *a posteriori* de θ. Como a distribuição *a posteriori* é uma combinação da função de Verossimilhança e uma distribuição *a priori*, a diferença básica entre os critérios já mencionados é que esta faz uso de uma distribuição *a priori*. O método da Máxima Informação Esperada baseia-se na análise preditiva. A análise preditiva em Estatística consiste em se fazer inferências probabilísticas sobre uma quantidade a ser observada no futuro Migon e Gamerman (2009). Em CAT, deseja-se prever a resposta aos itens ainda não administrados no teste, depois de $k-1$ respostas e, então, escolher o próximo item de acordo com as atualizações de uma quantidade *a posteriori* para essas respostas. O elemento chave dessa análise está na distribuição *a posteriori* preditiva para a resposta ao item s, com função de probabilidade dada por

$$P_s(u_s|u_1, \cdots, u_{k-1}) = \int P_s(u_s|\theta) g(\theta|u_1, \cdots, u_{k-1}) d\theta, \quad (2.2)$$

onde, $P_s(u_s|\theta)$ é a probabilidade preditiva da resposta u_s ao item s dado θ e $g(\theta|u_1, \cdots, u_{k-1})$ é a densidade *a posteriori* após $k-1$ itens.

Suponha que o item k será selecionado. O examinando responderá a esse item com probabilidade $P_k(1|u_1, \cdots, u_{k-1})$. Uma correta resposta irá atualizar as seguintes quantidades: a distribuição completa *a posteriori* de θ; a estimativa pontual do valor da habilidade do respondente $\hat{\theta}$; e a variância *a posteriori* de θ. Uma resposta incorreta tem probabilidade $P_k(0|u_1, \cdots, u_{k-1})$ e irá atualizar as mesmas quantidades.

A motivação para a adoção do critério MIE vem de Linden (1998). Como destaca o autor, se o k-ésimo item é selecionado, respostas para os $k-1$ itens já são conhecidas. Logo, os dados não podem ser considerados como variáveis aleatórias mas somente como valores fixos da realização dessa variável aleatória. Como consequência, a Informação de Fisher, definida como o valor esperado da variável aleatória U não é uma medida válida. Uma escolha Baysiana típica neste caso é o uso da medida de informação observada, expressa por

$$J_{u_i}(\theta) = -\frac{\partial^2}{\partial\theta^2} l(\theta; u_i).$$

que reflete a curvatura da função de Verossimilhança observada para o θ O objetivo do critério MIE consiste em maximizar a Informação Observada sobre as respostas preditas ao k-ésimo item. Formalmente, a escolha do próximo item que será administrado no CAT pelo critério MIE levará em conta a medida de Informação Observada dos itens no ponto $\hat{\theta}$. Dessa forma, seja i o i-ésimo item do banco, $i = 1, ..., I$, e k, a posição do i-ésimo item no teste adaptativo. Suponha que $k-1$ itens foram administrados no CAT. Os índices dos itens administrados formam o conjunto $S_{k-1} = \{1, 2, \cdots, k-1\}$, enquanto os itens restantes formam o conjunto $R_k = \{1, \cdots, I\} \backslash S_{k-1}$. A seleção do k-ésimo obedecerá à seguinte regra:

$$\begin{aligned} i_k = argmax_s \{ & P_s(0|u_1, \cdots, u_{k-1}) J_{u1,\cdots,uk-1}, U_s = 0(\hat{\theta}_{u1,\ldots,uk-1}, U_s = 0) \\ & + P_s(1|u_1, \cdots, u_{k-1}) J_{u1,\cdots,uk-1}, U_s = 1(\hat{\theta}_{u1,\ldots,uk-1}, U_s = 1) : s \in R_k \}. \end{aligned} \tag{2.3}$$

Parte II

Nova Modelagem e Aplicação com Dados Simulados

3 Modelo com a Covariável Tempo de Resposta

Após analisar os atuais métodos de construção de um CAT, especialmente os critérios de seleção do próximo item, percebemos que uma covariável não estava sendo levada em consideração: o Tempo de Resposta no item.

Isto é, nos atuais critérios (observar seção 2.3), após o candidato responder ao k-ésimo item, com base exclusivamente na sua resposta, escolhe-se o próximo item.

Não encontramos, até agora, nenhum trabalho que tenha levado em consideração a influência do tempo de resposta em um item, na habilidade do respondente e, consequentemente, na seleção da próxima questão de um CAT. Essa foi uma das grandes motivações do presente trabalho, afinal acredita-se que o tempo com que um indivíduo responde um item está fortemente ligado à sua habilidade e, por isso, essa covariável precisa, de alguma forma, ser considerada na modelagem.

Por exemplo, se dois candidatos C1 e C2 resolvem uma mesma questão k, ambos acertam e C1 for mais rápido que C2, então, agregando-se essa informação do tempo de resposta ($t_{C1} < t_{C2}$), reestimamos as habilidades dos candidatos (provavelmente, $\theta_{C1} > \theta_{C2}$) e definimos a questão $k+1$ mais apropriada para C1 e a mais apropriada para C2. Espera-se que a próxima questão de C1 possua o parâmetro de dificuldade (b_j) maior que a de C2.

Esse será o ponto chave do presente estudo, agregando-se essa covariável em um novo modelo para estimar a habilidade do candidato. Acredita-se que o tamanho do teste (consequentemente o tempo total do teste) será diminuído. Como essa é uma pesquisa nova, serão necessárias algumas simulações através de algoritmos construídos de maneira específica para se validar essas suposições. O Capítulo 4 tratará da simulação dos dados e o 5 da estrutura dos algoritmos utilizados. No anexo deste trabalho, colocou-se, na íntegra, os correspondentes algoritmos.

3.1 Modelo Proposto

Inicialmente, padronizou-se a notação. Imaginou-se que o j-ésimo respondente leva, para responder o i-ésimo item, o tempo t_{ij} e a sua resposta seja u_{ij}. Se o Tempo de Resposta no item não for levado em consideração, a modelagem é aquela apresentada na Introdução e Seção 1.1 deste trabalho, em que a saída é (u_{ij}) e $P(u_{ij}|\theta_j)$ segue o modelo

ML3. Com a covariável Tempo de Resposta, a saída é do tipo (u_{ij}, t_{ij}) e $P(u_{ij}, t_{ij}|\theta_j)$ precisa ser modelada. Pode-se escrever

$$P(u_{ij}, t_{ij}|\theta_j) = P(t_{ij}|\theta_j, u_{ij})P(u_{ij}|\theta_j). \tag{3.1}$$

Conforme apresentado na Introdução deste trabalho, usou-se o ML3 para $P(u_{ij}|\theta_j)$ e para simplificação de notação ela será denotada por $P_i(\theta)$.

Precisa-se agora estudar $P(t_{ij}|\theta_j, u_{ij})$. Assume-se que não existe informação no Tempo de Resposta do item quando ele é respondido de forma errada pelo candidato. Em outras palavras, $P(t_{ij}|\theta_j, u_{ij} = 0)$ não depende de θ_j. Por outro lado, tem-se informação no Tempo de Resposta quando o candidato acerta o item, isto é, $P(t_{ij}|\theta_j, u_{ij} = 1)$ depende de θ_j. Mais especificamente, imaginamos que, quanto maior θ_j, menor será t_{ij} e, portanto, precisa-se escolher um modelo razoável para essa relação. Por simplicidade, escolher-se-á a distribuição exponencial para tal modelagem, isto é

$$t_{ij}|\theta_j, u_{ij} = 1 \sim Exp(\lambda_{ij}), \tag{3.2}$$

com $\log(\lambda_{ij}) = r_i + s_i(\theta_j - b_i)$. Uma simplicação adicional pode ocorrer se fizermos $r_i = r$ e $s_i = s$. Nesse caso

$$t_{ij}|\theta_j, u_{ij} = 1 \sim Exp(\lambda_{ij} = e^{r+s(\theta_j - b_i)}) \tag{3.3}$$

e

$$P(t_{ij}|\theta_j, u_{ij} = 1) = \lambda_{ij} e^{-\lambda_{ij} t_{ij}}, \tag{3.4}$$

com

$$E(t_{ij}|\theta_j, u_{ij} = 1) = \frac{1}{\lambda_{ij}} = \frac{1}{e^{r+s(\theta_j - b_i)}}. \tag{3.5}$$

Assim, se $u_{ij} = 0$,

$$P(u_{ij} = 0, t_{ij}|\theta_j) = 1 - P_i(\theta)$$

e se $u_{ij} = 1$,

$$P(u_{ij} = 1, t_{ij}|\theta_j) = P_i(\theta)\lambda_{ij} e^{-\lambda_{ij} t_{ij}}.$$

3.1.1 Função de Verossimilhança do Novo Modelo

A Função de Verossimilhança dessa nova modelagem será expressa por

$$\begin{aligned} L(\theta|u_{ij}, t_{ij}) &= [P_i(\theta)\lambda_{ij} e^{-\lambda_{ij} t_{ij}}]^{u_i} [1 - P_i(\theta)]^{1-u_i} \\ &= [\lambda_{ij} e^{-\lambda_{ij} t_{ij}}]^{u_i} [1 - P_i(\theta)]^{1-u_i} [P_i(\theta)]^{u_i}. \end{aligned} \tag{3.6}$$

O Logaritmo da Verossimilhança será dado por

$$\begin{aligned} l(\theta|u_{ij}, t_{ij}) &= u_i[\log(\lambda_{ij}) - \lambda_{ij} t_{ij}] + (1 - u_i)\log(1 - P_i(\theta)) + u_i \log(P_i(\theta)) \\ &= u_i \log(P_i(\theta)) + (1 - u_i)\log(1 - P_i(\theta)) + u_i[r + s(\theta_j - b_i) - t_{ij} e^{r+s(\theta_j - b_i)}]. \end{aligned} \tag{3.7}$$

3.1.2 Informação de Fisher do novo modelo

A medida de informação observada $J_{u_{ij},t_{ij}}(\theta_j)$ é dada por

$$\begin{aligned} J_{u_{ij},t_{ij}}(\theta_j) &= -\frac{\partial^2}{\partial\theta_j^2} l(\theta - j|u_{ij}, t_{ij}) \\ &= -\frac{\partial^2}{\partial\theta_j^2}[u_i \log(P_i(\theta)) + (1-u_i)\log(1-P_i(\theta))] + u_i s^2 t_{ij} e^{r+s(\theta_j - b_i)}. \end{aligned} \tag{3.8}$$

3.2 Cálculos para os critérios de parada do CAT no novo modelo

Conforme apresentado na seção 2.3, em um CAT, precisa-se definir o critério de seleção dos próximos itens e contemplou-se 3 métodos: Máxima Informação (Informação de Fisher), Máxima Informação Global (Kullback Leibler) e Máxima Informação Esperada (Método Bayesiano). Nos algoritmos desenvolvidos nesse trabalho, utilizou-se apenas o primeiro método, mas a seguir apresenta-se o desenvolvimento teórico de todos esses três critérios para a nova modelagem, a fim de facilitar o estudo em futuros trabalhos.

3.2.1 Máxima Informação

Como visto na seção 2.3.1, esse método consiste em selecionar o próximo item no CAT com base na medida de Informação de Fisher avaliada na habilidade corrente. Apesar de já se ter apresentado definições sobre a medida de Informação, nesta seção dar-se-á maiores detalhes considerando a função de verossimilhança da nova modelagem (Equação 1.41). A medida de informação esperada ou informação de Fisher do i-ésimo item será dada por

$$\begin{aligned} I_{U_{ij},T_{ij}}(\theta_j) &= E_{U_{ij},T_{ij}|\theta_j}\left[-\frac{\partial^2}{\partial\theta_j^2} l(\theta_j; U_{ij}, T_{ij})\right] \\ &= \frac{[P_i'(\theta)]^2}{P_i(\theta)[1-P_i(\theta)]} + E_{U_{ij}} E_{T_{ij}}\left[u_{ij} s^2 t_{ij} e^{r+s(\theta_j - b_i)} | u_{ij} = 1\right] \\ &= \frac{[P_i'(\theta)]^2}{P_i(\theta)[1-P_i(\theta)]} + E_{U_{ij}}\left[u_{ij} s^2 E(t_{ij}|\theta_j, u_{ij}=1) e^{r+s(\theta_j - b_i)} | u_{ij} = 1\right] \\ &= \frac{[P_i'(\theta)]^2}{P_i(\theta)[1-P_i(\theta)]} + E_{U_{ij}}\left[u_{ij} s^2\right] \\ &= \frac{[P_i'(\theta)]^2}{P_i(\theta)[1-P_i(\theta)]} + P_i(\theta) s^2. \end{aligned} \tag{3.9}$$

A primeira parcela dessa equação é a medida de informação que se tinha obtido na equação 1.4, enquanto que a segunda parcela surgiu devido à covariável t_{ij}. É como se a Medida de Informação sofresse uma atualização quando se utiliza tal covariável.

3.2.2 Máxima Informação Global

Como visto na seção 2.3.2, esse critério utiliza a medida de informação de Kullback-Leibler. Utilizando a função de verossimilhança dada na equação 3.6 e denotando θ_0 como o valor verdadeiro da habilidade, para qualquer valor de θ, a informação de Kullback-Leibler para o i-ésimo item (com resposta u_i) é

$$\begin{aligned}
K_i(\theta||\theta_0) &= E_{\theta_0} \log\left[\frac{L_i(\theta_0;u_i)}{L_i(\theta;u_i)}\right] \\
&= E_{\theta_0} \log\left[\frac{[P_i(\theta_0)]^{u_i}[1-P_i(\theta_0)]^{1-u_i}[\lambda_{ij}(\theta_0)e^{-\lambda_{ij}(\theta_0)t_{ij}}]}{[P_i(\theta)]^{u_i}[1-P_i(\theta)]^{1-u_i}[\lambda_{ij}(\theta)e^{-\lambda_{ij}(\theta)t_{ij}}]}\right] \\
&= E_{\theta_0}\left[u_i \log\frac{P_i(\theta_0)}{P_i(\theta)} + (1-u_i)\log\frac{1-P_i(\theta_0)}{1-P_i(\theta)} + u_i\left[s(\theta_0-\theta) - t_i\left(\lambda_i(\theta_0)-\lambda_i(\theta)\right)\right]\right] \\
&= P_i(\theta_0)\log\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right] + [1-P_i(\theta_0)]\log\left[\frac{1-P_i(\theta_0)}{1-P_i(\theta)}\right] + \\
&+P_i(\theta_0)\left[s(\theta_0-\theta) - \frac{1}{e^{r+s(\theta_0-b_i)}}\left(e^{r+s(\theta_0-b_i)} - e^{r+s(\theta-b_i)}\right)\right] \\
&= P_i(\theta_0)\log\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right] + [1-P_i(\theta_0)]\log\left[\frac{1-P_i(\theta_0)}{1-P_i(\theta)}\right] + \\
&+P_i(\theta_0)\left[s(\theta_0-\theta) + e^{-s(\theta_0-\theta)} - 1\right] \qquad (3.10)
\end{aligned}$$

As duas primeiras parcelas dessa equação são a Medida de Informação Global que se tinha obtido na equação 2.1, enquanto que a terceira parcela surgiu devido à covariável t_{ij}. É como se a Medida de Informação Global sofresse uma atualização com a nova modelagem.

3.2.3 Máxima Informação Esperada

Como visto na seção 2.3.3, a seleção do k-ésimo obedecerá à seguinte regra:

$$\begin{aligned}
i_k = argmax_s\{&P_s(0|u_1,\cdots,u_{k-1})J_{u1,\cdots,uk-1}, U_s = 0(\hat{\theta}_{u1,\ldots,uk-1}, U_s = 0) \\
&+P_s(1|u_1,\cdots,u_{k-1})J_{u1,\cdots,uk-1}, U_s = 1(\hat{\theta}_{u1,\ldots,uk-1}, U_s = 1) : s \in R_k\}, \qquad (3.11)
\end{aligned}$$

em que $J_{u_{ij},t_{ij}}(\theta_j) = -\frac{\partial^2}{\partial\theta_j^2}[u_i\log(P_i(\theta)) + (1-u_i)\log(1-P_i(\theta))] + u_i s^2 t_{ij} e^{r+s(\theta_j-b_i)}$

3.2.4 Considerações sobre o CAT com o novo modelo

O objetivo do nosso trabalho é estudar a influência do Tempo de Resposta de em um item na seleção dos próximos itens do CAT. Para isso, o ideal seria contar com um banco de itens real que contemplasse todas as propriedades citadas na seção 1.2.1 e

também que tivesse armazenado o Tempo de Resposta dos itens para toda a amostra que foi utilizada para calibrar o banco.

No entanto, considerando que a construção de um banco de itens com essas características levaria um tempo incompatível para a conclusão e defesa da Dissertação que este trabalho gerará, utilizar-se-á, portanto, um banco de itens simulado.

4 Aplicação com Dados Simulados

A partir de um banco simulado com 500 itens distintos, fizeram-se dois estudos: no primeiro, Estudo I, estruturou-se um algoritmo de CAT sem levar em consideração a covariável Tempo de Resposta (CAT tradicional) e submeteu-se uma amostra de 100 candidatos também simulados e colheu-se o número médio de itens nos diversos CATs realizados (cada respondente foi submetido a 6 testes adaptativos, variando-se o critério de parada em 6 precisões específicas para o estimador). No segundo, Estudo II, estruturou-se outro algoritmo levando-se em conta a covariável Tempo de Resposta que os candidatos levaram em cada item acertado ao longo do teste, colhendo-se também o número médio de itens nos diversos CATs realizados.

Para se cumprir o objetivo do presente estudo é necessário comparar os resultados entre os dois primeiros estudos e perceber a convergência dos dois algoritmos. A grande motivação dessa pesquisa consiste na otimização do algoritmo do CAT, pois acreditou-se que a inserção da covariável Tempo de Resposta reduzirá de maneira significativa o tamanho do teste e sabemos que quanto menor é um teste, mais atrativo ele é. E se isso for feito de maneira que a precisão do exame fique controlada, o objetivo do estudo será cumprido.

Além de tudo isso, foi proposto um estudo especial, Estudo III, para a repetição dos algoritmos para um mesmo examinando. Para isso, escolheram-se 3 alunos com habilidades verdadeiras distintas ($-0,8$, 0 e $0,8$) para realizarem 100 testes cada um com os dois programas (com e sem a utilização da covariável tempo de resposta), estimando, assim, suas respectivas habilidades.

4.1 Estudo I - CAT sem a covariável Tempo de Resposta

Simularam-se os parâmetros dos 500 itens da seguinte forma:

- Parâmetro a_i: As distribuições mais adotadas para o parâmetro a_i são Log-Normal e Qui-Quadrado. A justificativa teórica para o uso dessas distribuições reside no fato de que valores de a_i são tipicamente maiores que zero, sugerindo que a distribuição de a_i pode ser modelada por uma distribuição unimodal e positivamente assimétrica (Mislevy, 1986). Neste estudo, será assumida a distribuição Log-Normal com parâmetros $(0, 0.35)$

- Parâmetro b_i: Como o parâmetro de dificuldade do item pertence ao intervalo $-\infty < b_i < +\infty$ e este está medido na mesma escala de distribuição das habilidades dos

candidatos, pode-se adotar a distribuição Normal $N(0,1)$

- Parâmetro c_i: Como este parâmetro representa a probabilidade de acerto ao acaso, seu valor só pode pertencer ao intervalo $[0,1]$. No presente estudo, adotou-se a distribuição Beta $(2,5)$.

Com os respectivos parâmetros dos itens simulados $a_1 \ldots a_{500}$, $b_1 \ldots b_{500}$ e $c_1 \ldots c_{500}$, simularam-se as habilidades de 100 alunos, aleatoriamente atribuídas, a partir da distribuição Normal padrão, isto é, $\theta_j \sim N(0,1), j = 1 \ldots 100$.

A aplicação foi implementada a partir de um programa desenvolvido na linguarem R. Na primeira parte do programa é criada uma função para calcular os pontos de quadratura e seus respectivos pesos. Esses comandos foram retirados do trabalho de Gray (2001) e constam no Anexo A do presente trabalho.

As habilidades dos 100 alunos são geradas, bem como os parâmetros dos itens. Com essas informações, os acertos e erros de cada item por respondente são possíveis de serem obtidos, pois utilizou-se o ML3, descrito pela Equação 1, em que $D = 1,7$ para que os resultados sejam análogos à Ogiva Normal e, assim, fiquem equivalentes ao modelo utilizado para estimar os parâmetros dos itens. A partir das probabilidades geradas, aplica-se a distribuição Bernoulli para se obter os zeros e uns, definindo o acerto ou erro de cada item por respondente. E isso será feito à medida que o programa for rodando, isto é, em tempo real.

As estimativas iniciais das habilidades de todos os respondentes são igualadas a zero (média da distribuição). Para cada respondente, o programa inicia um *loop*, que é encerrado quando o critério de parada for atingido. Na primeira iteração do *loop*, cinco itens com dificuldades próximos à média são selecionados aleatoriamente (itens cujos parâmetros de dificuldade, b, estejam entre $-0,5$ e $0,5$). Já nas demais iterações, a informação de cada item é calculada pelo Critério de Máxima Informação (Equação 1.6), e o item de maior informação, dada a atual habilidade estimada do respondente, é selecionado. Vale ressaltar que não há repetição de itens para um mesmo aluno e, dessa forma, os itens que já foram expostos são retirados do banco antes do referido cálculo. Obtém-se, em tempo real (online), os acertos ou erros do examinando e guarda-os em um vetor cujo comprimento é igual a quantidade de itens respondidos pelo aluno.

A habilidade do examinando é estimada pelo método EAP, levando em consideração o método da quadratura (equação 1.41). Para a mensuração da habilidade, consideram-se todos os itens, com seus respectivos parâmetros e respostas previamente estimadas, já expostos aos respondentes. Junto com o cálculo da habilidade, também é calculada a variância *a posteriori* associada à estimativa obtida, equação 1.42. Uma vez atendido o critério de parada, finaliza-se o programa e a estimativa da habilidade do candidato é a última obtida.

Tabela 1: Simulação I

Precisão do Estimador	Erro verdadeiro	Número Médio de Itens	% de não convergência
0,35	0,015	12,6 (3,34)	0
0,30	0,010	18,3 (7,13)	0
0,25	0,008	26,9 (12,49)	1
0,20	0,010	42,1 (17,59)	7
0,15	0,021	58,9 (22,14)	33
0,10	0,034	68,1 (19,51)	62

O critério de parada utilizado no algoritmo foi a precisão do estimador (ou o limite de 100 itens para se concluir o teste), que é a raiz quadrada da variância *a posteriori* (equação 1.42). Foram feitos 6 testes para cada candidato com precisões predefinidas em 0,35, 0,30, 0,25, 0,20, 0,15 e 0,10. Desse modo, a habilidade de cada respondente foi estimada uma única vez para cada uma dessas 6 precisões e se obteve a quantidade de itens necessários para a convergência do algoritmo, a medida erro verdadeiro[1] e também a taxa de não convergência do algoritmo, isto é, representa o % de candidatos que precisaram responder as 100 questões limites do teste, ou seja, a precisão do estimador não foi o critério de parada para esses respondentes. Com isso, geraram-se os dados apresentados na tabela 1, que estarão representados nas linhas vermelhas dos gráficos da seção 4.3.

4.2 Estudo II - CAT com a Covariável Tempo de Resposta

O grande objetivo de nosso estudo é a melhora do algoritmo de um CAT. Para isso estabeleceu-se uma nova modelagem (Capítulo 3), que leva em conta o Tempo de Resposta no item.

Para esse estudo, simularam-se os t_{ij} a partir dos parâmetros r e s da modelagem proposta na equação 3.3. Para tanto, precisou-se fixar valores para os parâmetros e utilizou-se o seguinte critério:

Imaginou-se um candidato respondendo o CAT e encontrando um item com dificuldade muito próxima à sua habilidade ($\theta_j \approx b_i$). Imaginou-se, de maneira subjetiva, que o tempo aproximado para o respondente resolver o item está entre 3 e 10 min. Ou seja,

$$3 \leq E(t_{ij}|u_{ij} = 1; \theta_j = b_i) \leq 10.$$

[1] A medida erro verdadeiro, mostrada na tabela 1 a seguir, foi calculada da seguinte forma $erro = \frac{1}{N}\sqrt{\sum_{j=1}^{N}(\hat{\theta} - \theta_j)^2}$, onde N é o total de respondentes que fizeram o teste sem atingir o limite de 100 questões, $\hat{\theta}$ é a estimação da habilidade do respondente e θ_j é a habilidade verdadeira, que só se conhece porque houve a simulação dos dados. Na prática, em um estudo com dados reais, não se conhecerá tal informação.

Tabela 2: Parâmetros r e s fixados para a Simulação II

r	s
-2.3	1.3
-2.1	1.1
-1.9	0.9
-1.7	0.7
-1.5	0.5
-1.3	0.3
-1.1	0.1

Como $E(t_{ij}|\theta_j, u_{ij} = 1) = \frac{1}{e^{r+s(\theta_j - b_i)}}$, podemos concluir, fazendo $\theta_j = b_i$ que

$$3 \leq \frac{1}{e^r} \leq 10.$$

Isso significa que

$$-2,3 \leq r \leq -1,1.$$

Em seguida, imaginou-se um candidato com habilidade superior à dificuldade do item em uma unidade de desvio-padrão ($\theta_j - b_i = 1$). Imaginou-se, de maneira subjetiva, que o tempo aproximado será menor que o caso anterior. Ou seja,

$$E(t_{ij}|u_{ij} = 1; \theta_j - b_i = 1) \leq 3.$$

De onde extrai-se que

$$\frac{1}{e^{r+s}} \leq 3,$$

que pode ser equacionada, para facilitar os cálculos, da seguinte forma

$$\frac{1}{e^{r+s}} = e.$$

Isto é

$$s = -1 - r.$$

Com isso e fixando os valores de r entre $-2,3$ a $-1,1$ obtem-se os seguintes valores para s, constantes na tabela 2.

Consideraram-se os 7 pares de valores da tabela 2 para fixar os parâmetros da nova modelagem, obtendo assim os dados simulados dos t_{ij}.

Os resultados encontram-se nas tabelas 3-9.

Tabela 3: Caso 1

Precisão do Estimador	Erro verdadeiro	Número Médio de Itens	% de não convergência
0,35	0,027	5 (0,17)	0
0,30	0,025	5,3 (0,53)	0
0,25	0,021	7,3 (0,91)	0
0,20	0,018	14,7 (7,68)	0
0,15	0,021	27,4 (19,63)	12
0,10	0,023	37,3 (22,25)	21

Tabela 4: Caso 2

Precisão do Estimador	Erro verdadeiro	Número Médio de Itens	% de não convergência
0,35	0,031	5,1 (0,31)	0
0,30	0,026	6,6 (0,96)	0
0,25	0,021	10,1 (1,14)	0
0,20	0,019	19,4 (10,03)	0
0,15	0,023	34,9 (20,77)	21
0,10	0,029	47,1 (25,15)	39

Tabela 5: Caso 3

Precisão do Estimador	Erro verdadeiro	Número Médio de Itens	% de não convergência
0,35	0,027	6,3 (1,61)	0
0,30	0,018	9,9 (2,16)	0
0,25	0,015	14,3 (3,12)	0
0,20	0,015	23,4 (7,68)	0
0,15	0,023	38,1 (19,11)	22
0,10	0,027	52,2 (21,54)	42

Tabela 6: Caso 4

Precisão do Estimador	Erro verdadeiro	Número Médio de Itens	% de não convergência
0,35	0,015	9,4 (2,85)	0
0,30	0,009	13,1 (3,81)	0
0,25	0,008	18,8 (5,71)	0
0,20	0,009	31,4 (15,19)	0
0,15	0,015	43,6 (18,04)	22
0,10	0,022	56,8 (20,86)	45

Tabela 7: Caso 5

Precisão do Estimador	Erro verdadeiro	Número Médio de Itens	% de não convergência
0,35	0,017	11,2 (3,11)	0
0,30	0,008	15,8 (5,37)	0
0,25	0,008	22,8 (8,70)	0
0,20	0,006	35,8 (13,88)	1
0,15	0,012	53,8 (21,72)	23
0,10	0,017	65,2 (21,50)	44

Tabela 8: Caso 6

Precisão do Estimador	Erro verdadeiro	Número Médio de Itens	% de não convergência
0,35	0,019	11,9 (3,50)	0
0,30	0,014	16,9 (6,63)	0
0,25	0,012	25,1 (11,58)	0
0,20	0,009	41,1 (16,81)	4
0,15	0,012	53,2 (22,31)	29
0,10	0,020	61,9 (19,32)	56

Tabela 9: Caso 7

Precisão do Estimador	Erro verdadeiro	Número Médio de Itens	% de não convergência
0,35	0,021	12,8 (3,79)	0
0,30	0,016	17,9 (7,55)	0
0,25	0,014	26,1 (14,19)	1
0,20	0,011	40,1 (18,05)	5
0,15	0,013	59,5 (22,59)	37
0,10	0,020	66,7 (18,38)	58

4.3 Comparação Gráfica dos Estudos I e II

Os resultados obtidos no Estudo I (CAT sem a covariável Tempo de Resposta) são representados pelo gráfico vermelho e serão comparados com os resultados dos 7 casos do Estudo II (CAT com a covariável Tempo de Resposta), linha azul dos gráficos.

Nesses gráficos, o eixo das abscissas representa a precisão do estimador, que, nos estudos, foi o critério de parada do algoritmo; já o eixo das ordenadas representa o número médio de questões que os respondentes tiveram ao atingirem o critério de parada.

Nota-se, em todos os gráficos, a linha azul bem abaixo da linha vermelha. Demonstrando como o algoritmo utilizado no Estudo II é mais eficiente, pois convergiu utilizando

um número significativamente menor de questões quando comparado com o algoritmo do Estudo I.

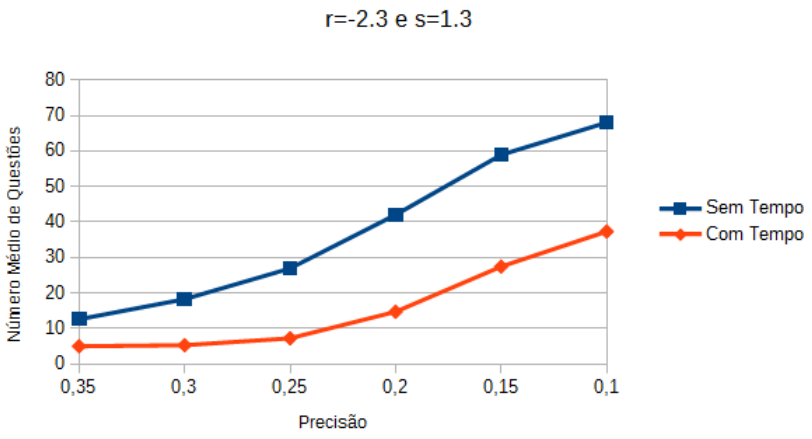


Figura 7: Comparação entre o Estudo I e o caso 1 do Estudo II

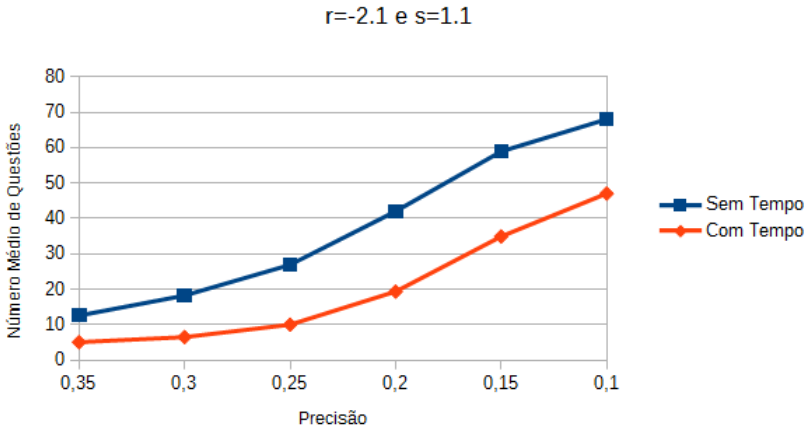


Figura 8: Comparação entre o Estudo I e o caso 2 do Estudo II

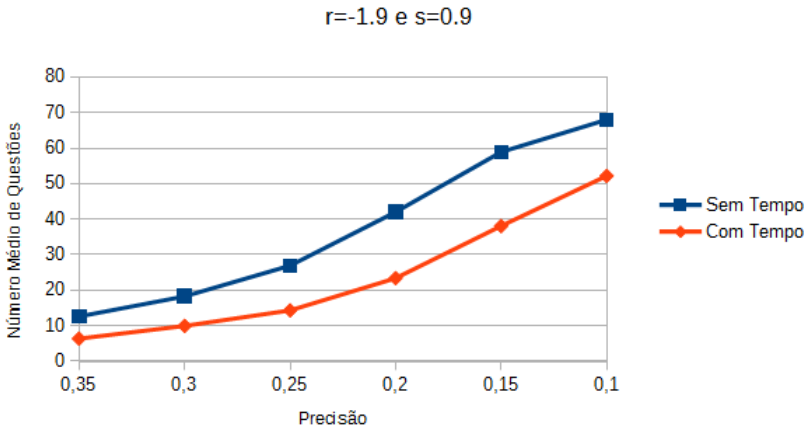


Figura 9: Comparação entre o Estudo I e o caso 3 do Estudo II

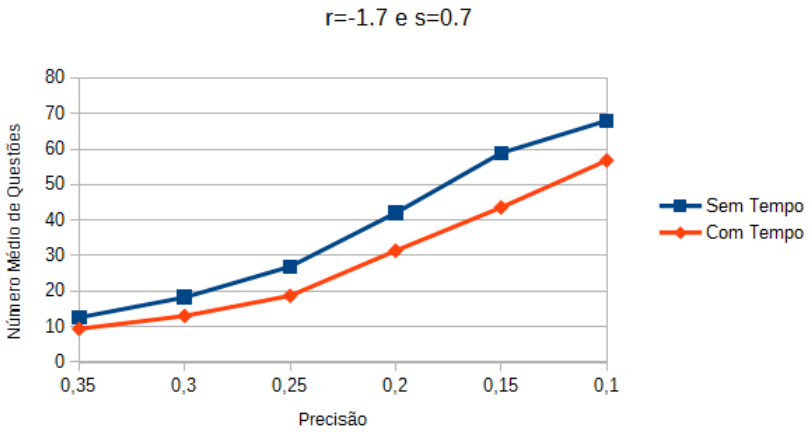


Figura 10: Comparação entre o Estudo I e o caso 4 do Estudo II

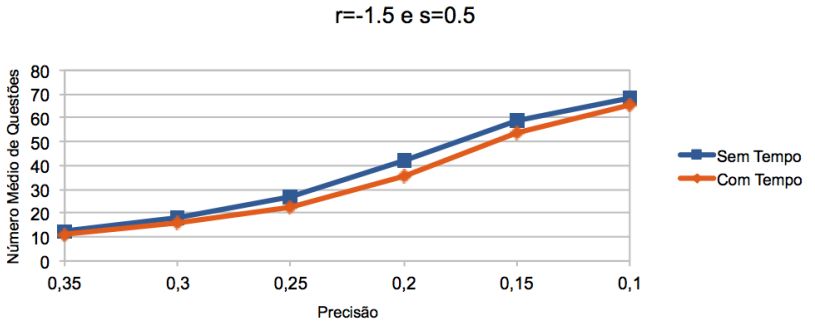


Figura 11: Comparação entre o Estudo I e o caso 5 do Estudo II

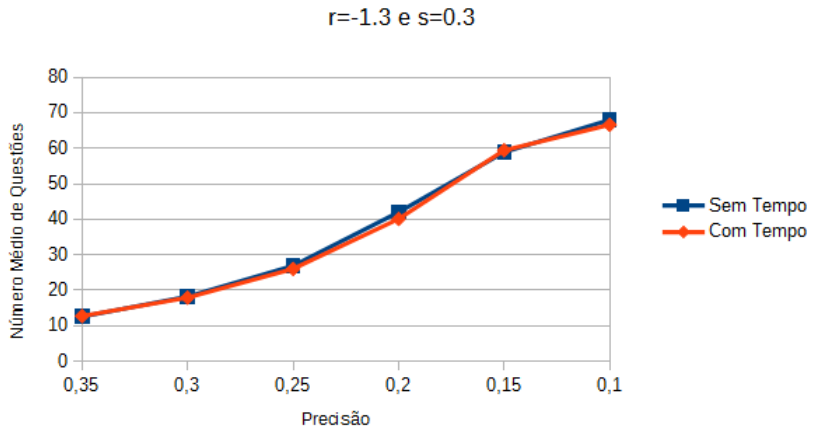


Figura 12: Comparação entre o Estudo I e o caso 6 do Estudo II

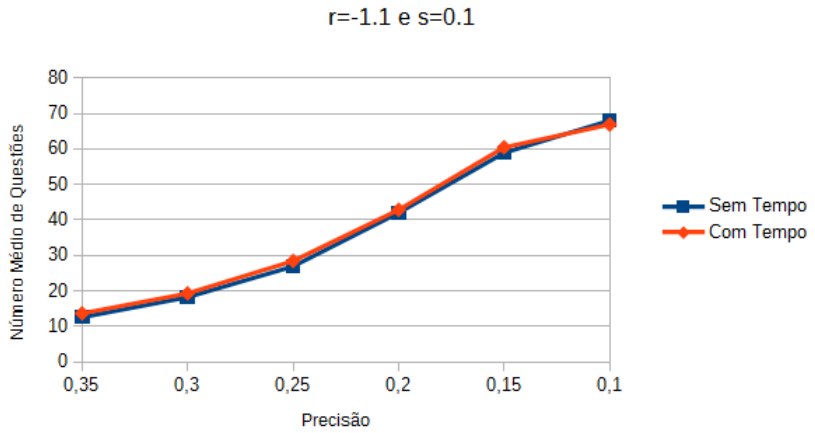


Figura 13: Comparação entre o Estudo I e o caso 7 do Estudo II

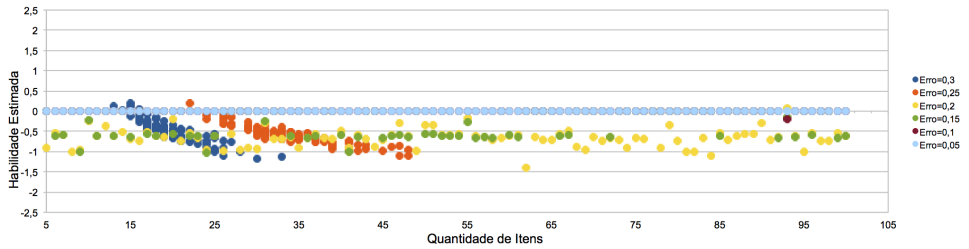
4.4 Estudo III

Após a análise desses dados (Estudos I e II), sentiu-se a necessidade de repetir os testes para o mesmo candidato algumas vezes a fim de perceber a consistência da convergência dos dois algoritmos desenvolvidos nesse trabalho (um com e o outro sem a Covariável Tempo de Resposta). Nesse sentido, escolheram-se 3 candidatos com habilidades verdadeiras conhecidas (Aluno 1: $\theta = -0,8$, Aluno 2: $\theta = 0$ e Aluno 3: $\theta = 0,8$) e repetiram-se as simulações dos testes adaptativos 100 vezes, utilizando como critério de parada 6 precisões distintas ($0,3$, $0,25$, $0,2$, $0,15$, $0,1$ e $0,05$) para os dois programas estudados. Para a simulação dos testes com o algoritmo que utilizou a nova modelagem, foram utilizados os parâmetros $r = -2,1$ e $s = 1,1$.

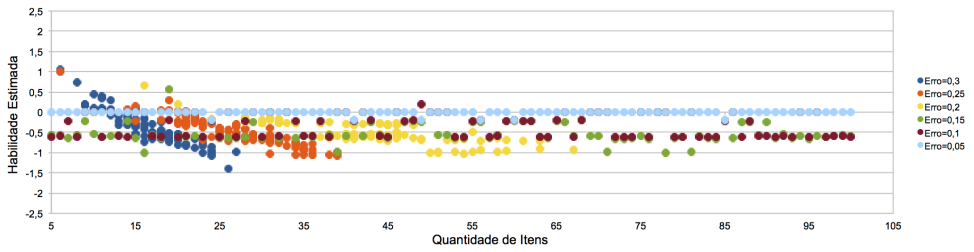
Para apresentar os resultados, fizeram-se 2 tipos de gráficos. No primeiro tipo (gráficos 14a-b, 15a-b, 16a-b), esboçaram-se as 100 habilidades estimadas versus a quantidade de itens administrados nesses 100 testes, para cada uma das 6 precisões, para cada um dos programas. No segundo tipo (gráficos 14c, 15c e 16c), esboçou-se a evolução da habilidade estimada à medida que os itens eram admistrados no CAT. Nesse caso utilizou-se como critério de parada o número limite de 100 questões. Como foram 100 repetições, esboçou-se uma linha contínua representando a média das estimativas das habilidades e uma linha tracejada com o correspondente Intervalo de Confiança de 90%. Naturalmente, os dois programas foram utilizados. A cor azul representa os resultados do algoritmo com a covariável tempo de resposta e a cor vermelha o algoritmo sem a covariável tempo de resposta.

4.4.1 Estudo III, Aluno 1 ($\theta = -0,8$)

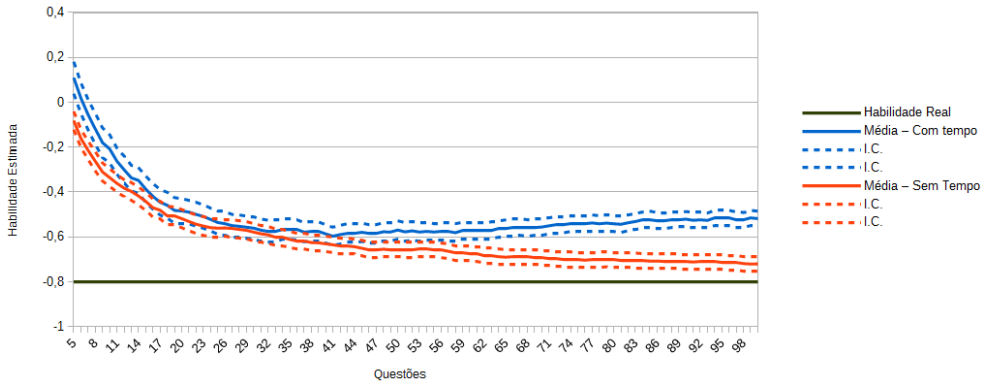
Aluno 1	Com a Covariável Tempo		Sem a Covariável Tempo	
Precisão (Critério de parada)	Número médio de questões	Média da estimativa de θ	Número médio de questões	Média da estimativa de θ
0,30	15,9	-0,317 (0,427)	20,1	-0,492 (0,277)
0,25	26,3	-0,472 (0,349)	33,7	-0,573 (0,242)
0,20	42,9	-0,515 (0,27)	63,6	-0,663 (0,222)
0,15	65,1	-0,517 (0,228)	84,8	-0,61 (0,154)
0,10	84,8	-0,469 (0,201)	-	- (-)
0,05	96	-0,201 (0,003)	-	- (-)



(a) Sem a Covariável Tempo de Resposta



(b) Com a Covariável Tempo de Resposta

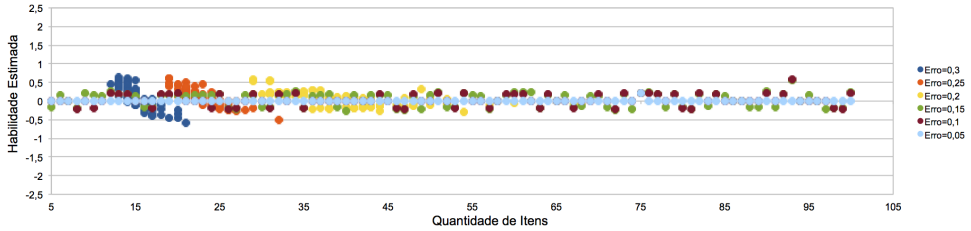


(c) Evolução do CAT para o Aluno 1

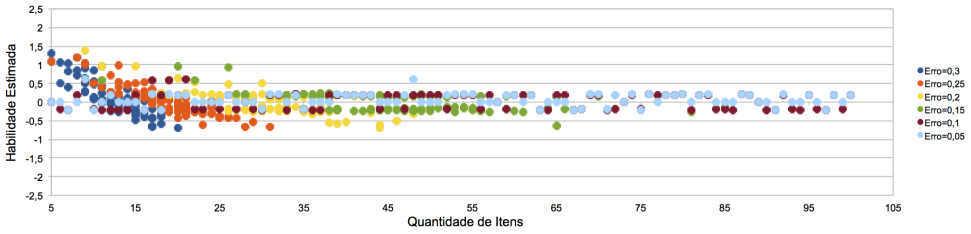
Figura 14: Estudo III, Aluno 1 ($\theta = -0,8$)

4.4.2 Estudo III, Aluno 2 ($\theta = 0$)

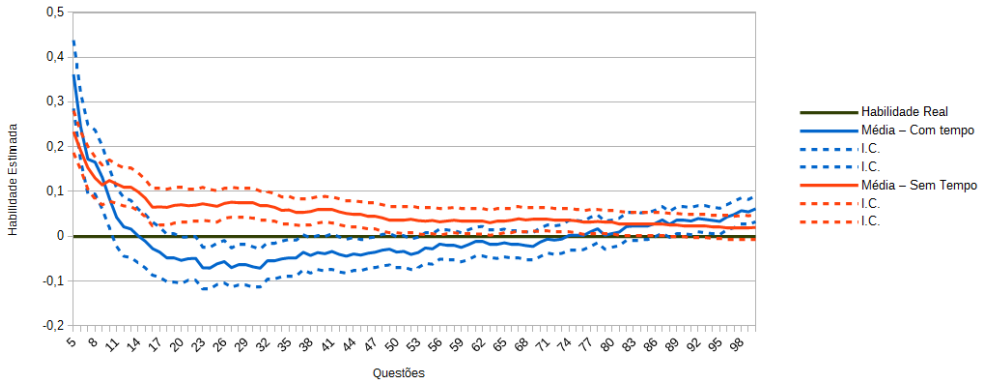
Aluno 2	Com a Covariável Tempo		Sem a Covariável Tempo	
Precisão (Critério de parada)	Número médio de questões	Média da estimativa de θ	Número médio de questões	Média da estimativa de θ
0,30	12,5	0,097 (0,439)	15,1	0,102 (0,262)
0,25	18,7	0,034 (0,386)	23,1	0,088 (0,203)
0,20	29,7	-0,005 (0,327)	39,6	0,069 (0,187)
0,15	45,1	0,018 (0,259)	63,2	0,053 (0,183)
0,10	64,1	0,034 (0,224)	84,1	0,057 (0,205)
0,05	84,8	0,113 (0,197)	-	- (-)



(a) Sem a Covariável Tempo de Resposta



(b) Com a Covariável Tempo de Resposta

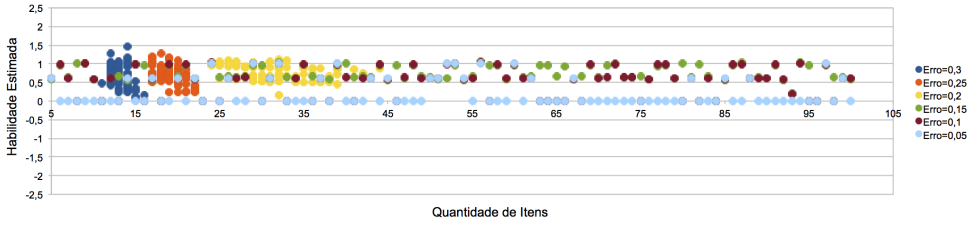


(c) Evolução do CAT para o Aluno 2

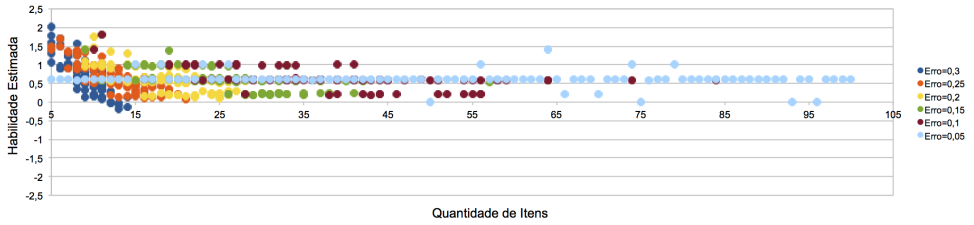
Figura 15: Estudo III, Aluno 2 ($\theta = 0$)

4.4.3 Estudo III, Aluno 3 ($\theta = 0,8$)

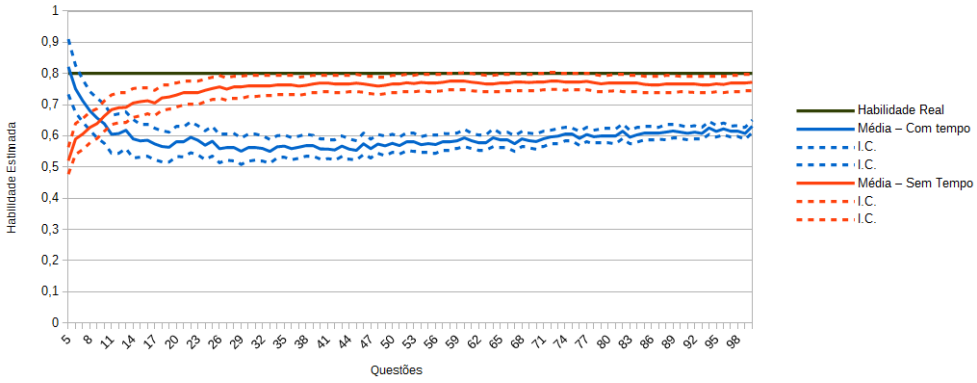
Aluno 3	Com a Covariável Tempo		Sem a Covariável Tempo	
Precisão (Critério de parada)	Número médio de questões	Média da estimativa de θ	Número médio de questões	Média da estimativa de θ
0,30	9	0,719 (0,478)	13,2	0,704 (0,269)
0,25	12,4	0,67 (0,364)	18,9	0,736 (0,223)
0,20	18,3	0,623 (0,317)	30,5	0,769 (0,194)
0,15	26,1	0,603 (0,304)	51,7	0,764 (0,187)
0,10	38,4	0,627 (0,262)	68,3	0,748 (0,198)
0,05	57,1	0,632 (0,147)	91,4	0,743 (0,195)



(a) Sem a Covariável Tempo de Resposta



(b) Com a Covariável Tempo de Resposta



(c) Evolução do CAT para o Aluno 3

Figura 16: Estudo III, Aluno 3 ($\theta = 0,8$)

5 Conclusão e Trabalhos Futuros

A possibilidade de se obter um teste personalizado para estimarmos, com precisão controlada, a habilidade de cada candidato e que elas sejam comparáveis entre si, faz com que o número de pesquisas em Testes Adaptativos Informatizados (CAT) venham crescendo. Diferentemente dos testes tradicionais (papel e caneta), o CAT administra itens adequados a cada respondente. Essa seleção baseia-se na característica dos itens e na estimativa da habilidade do examinando. Para isso, a construção de um banco de itens e o desenvolvimento de um algoritmo para a seleção adaptativa dos itens se fazem necessários. Este trabalho se propôs a discutir métodos estatísticos que envolvam esses assuntos, especialmente a seleção adaptativa de itens no CAT.

A seleção adaptativa de itens depende da estimativa da habilidade corrente do candidato. E esse foi o grande foco da pesquisa.

Inicialmente, criou-se um modelo estatístico que levou em conta a covariável Tempo de Resposta. Fez-se todo o estudo teórico necessário para utilização desse modelo. Implementou-se dois algoritmos de Testes Adaptativos Informatizados: o primeiro, um programa tradicional de CAT, que utilizava apenas a resposta do candidato para a estimação iterativa de sua habilidade, escolhendo as próximas questões do teste com o critério da Máxima Informação, até a convergência do algoritmo. O segundo, que representa a aplicação do estudo principal dessa pesquisa, que, além de levar em conta a resposta do respondente, também considerou o Tempo de Resposta das questões acertadas por ele, estimando, mais eficientemente, a habilidade corrente do respondente, escolhendo melhor a próxima questão do teste com o critério de Máxima Informação, fazendo isso de maneira iterativa até que o critério de parada fosse atingido mais rapidamente em comparação com o primeiro algoritmo.

Nos Estudos I e II, fez-se uma simulação dos respondentes e do banco de itens, aplicando-se esses dois algoritmos e comparou-se a convergência desses programas por meio do número médio de questões necessárias para finalizar o teste, utilizando como critério de parada a precisão do estimador. Percebeu-se uma considerável melhora nos resultados do segundo algoritmo em comparação aos do primeiro, pois foram necessárias menos questões para se estimar as habilidades dos respondentes. No entanto, quando os dados foram simulados com o parâmetro s próximo a 0 (como $0,1$ e $0,3$), os resultados entre os dois algoritmos foram muito próximos, com uma discreta melhora do programa que utilizou a covariável tempo de resposta. E, de certa forma, isso já era esperado, pois no estudo teórico da nova modelagem, percebeu-se que as novas equações (3.9 e 3.10) sofriam uma "atualização", em relação à modelagem tradicional, acrescentando-se uma

parcela que dependia diretamente do parâmetro s. Portanto, quanto mais próximo de 0 é o valor de s, menor é a "atualização" sofrida na nova modelagem, fazendo com que os dois modelos se aproximem.

Já no Estudo III, estimou-se a habilidade de 3 determinados examinandos 100 vezes nos dois programas e percebeu-se que o algoritmo da nova modelagem convergia mais rápido do que o tradicional para os 3 alunos, pois o número médio de questões era consideravelmente menor. No entanto, comparando-se a média das estimativas das habilidades, com as respectivas habilidades verdadeiras, percebeu-se que para o aluno 2, o programa que utilizava a covariável tempo de resposta era ligeiramente melhor. Já, para os alunos 1 e 3, o outro programa apresentava melhores estimativas médias. Observou-se também que ao utilizar como critério de parada estimadores mais precisos (precisão $0,10$ ou $0,05$), os algoritmos tradicionais não convergiam até o número limite de 100 questões. Ainda nesse estudo, os gráficos 14c, 15c e 16c apresentam a evolução das médias das estimativas das habilidades dos alunos 1, 2 e 3, respectivamente, em função da administração dos itens. Percebeu-se, de maneira geral, que se o critério de parada é o número de itens administrados, quanto menor esse número (testes mais curtos), a média das estimativas quando comparada ao valor verdadeiro fica melhor no programa que leva em conta a covariável tempo de resposta. E quanto maior aquele número (testes mais longos), o programa tradicional leva vantagem.

Nessa perspectiva, conclui-se que a utilização da covariável tempo de resposta, indica um caminho de que pesquisas nessa área podem melhorar a convergência dos algoritmos de Testes Adaptativos Informatizados, no entanto há necessidade de se aprofundar os estudos, implementando novos modelos com a covariável tempo de resposta, comparando-se os resultados obtidos neste trabalho. Entende-se também que a utilização de dados reais é fundamental para a evolução desse estudo.

Com isso os objetivos do presente trabalho foram cumpridos.

Para futuros trabalhos, sugere-se o aprofundamento nos estudos ligados ao novo modelo, que, por simplicidade, adotamos a distribuição exponencial e cujos parâmetros ainda foram simplificados. Sugere-se também que sejam desenvolvidos algoritmos que além de utilizarem a Máxima Informação como critério de seleção dos próximos itens, utilizem também a Máxima Informação Global e a Máxima Informação Esperada.

Referências

ABAD, F. J. et al. Efectos de las omisiones en la calibracion de un test adaptativo informatizado. *Metodologia de las Ciencias del Comportamiento*, p. 1–6, 2004. Citado na página 35.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da Resposta ao Item: conceitos e aplicações*. [S.l.]: São paulo: ABE - Associação Brasileira de Estatística, 2000. Citado 12 vezes nas páginas 7, 14, 19, 22, 23, 24, 25, 26, 27, 30, 31 e 75.

AZEVEDO, C. L. N. Modelos longitudinais de grupos múltiplos multiníveis na teoria da resposta ao item: Métodos de estimação e seleção estrutural sob uma perspectiva bayesiana. *Tese de Doutorado em Ciencias - USP/SP*, p. 265p, 2008. Citado na página 31.

BAZAN, J. L. Uma família de modelos de resposta ao item normal assimétrica. *Tese de Doutorado em Estatística - USP/SP*, p. 133p, 2005. Citado na página 31.

CHANG, H. H.; YING, Z. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, n. 20, p. 213–229, 1996. Citado na página 39.

COSTA, D. R. Métodos estatísticos em testes adaptativos informatizados. *Dissertação de Mestrado em Estatística - UFRJ*, p. 107p, 2009. Citado 3 vezes nas páginas 15, 37 e 38.

EMBRETSON, S. E. *Item response theory for psychologists*. [S.l.]: Lawrence Erlbaum Associates, Inc, 2013. Citado na página 14.

GEORGIADOU, E. et al. A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 2007. Citado na página 37.

GRAY, R. advanced statistical computing. *BIO 248*, p. 342p, 2001. Citado 3 vezes nas páginas 31, 49 e 75.

HAMBLETON, R. K. et al. *Fundamentals of Item Response Theory*. [S.l.]: Newbury Park : Sage Publications, 2001. Citado na página 20.

HERRANDO, S. Tests adaptativos computerizados: una sencilla solucion al problema de la estimacion con puntuaciones perfectas y cero. In: BIOMETRIC SOCIETY, SEGOVIA, ESPANA. *II Conferencia Espanola de Biometria*. [S.l.], 1989. Citado na página 35.

KINGSBURY, G. G.; ZARA, A. R. Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, p. 359–375, 1989. Citado na página 38.

LABARRERE, J. G. et al. Testes adaptativos computadorizados. *Revista Brasileira de Biometria*, v. 29, n. 2, p. 229–261, 2011. Citado na página 74.

LINDEN, W. J. v. d.; HAMBLETON, R. K. *Handbook of modern item response theory*. [S.l.]: Springer science Business Media, LLC, 2013. Citado na página 18.

LINDEN, W. J. Van der. Baysian iten selection criteria for adaptative testing. *Psychometrika, 63*, 1998. Citado 2 vezes nas páginas 38 e 41.

LINDEN, W. J. Van der; GLAS, C. A. W. Elements of adaptive testing. *Statistical for Social and Behavioral Sciences*, 2010. Citado 3 vezes nas páginas 7, 38 e 39.

LORD, F. M. Applications of item response theory to pratical testing problems. *Hillsdale: Lawrence Erlbaum Associates, Inc.*, 1980. Citado 2 vezes nas páginas 34 e 38.

MIGON, H. S.; GAMERMAN, D. *Statistical Inference - an integrated approach*. [S.l.]: Edward Arnold, 2009. Citado na página 40.

MISLEVY, R. J.; STOCKING, M. L. *Applied Psychological Measurement*. [S.l.]: A Consumer's Guide to logistic and BILOG, 1989. Citado na página 30.

MOREIRA, F. J. Sistemática para a implantação de testes adaptativos informatizados baseados na teoria da resposta ao item. *Tese de Doutorado*, 2011. Citado na página 23.

NAVAS, M. J. Equiparacion de puntuaciones. *Psicometría*, p. 293–369, 1996. Citado na página 23.

OLEA, J. et al. *Tests informatizados:Fundamentos y aplicationes*. [S.l.]: Pirámide, 1999. Citado 2 vezes nas páginas 21 e 39.

PASQUALI, L. *Teoria e Métodos de Medida em Ciências do Comportamento*. [S.l.]: Instituto de Psicologia / UnB: INEP, 1996. Citado na página 21.

PASQUALI, L. Princípios de elaboração de escalas psicológicas. *Revista de Psiquiatria Clínica*, v. 5, n. 25, p. 206–213, 1998. Citado na página 21.

SEGALL, D. O. Computerized adaptive testing. *Encyclopedia of Social Measurement*, Elsevier Inc., v. 1, n. 1, p. 429–438, 2005. Citado 2 vezes nas páginas 21 e 35.

WAINER, H. Computerized adaptive testing: A primer. *New Jersey: Lawrence Erlbaum Associates*, 2000. Citado na página 15.

Anexos

ANEXO A – Algoritmos Utilizados

A.1 Algoritmo da Função Gauher

```
gauher <- function(n) {# Gauss-Hermite: returns x,w so that
#\int_-\infty^\infty exp(-x^2) f(x) dx \doteq \sum w_i f(x_i)
EPS <- 3e-14
PIM4 <- .7511255444649425
MAXIT <- 10
m <- trunc((n+1)/2)
x <- w <- rep(-1,n)
for (i in 1:m) {
if (i==1) {
z <- sqrt(2*n+1)-1.85575*(2*n+1)^(-.16667)
} else if(i==2) {
z <- z-1.14*n^.426/z

} else if (i==3) {
z <- 1.86*z-.86*x[1]
} else if (i==4) {
z <- 1.91*z-.91*x[2]
} else {
z <- 2.*z-x[i-2]
}
for (its in 1:MAXIT) {
p1 <- PIM4
p2 <- 0
for (j in 1:n) {
p3 <- p2
p2 <- p1
p1 <- z*sqrt(2/j)*p2-sqrt((j-1)/j)*p3
}
pp <- sqrt(2*n)*p2
z1 <- z
z <- z1-p1/pp
if(abs(z-z1) <= EPS) break
}
x[i] <- z
x[n+1-i] <- -z
w[i] <- 2/(pp*pp)
w[n+1-i] <- w[i]
}
list(x=x,w=w)
}
```

A.2 Algoritmo de um CAT sem a Covariável Tempo de Resposta

```
#1) Quantidade de Alunos
na<-100

#2) Habilidades
seed<-123
set.seed(seed)
theta<-rnorm(na)

#3) Precisao
preci<-seq(from=0.35,to=0.10,by=-0.05)

#4) Simulacao dos parametros dos itens
ni<-500
set.seed(seed)
par.a<-rlnorm(ni,0,0.35)
set.seed(seed)
par.b<-rnorm(ni)
set.seed(seed)
par.c<-rbeta(ni,2,5)
Item<-seq(1,ni, by=1)
quest<-data.frame(cbind(par.a,par.b,par.c,Item))

ni<-nrow(quest) # Quantidade de Itens

#5) Matrizes importantes
mp<-matrix(NA,ncol=na,nrow=ni)
ma<-matrix(NA,ncol=na,nrow=ni)

#6) Numero de pontos de Quadratura e Funcao Gauher
nn<-30
source("gauher.R")
u<-gauher(nn)

#7) Modelo normal
d <- 1.7

#8) Calculo das probabilidades de acertos
for (i in 1:ni) {
  for (j in 1:na) {
    mp[i,j]<-quest[i,3]+(1-quest[i,3])/(1+exp(-d*quest[i,1]*(theta[j]-quest[i,2])))
  }
}

#9) Matriz de acertos/erros
for (i in 1:ni) {
  for (j in 1:na) {
    set.seed(seed)
    ma[i,j]<-rbinom(1,1,mp[i,j])
  }
}

theta_mat<-matrix(NA,nrow=na,ncol=6) #tabela de apoio

#10) Inicializacao do teste
ninit<-5
```

```
matriz<-matrix(NA,nrow=length(preci),ncol=4)
for (e in 1:length(preci)){
  pp<-preci[e]
    for (j in 1:na){
    nq<-ninit
    quest_j<-subset(quest, par.b > -0.5 & par.b < 0.5)
    iq<-sample(nrow(quest_j),size=ninit,replace=FALSE)
    a<-which(quest$Item %in% quest_j[iq,]$Item)
    quest_jj<-quest[-a,]
    resp<-ma[a,j]

#11) Estimacao inicial de theta
    L<-rep(0,nn)
    A<-0
    A2<-0
    B<-0
    R<-0
    R2<-0
    V<-0
    for (k in 1:nn){
      for (c2 in 1:nq){
        pij<-quest$par.c[a[c2]]+(1-quest$par.c[a[c2]])/(1+exp(-d*quest$par.a[a[c2]]
        *(u$x[k]-quest$par.b[a[c2]])))
        L[k]<-L[k]+resp[c2]*log(pij)+(1-resp[c2])*log(1-pij)}
      L[k] <- exp(L[k])
      B <- B+L[k]*dnorm(1,u$x[k])*u$w[k]
      A <- A+u$x[k]*L[k]*dnorm(1,u$x[k])*u$w[k]
      A2 <- A2+(u$x[k])^2*L[k]*dnorm(1,u$x[k])*u$w[k]}
    R=A/B
    R2=A2/B
    V=R2-(R)^2
    theta_est<-R
    prec<-V
    erro<-sqrt(V)

#12) Criterio de parada:
    while ((erro> pp | erro< -pp) & nq<101){
      nq<-nq+1

#13) Informacao de Fisher e escolha da proxima questao
      Ii <- rep(NA,(nrow(quest_jj)))
      for (i2 in 1:nrow(quest_jj)){
        pij<-quest_jj[i2,3]+(1-quest_jj[i2,3])/
        (1+exp(-d*quest_jj[i2,1]
        *(theta_est-quest_jj[i2,2])))
        Ii[i2] <- d^2*(quest_jj[i2,1])^2*((1-pij)/pij)
        *((pij-quest_jj[i2,3])/(1-quest_jj[i2,3]))^2
      }
      lin <- which(Ii==max(Ii, na.rm=T))
      a<-c(a,which(quest$Item %in% quest_jj[lin,]$Item))
      quest_jj<-quest_jj[-lin,]
      resp<-ma[a,j]

#14) Estimacao de theta
      L<-rep(0,nn)
      A<-0
      A2<-0
```

```
      B<-0
      R<-0
      R2<-0
      V<-0
      for (k in 1:nn){
        for (c2 in 1:nq){
          pij<-quest$par.c[a[c2]]+(1-quest$par.c[a[c2]])/(1+exp(-d*quest$par.a[a[c2]]
          *(u$x[k]-quest$par.b[a[c2]])))
          L[k]<-L[k]+resp[c2]*log(pij)+(1-resp[c2])*log(1-pij)}
        L[k] <- exp(L[k])
        B <- B+L[k]*dnorm(1,u$x[k])*u$w[k]
        A <- A+u$x[k]*L[k]*dnorm(1,u$x[k])*u$w[k]
        A2 <- A2+(u$x[k])^2*L[k]*dnorm(1,u$x[k])*u$w[k]}
      R=A/B
      R2=A2/B
      V=R2-(R)^2
      theta_est<-R
      prec<-V
      erro<-sqrt(V)
    }

    theta_mat[j,1]<-theta[j]
    theta_mat[j,2]<-theta_est
    theta_mat[j,3]<-(theta_est-theta[j])^2
    theta_mat[j,4]<-erro
    theta_mat[j,5]<-prec
    theta_mat[j,6]<-nq

  }

  yyy<-which(theta_mat[,4]<pp) #Alunos que alcancaram a precisao

  matriz[e,1]<-preci[e]
  matriz[e,2]<-sqrt(sum(theta_mat[yyy,3]))/length(yyy)
  matriz[e,3]<-mean(theta_mat[yyy,6])
  matriz[e,4]<-1-(length(yyy)/na)

  e<-e+1
}
```

A.3 Algoritmo de um CAT com a Covariável Tempo de Resposta

```
#1) Quantidade de Alunos
na<-100

#2) Habilidades
seed<-123
set.seed(seed)
theta<-rnorm(na)

#3) Precisao
preci<-seq(from=0.35,to=0.1,by=-0.05)

#4) Simulacao dos parametros dos itens
ni<-200
set.seed(seed)
par.a<-rlnorm(ni,0,0.35)
set.seed(seed)
par.b<-rnorm(ni)
set.seed(seed)
par.c<-rbeta(ni,2,5)
Item<-seq(1,ni, by=1)
quest<-data.frame(cbind(par.a,par.b,par.c,Item))


ni<-nrow(quest) #Quantidade de Itens

#5) Matrizes importantes
mp<-matrix(NA,ncol=na,nrow=ni)
ma<-matrix(NA,ncol=na,nrow=ni)
mt<-matrix(NA,ncol=na,nrow=ni) #matriz dos tempos
mlam<-matrix(NA,ncol=na,nrow=ni) #matriz dos lambdas

#6) Numero de pontos de Quadratura e Funcao Gauher
nn<-30
source("gauher.R")
u<-gauher(nn)

#7) Modelo normal
d <- 1.7

#8) Calculo das probabilidades de acertos
for (i in 1:ni) {
  for (j in 1:na) {
    mp[i,j]<-quest[i,3]+(1-quest[i,3])/(1+exp(-d*quest[i,1]*(theta[j]-quest[i,2])))
  }
}

#9) Matriz de acertos/erros
for (i in 1:ni) {
  for (j in 1:na) {
    set.seed(seed)
    ma[i,j]<-rbinom(1,1,mp[i,j])
  }
}
theta_mat<-matrix(NA,nrow=na,ncol=6) #Tabela de apoio
```

```
#10) Indice das questoes acertadas
I<-which(ma==1, arr.ind=TRUE)

#11) Simulacao dos tempos de resposta para as questoes acertadas
r<- seq(-2.3,-1.1,0.2) #parametro r
matriz<-matrix(NA,nrow=length(preci)*length(r),ncol=6) #Tabela final
row<-1
for (rr in 1:length(r)) {
s<-1-r[rr] #parametro s

for (z in 1:nrow(I)){
  bj<-quest$par.b[I[z,1]] #parametro b das questoes acertadas
  lambda<- r[rr]+s*(theta[I[z,2]]-bj) #lambda
  mlam[I[z,1],I[z,2]]<-exp(lambda)
  set.seed(seed)
  mt[I[z,1],I[z,2]]<-rexp(1,exp(lambda)) #simulacao dos tempos para itens corretos
}

mt[which(mt>500,arr.ind=TRUE)]<-500 #limitacao com o tempo da prova


#12) Inicializacao do teste
ninit<-5 #numero inicial de questoes
for (e in 1:length(preci)){ #precisoes/criterio de parada
  pp<-preci[e]

  for (j in 1:na){ #por aluno

    nq<-ninit
    quest_j<-subset(quest, par.b > -0.5 & par.b < 0.5)
    set.seed(seed)
    iq<-sample(nrow(quest_j),size=ninit,replace=FALSE) #selecao das questoes iniciais

    a<-which(quest$Item %in% quest_j[iq,]$Item)
    quest_jj<-quest[-a,] #retirar as questoes iniciais do banco

    resp<-ma[a,j] #respostas
    t<-mt[a,j] #tempos
    lam<-mlam[a,j] #lambdas

#13) Estimacao inicial de theta
    L<-rep(0,nn)
    A<-0
    A2<-0
    B<-0
    R<-0
    R2<-0
    V<-0
    for (k in 1:nn){
      for (c2 in 1:nq){
        if (resp[c2]==0){ #se errou
          pij<-quest$par.c[a[c2]]+(1-quest$par.c[a[c2]])/(1+exp(-d*quest$par.a[a[c2]]
          *(u$x[k]-quest$par.b[a[c2]]))) #p_i(\theta)
          L[k]<-L[k]+resp[c2]*log(pij)+(1-resp[c2])*log(1-pij) #log-verossimilhanca
         } else { #se acertou
            pij<-quest$par.c[a[c2]]+(1-quest$par.c[a[c2]])/(1+exp(-d*quest$par.a[a[c2]]
```

```
            *(u$x[k]-quest$par.b[a[c2]]))) #p_i(\theta)
            aaa<-r[rr]+s*(u$x[k]-quest$par.b[a[c2]])-mt[a[c2],j]*exp(r[rr]
            +s*(u$x[k]-quest$par.b[a[c2]]))
            L[k]<-L[k]+resp[c2]*log(pij)+resp[c2]*(aaa)+(1-resp[c2])*log(1-pij)
          }}
      L[k] <- exp(L[k])
      B <- B+L[k]*dnorm(1,u$x[k])*u$w[k]
      A <- A+u$x[k]*L[k]*dnorm(1,u$x[k])*u$w[k]
      A2 <- A2+(u$x[k])^2*L[k]*dnorm(1,u$x[k])*u$w[k]}
    R=A/B
    R2=A2/B
    V=R2-(R)^2
    theta_est<-R
    prec<-V
    erro<-sqrt(V)

#14) Criterio de parada:
    while ((erro> pp | erro< -pp) & nq<101){
      nq<-nq+1

#15) Informacao de Fisher e escolha da proxima questao
      Ii <- rep(NA,(nrow(quest_jj))) #Informacao de Fisher
      for (i2 in 1:nrow(quest_jj)){
        pij<-quest_jj[i2,3]+(1-quest_jj[i2,3])/(1+exp(-d*quest_jj[i2,1]
        *(theta_est-quest_jj[i2,2])))
        Ii[i2] <- d^2*(quest_jj[i2,1])^2*((1-pij)/pij)
        *((pij-quest_jj[i2,3])/(1-quest_jj[i2,3]))^2 + pij*(s^2)
      }
      lin <- which(Ii==max(Ii, na.rm=T))
      a<-c(a,which(quest$Item %in% quest_jj[lin,]$Item))
      quest_jj<-quest_jj[-lin,]
      resp<-ma[a,j]
      t<-mt[a,j]
      lam<-mlam[a,j]

#16) Estimacao de theta
      L<-rep(0,nn)
      A<-0
      A2<-0
      B<-0
      R<-0
      R2<-0
      V<-0
      for (k in 1:nn){
        for (c2 in 1:nq){
          if (resp[c2]==0){
            pij<-quest$par.c[a[c2]]+(1-quest$par.c[a[c2]])/(1+exp(-d*quest$par.a[a[c2]]
            *(u$x[k]-quest$par.b[a[c2]])))
            L[k]<-L[k]+resp[c2]*log(pij)+(1-resp[c2])*log(1-pij)} else {
              pij<-quest$par.c[a[c2]]+(1-quest$par.c[a[c2]])/(1+exp(-d*quest$par.a[a[c2]]
              *(u$x[k]-quest$par.b[a[c2]]))) #p_i(\theta)
              aaa<-r[rr]+s*(u$x[k]-quest$par.b[a[c2]])-mt[a[c2],j]*exp(r[rr]
              +s*(u$x[k]-quest$par.b[a[c2]]))
              L[k]<-L[k]+resp[c2]*log(pij)+resp[c2]*(aaa)+(1-resp[c2])*log(1-pij)
            }}
        L[k] <- exp(L[k])
        B <- B+L[k]*dnorm(1,u$x[k])*u$w[k]
```

```
        A <- A+u$x[k]*L[k]*dnorm(1,u$x[k])*u$w[k]
        A2 <- A2+(u$x[k])^2*L[k]*dnorm(1,u$x[k])*u$w[k]}
      R=A/B
      R2=A2/B
      V=R2-(R)^2
      theta_est<-R
      prec<-V
      erro<-sqrt(V)
    }
  theta_mat[j,1]<-theta[j]
  theta_mat[j,2]<-theta_est
  theta_mat[j,3]<-(theta_est-theta[j])^2
  theta_mat[j,4]<-erro
  theta_mat[j,5]<-prec
  theta_mat[j,6]<-nq

  }
yyy<-which(theta_mat[,4]<pp) #Alunos que alcancaram a precisao

matriz[row,1]<-preci[e]
matriz[row,2]<-sqrt(sum(theta_mat[yyy,3]))/length(yyy)
matriz[row,3]<-mean(theta_mat[yyy,6])
matriz[row,4]<-1-(length(yyy)/na)
  matriz[row,5]<-r[rr]
  matriz[row,6]<-s
  e<-e+1
  row<-row+1
}}
```

B Estrutura dos Algoritmos Utilizados

A grande dificuldade, no primeiro momento de nossa pesquisa, esteve pautada em encontrar algum algoritmo de CAT para que pudéssemos inserir a covariável Tempo de Resposta, criando assim, outro algoritmo. Já existem, atualmente, pacotes no R para implementar Testes Adaptativos Informatizados. O mais completo e robusto é o "catSim". No entanto, ele não contempla a covariável Tempo de Resposta.

Continuando com a nossa pesquisa, encontramos o artigo Labarrere et al. (2011), em que os autores compararam a convergência do algoritmo proposto por eles, à medida que se alterava a precisão do estimador. E esse foi o início de nossos trabalhos com a programação.

Para contribuir com as futuras pesquisas nessa área, diponibilizou-se, no anexo desse trabalho, os algoritmos utilizados e, nesse capítulo, comentar-se-á as principais estruturas, parâmetros, variáveis e funções utilizadas neles.

B.1 Algoritmo do CAT sem a Covariável Tempo de Resposta

No anexo A.2, colocou-se o algoritmo (em linguagem R) na íntegra. É o algoritmo de simulação de Testes Adaptativos Informatizados sem a covariável Tempo de Resposta. Para se entender bem o programa, sugere-se que a explicação a seguir seja acompanhada pelo código que se encontra no anexo A.2.

- 1) Quantidade de alunos: Por meio da variável "na", define-se a quantidade de respondentes que serão submetidos aos testes.
- 2) Habilidades: Adotou-se que a habilidade dos mesmos, representada no algoritmo por "theta", segue uma distribuição $\theta \sim N(0, 1)$
- 3) Precisão: Realizaram-se 6 testes para cada respondente, utilizando como critério de parada a precisão do estimador, variando-a de 10% a 35%.
- 4) Simulação dos parâmetros dos itens: Simularam-se 500 itens, com os seguintes parâmetros $a_i \sim LOGNORM(0, 0.35)$, $b_i \sim N(0, 1)$ e $c_i \sim BETA(2, 5)$.
- 5) Matrizes importantes: Criou-se duas matrizes fundamentais, "mp" e "ma". Na primeira guardaram-se as probabilidades de acertos dos 500 itens pelos 100 respondentes, segundo o ML3 (equação, 1). Na segunda guardaram-se as respostas (0 para itens errados e 1 para itens acertados) dos 100 respondentes nos 500 itens.

- 6) Número de pontos de quadratura e função Gauher: Definiu-se a quantidade de pontos de quadratura por meio da variável "nn". Nesse momento do algoritmo, habilita-se a função Gauher, retirada de Gray (2001). Ela calcula a estimação da habilidade, com base no método de quadratura gaussiana. Para maiores detalhes, ver Andrade, Tavares e Valle (2000), a partir da página 59.

- 7) Modelo normal: Fixa-se a variável "d" em 1,7 para que o curva logística se assemelhe à Ogiva Normal.

- 8) Cálculo das probabilidades de acertos: Preencheu-se a matriz "mp", definida anteriormente, com as probabilidades de acertos de todos os respondentes (de 1 a "na") para todas as questões (de 1 a "ni") do banco, por meio do Modelo Logístico de 3 parâmetros, ML3 (equação, 1).

- 9) Matriz de acertos/erros: Preencheu-se a matriz "ma", definida anteriormente, com zeros e uns. A obtenção desses dados foi feita através da função "rbinom(1,1,mp[i,j])". Essa matriz será muito utilizada na simulação, pois ela informa se o aluno "j" acertou ou errou a questão "i".

- 10) Inicialização do teste: A variável "ninit" define a quantidade de questões que iniciarão o CAT antes de se fazer a primeira estimativa da habilidade do respondente. Elas são escolhidas aleatoriamente do banco, dentre as questões que possuem o parâmetro "b" entre -0,5 e 0,5. Essas questões são retiradas do banco e é feita a estimação inicial da habilidade.

- 11) Estimação inicial de theta: Com as respostas das 5 primeiras questões (variável "resp" do código), estimou-se a habilidade do candidato ("theta.est") e a precisão do estimador ("erro") com base no método de quadratura.

- 12) Critério de parada: O teste avança enquanto a precisão do estimador (variável "erro" do código) está superior ao critério de parada fixado (variável "pp", que, em nosso estudo, assume os valores 10%, 15%, 20%, 25%, 30% e 35% para cada um dos respondentes). Caso o teste não pare até 100 questões, o algoritmo também para o teste e a habilidade do candidato assume o valor da última iteração.

- 13) Informação de Fisher e escolha da próxima questão: Com a estimativa inicial da habilidade do respondente e excluindo-se as questões utilizadas até então, calculam-se as medidas de Informação de Fisher para todas as demais questões do banco, escolhendo como próxima questão aquela que tem a maior Informação de Fisher. Isso é feito de maneira iterativa até atingir o critério de parada.

- 14) Estimação de theta: Uma vez atingido o critério de parada, a última estimativa obtida será a estimação considerada da habilidade, com sua respectiva precisão.

B.2 Algoritmo do CAT com a Covariável Tempo de Resposta

No anexo A.3, colocou-se o algoritmo (em linguagem R) na íntegra. É o algoritmo de simulação de Testes Adaptativos Informatizados com a covariável Tempo de Resposta. Para se entender bem o programa, sugere-se que a explicação a seguir seja acopanhada pelo código que se encontra no anexo A.3. Boa parte do programa é idêntico ao já mostrado anteriormente. Portanto, comentar-se-á os novos códigos.

- 1), 2), 3) e 4) Esses itens são idênticos aos mesmos itens do algoritmo anterior.
- 5) Matrizes importantes: Além das matrizes "mp" e "ma", estrutura-se também as matrizes "mt" e "mlam". Em "mt" guardaram-se os tempos dos itens acertados pelos 100 respondentes. Em "mlam", guardaram-se os parâmetros da função exponencial utilizada para simular os tempos.
- 6), 7), 8) e 9) Esses itens são idênticos aos mesmos itens do algoritmo anterior.
- 10) Índice das questões acertadas: Como a informação do Tempo de Resposta só será considerada para as questões em que o respondente acertou, precisou-se marcá-las com o índice "I".
- 11) Simulação dos tempos de resposta para as questões acertadas: De acordo com o estudo feito no capítulo anterior (Seção 4.2), a simulação dos tempos de respostas para as questões acertadas depende dos parâmetros "r" e "s". Consideraram-se, portanto, 7 pares (r, s) para simular os tempos de respostas. Estabeleceu-se também o tempo máximo de resposta a uma questão sendo 500, evitando assim algumas distorções na simulação dos tempos.
- 12) Inicialização do teste: Esse item é idêntico ao item 10 do algoritmo anterior, com uma ligeira alteração no final do código para habilitar os tempos de resposta das questões iniciais do teste, com as matrizes "t" e "lam".
- 13) Estimação inicial de theta: Com as respostas das 5 primeiras questões (matriz "resp" do código) e o Tempo de Resposta das questões acertadas (matriz "t" do código), estimou-se a habilidade do candidato ("theta.est") e a precisão do estimador ("erro") com base no método de quadratura. Vale a pena ressaltar que quando o respondente errava a questão, a estimativa da habilidade não levava em consideração o Tempo de Resposta, ou seja, o método de quadratura ficou idêntico ao do algoritmo anterior. No entanto, quando o respondente acertava a questão, o Tempo de Resposta foi levado em consideração, atualizando a função de verossimilhança "L(k)" com a variável "aaa".
- 14) Critério de parada: Esse item é idêntico ao item 12 do algoritmo anterior.

- 15) Informação de Fisher e escolha da próxima questão: Com a estimativa inicial da habilidade do respondente e excluindo-se as questões utilizadas até então, calculam-se as medidas de Informação de Fisher para todas as demais questões do banco, escolhendo como próxima questão aquela que tem a maior Informação de Fisher. Isso é feito de maneira iterativa até atingir o critério de parada.

- 16) Estimação de theta: Uma vez atingido o critério de parada, a última estimativa obtida será a estimação considerada da habilidade, com sua respectiva precisão.