Denise Reis Costa

Métodos Estatísticos em Testes Adaptativos Informatizados

Rio de Janeiro
2009

Denise Reis Costa

Métodos Estatísticos em Testes Adaptativos Informatizados

Dissertação apresentada ao Instituto de Matemática - Departamento de Métodos Estatísticos da Universidade Federal do Rio de Janeiro, UFRJ - para a obtenção do Título de Mestre em Estatística.

Orientador: Prof. Dr. Fernando A. S. Moura

Co-orientador: Prof. Dr. Dalton F. Andrade

Rio de Janeiro

2009

Métodos Estatísticos em Testes Adaptativos Informatizados

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Denise Reis Costa e aprovada pela banca examinadora.

Rio de Janeiro, 9 de Março de 2009.

Banca Examinandora:

- Prof. Fernando A. S. Moura (orientador) - DME/UFRJ
- Prof. Dalton F. Andrade (co-orientador) - INE-CTC/UFSC
- Prof. Dani Gamerman - DME/UFRJ
- Prof. Héliton Ribeiro Tavares - CCEN/UFPA

A todos os educadores e pesquisadores que acreditam na mudança do cenário da Educação no Brasil...

"Mudança não necessariamente implica progresso, mas o progresso é impossível sem mudança. A Educação, por sua vez, é essencial à mudança, uma vez que educação cria novos desejos e habilidades para satisfazê-los."

Henry Steele Commage

Agradecimentos

Antes do leitor apreciar as páginas desta dissertação, dedico algumas palavras de gratidão às pessoas que, desde o início, confiaram no meu esforço e dedicação, e que de variadas formas contribuíram para a realização deste trabalho. Para evitar cometer injustiças a quem sempre esteve (mesmo à distância) ao meu lado, mencionarei meus sinceros agradecimentos não por ordem de preferência, mas, para facilitar minhas recordações, de acordo com a região geográfica a qual foi possível o contato direto/indireto para o desenvolvimento destes capítulos.

Em Brasília/DF, agradeço à minha família e amigos que tanto deram apoio e acreditaram nos meus sonhos. Obrigada, queridos e amados pais, Matias e Marlene, pela fonte inesgotável de carinho e amor. Às minhas irmãs também agradeço por sempre desejarem meu sucesso tanto na vida pessoal quanto na acadêmica/profissional.

Ao Waldir, meu lindo namorado, pelo amor a mim dedicado. Amor, obrigada por estar sempre ao meu lado, ajudando-me em tudo que preciso e fazendo-me muito feliz. Sou muito grata por ter você!

Aos amigos e colegas do Cespe/UnB por todo suporte profissional. Agradeço, em especial, ao Prof. Joaquim José Soares Neto, atual diretor do Cespe/UnB, pela amizade e incentivo para que eu fizesse o mestrado.

Gostaria de agradecer também à equipe de Línguas desse Centro pela disponibilidade e atenção prestadas para o desenvolvimento desta dissertação.

No Rio de Janeiro/RJ, agradeço ao meu orientador Fernando A. S. Moura, pelo voto de confiança, pelo incentivo e por todo o conhecimento passado. Agradeço também a todos os meus professores da UFRJ pelo apoio e amizade.

Aos meus amigos da UFRJ pela amizade e, principalmente, por terem compartilhado momentos difícies e felizes ao longo desses dois anos.

Em Florianópolis/SC, agradeço ao meu co-orientador Dalton F. de Andrade pela atenção dispensada em todos os momentos da dissertação. Obrigada pelos ensinamentos e por toda confiança no meu trabalho.

Em São Paulo/SP, agradeço ao Caio pelas boas discussões sobre a TRI. Obrigada Caio, pela torcida sincera.

E por fim e não menos importante, agradeço a Deus pelo dom da vida, por me dar forças a cada dia para lutar pelos meus sonhos.

A todos, meu sincero MUITO OBRIGADA por serem exatamente o que preciso no momento em que mais preciso!

Resumo

Um Teste Adaptativo Informatizado (CAT) é aquele administrado pelo computador, que procura ajustar as questões do teste ao nível de habilidade de cada examinando. Basicamente, existem dois principais elementos para implementação do CAT: construir um banco de itens e desenvolver um algoritmo para seleção adaptativa dos itens. Para construir o banco, uma das exigências do CAT é que as questões que o comporão possuam boa qualidade tanto do ponto de vista pedagógico como psicométrico. Na avaliação psicométrica dos itens, a Teoria de Resposta ao Item (TRI) surge como um bom suporte estatístico, pois oferece uma análise quantitativa de certas características dos itens, tais como dificuldade e discriminação. Em relação ao desenvolvimento do algoritmo, faz-se necessário avaliar medidas estatísticas para uma administração adequada dos itens no teste. Nesse algoritmo, as proficiências dos indivíduos devem ser estimadas iterativamente de forma que só serão selecionados os itens que mensurem eficientemente as proficiências dos examinandos. Com isso em mente, a presente dissertação se propôs a apresentar alguns métodos estatísticos referentes a esses dois componentes do CAT: construção do banco de itens e construção do teste adaptativo. Adicionalmente, avaliou-se a adequabilidade do banco de itens da Proficiência em Inglês Instrumental I da Universidade de Brasília à implementação do CAT.

Palavras-chave: Teste Adaptativo Informatizado, Teoria de Resposta ao Item, seleção adaptativa.

Abstract

A Computerized Adaptive Test (CAT) is a computer-administered evaluation which tries to adjust the test questions to the examinee's skill level. Basically, CAT implementation requires two main steps: to build an item pool and development of an algorithm for adaptive item selection. In order to muster items into a pool, a CAT demands that every question has good quality, both on the pedagogical and the psychometrical sense. To psychometrically evaluate those items, the Item Response Theory (IRT) comes along as a good statistical support, for it offers quantitative analysis of certain item characteristics, such as difficulty and discrimination. With relation to algorithm development, it is necessary to check statistical measures in order to adequately administrate the test items. On this algorithm, individual proficiencies must be iteratively estimated so that only the items that most efficiently measure the examinee's proficiency are selected. With that in mind, this dissertation proposes to present some statistical methods which refer to those two components of CAT: item pool assembly and adaptive test construction. Moreover, the evaluation of the adequacy of the item pool of the Instrumental English test of the University of Brasilia for CAT implementation was done.

Keywords: Computerized Adaptive Testing, Item Response Theory, adaptive selection.

Sumário

Lista de Tabelas

Lista de Figuras

Glossário

CAT *Computerized Adaptive Test*

CCI Curva Característica do Item

EAP Estimador bayesiano da média a posteriori

KL Kullback-Leibler

MAP Estimador bayesiano da moda a posteriori

MI Máxima Informação de Fisher

MIE Máxima Informação Esperada

ML1P Modelo Logístico de um parâmetro

ML2P Modelo Logístico de dois parâmetros

ML3P Modelo Logístico de três parâmetros

MV Estimador de Máxima Verossimilhança

MVM Máxima Verossimilhança Marginal

PSD *Posterior Standard Deviation*

TCT Teoria Clássica dos Testes

TRI Teoria de Resposta ao Item

Capítulo1

Introdução

1.1 Tema de Estudo

Não se pode negar que nas últimas décadas o uso de computador tornou-se imprescindível ao cotidiano de muitas pessoas. Seja para a elaboração de um simples relatório ou para movimentar grandes quantidades de dinheiro entre contas bancárias, a disseminação em larga escala de computadores tornou o uso dessa tecnologia fundamental nos mais diversos setores de atividades. Na Educação, por exemplo, existe uma grande preocupação em investir novas tecnologias dentro das salas de aula. Para tanto, o governo tem criado programas de inclusão digital que reúne iniciativas como a de instalar computadores nas escolas, ter acesso à internet, capacitação de professores e oferta de conteúdos educacionais *online*.

Com a inserção de um ambiente informatizado nas escolas, o desenvolvimento de novas ferramentas de ensino-aprendizagem tornou-se propício. A criação de testes para serem aplicados pelo computador é um exemplo de iniciativas que estão obtendo êxito. Os testes por computador possuem diversas vantagens, como a criação de itens em formatos multimídia, permitindo, ainda, que o teste seja aplicado simultaneamente em diferentes examinandos de diversos lugares do país. As crescentes pesquisas para a implementação desses testes fizeram surgir na literatura o conceito de testes adaptativos informatizados.

1.2 Testes adaptativos informatizados

Um teste adaptativo informatizado, *Computerized Adaptive Test* (CAT) em inglês, é aquele administrado pelo computador, que procura encontrar um teste ótimo para cada examinando. Para atingir isso, a proficiência do indivíduo (também conhecido como o traço latente ou a habilidade do indivíduo naquela área de conhecimento) é estimada iterativamente durante a administração do teste e, assim, só são selecionados os itens que mensurem eficientemente a proficiência do examinado.

Como citado por Wainer (2000), "a noção básica de um teste adaptativo é imitar automaticamente o que um sábio examinador faria". Um teste adaptativo informatizado tem por finalidade administrar itens, de um banco de itens previamente calibrados, que correspondam ao nível de capacidade do examinando. Geralmente, esses itens são selecionados de acordo com o modelo da Teoria de Resposta ao Item (TRI), que é assumido para descrever o comportamento da resposta do indivíduo. Ao contrário dos testes papel-e-caneta, diferentes examinandos podem receber diferentes testes de tamanhos variados. Esse mesmo autor ainda destaca que o resultado traz uma medição mais precisa da proficiência, além da redução do tamanho do teste (geralmente em 50%).

As primeiras pesquisas sobre testes adaptativos computacionais foram realizadas na década de 70 por Lord (1971) e Owen (1975). Desde então, diversos testes adaptativos informatizados têm sido operacionalizados, tais como: o *Graduate Record Examination* (GRE), desenvolvido pela Educational Testing Service (ETS) em 1996; *Test of English as a Foreign Language* (TOEFL), também desenvolvido pela ETS; *Armed Services Vocational Aptitude Test Battery* (ASBAV), desenvolvido pelo Departamento de Defesa dos Estados Unidos para selecionar potenciais recrutas para o serviço militar.

Diferentemente dos testes papel-e-caneta, em que cada examinando, geralmente, responde aos mesmos itens, tipicamente na mesma ordem, os testes adaptativos informatizados administram adequadamente os itens que irão compor o teste para cada examinando. Como destacam Sands e Waters (1997), administrar itens fáceis para examinandos de alta habilidade é desgastante e, ainda, as respostas corretas a esses itens agregam pouca informação para a estimação da proficiência desses indivíduos. Além disso, o examinando pode ficar entediado com os itens do teste que não oferecem nenhum desafio a eles e podem responder sem maiores cuidados os demais itens, introduzindo uma medida adicional de erro na estimação da proficiência. Similarmente, a administração de itens difíceis

para indivíduos de baixa proficiência é desgastante e as respostas incorretas não ofereceram muita informação às estimativas. Diante de itens difíceis, os indivíduos de baixa proficiência estão mais propícios a se sentirem frustrados e acabam por responder aleatoriamente aos itens, incorporando erro adicional ao processo de estimação. Por outro lado, um instrumento como o CAT procura ajustar o teste a cada diferente examinando.

Para ilustrar o processo adaptativo desses testes, a Figura 1.1 apresenta um teste hipotético com cinco itens. Como a proficiência θ é uma variável latente[1], faz-se necessário definir uma métrica para θ. Suponha para esse exemplo que θ esteja definida numa escala cuja média é 0 e desvio igual a 1. Suponha também que no início do teste, não exista nenhuma informação sobre o nível de habilidade do examinando e que se assumiu um nível de proficiência igual a $\theta = 0$. Em seguida, um item de dificuldade neste ponto é escolhido e administrado. Suponha que o examinando respondeu corretamente ao primeiro item. Dessa forma, a estimativa da habilidade é atualizada (neste caso, é aumentada para $\theta = 1,0$) e um segundo item (mais difícil) é escolhido para ser administrado. Suponha agora que o sujeito selecionou a opção errada para esse segundo item, sugerindo que o item é "muito difícil" para seu nível de proficiência. Novamente, o computador atualiza a estimativa da habilidade (neste momento, a estimativa diminuiu para $\theta = 0,6$). Então, o próximo item que será administrado será menos difícil que o segundo, refletindo a última estimativa da proficiência do indivíduo. Supondo que o examinando também respondeu incorretamente a este item, a estimativa da habilidade nesse passo será atualizada e sofrerá um decréscimo ($\theta = 0,3$). O quarto item será escolhido de tal maneira que seja mais fácil que o terceiro item. Se o examinando responder corretamente a este item, a estimativa de sua habilidade aumentará ($\theta = 0,5$) e um item mais difícil para esse nível de habilidade será apresentado como o último item do teste adaptativo informatizado hipotético.

Os procedimentos de seleção e administração de itens no teste e atualização das estimativas das proficiências em cada fase são feitas iterativamente até que algum critério de parada seja satisfeito.

[1] Variável que não pode ser observada diretamente

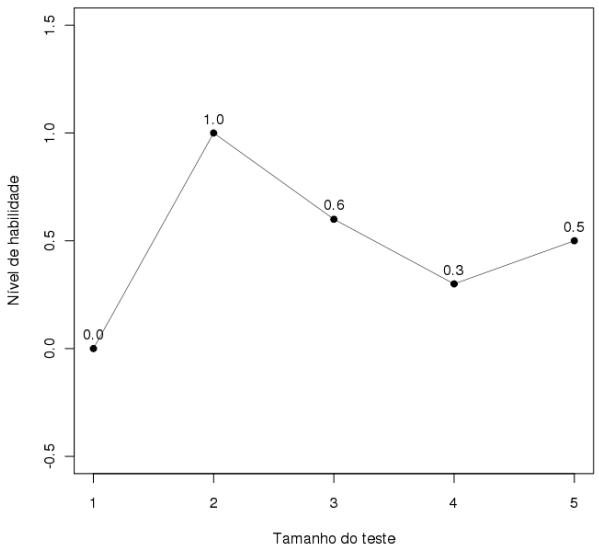


Figura 1.1: Teste adaptativo hipotético com cinco itens.

1.2.1 Vantagens do CAT

Administração e correção do teste

Como citado por Sands e Waters (1997), a versão CAT de um teste oferece diversas vantagens em relação a sua versão tradicional (papel-e-caneta). A redução do tamanho do teste é uma das vantagens. Como cada item apresentado a um indivíduo é adequado à sua específica habilidade, nenhum item administrado no teste é irrelevante. O número de itens de um teste adaptativo é substancialmente menor do que em um teste tradicional para um mesmo nível de precisão. Esta redução só é possível graças à informação sobre as estimativas da habilidade em cada item administrado. A Figura 1.2 exemplifica essa relação.

Tipo de teste	Habilidades dos examinandos	Ordem de dificuldade dos itens																			
		Fácil	----------																		Difícil
Papel-e-caneta	Todas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
CAT	Baixas	1	2	3	4	5	6	7	8	9	10										
	Médias						6	7	8	9	10	11	12	13	14	15					
	Altas											11	12	13	14	15	16	17	18	19	20

Figura 1.2: Comparação de um teste na forma papel-e-caneta e CAT.

No exemplo, todos os examinandos na versão papel-e-caneta respondem a todos os 20 itens do teste, independente de sua habilidade. Por outro lado, no CAT, um examinando de baixa habilidade responderá aos 10 itens relativamente mais fáceis; um examinando de média habilidade responderá aos 10 itens medianos do teste e um examinando de alta habilidade responderá aos 10 mais difíceis. Nessa situação hipotética, a metodologia do

CAT requer somente metade dos itens do teste clássico para produzir resultados com a mesma precisão.

Outra vantagem é a flexibilidade para realizar baterias de testes. Ao contrário do exame tradicional, um teste adaptativo informatizado não requer que todos os examinandos façam a prova ao mesmo tempo. Em uma bateria de testes, por exemplo, o examinando que terminar a prova pode passar diretamente para a prova seguinte sem precisar aguardar os outros. Além disso, o administrador do teste pode fornecer as instruções do teste virtualmente.

Maior rigidez no controle das regras do teste é outra vantagem. Um exame feito pelo computador está menos sujeito à burla de regras como a do tempo-limite para a realização do teste.

Sands e Waters (1997) mencionam também que o CAT simplifica o processo de correção do teste. A correção de um teste papel-e-caneta é dispendiosa e demanda tempo. Por meio de rotinas computacionais, o CAT traz agilidade, objetividade e transparência ao processo. Aliás, os resultados de uma avaliação por meio do CAT podem ser publicados quase que imediatamente após sua realização.

Um sistema de avaliação por computador reduz os erros que podem ocorrer em processos de correção que utilizam *scanners* ópticos. Além disso, não há possibilidade de erros de transcrição como as que ocorrem em testes que são corrigidos à mão.

Precisão das estimativas

Uma das estratégias para desenvolvimento de testes em papel é apresentar uma maior proporção de itens de dificuldade mediana e poucos itens de alta e baixa dificuldade. Esse procedimento torna o teste mais adequado para avaliar indivíduos de habilidade média do que indivíduos situados nos extremos da distribuição de habilidade, que possuirão menor precisão de suas estimativas. Em contraste, testes adaptativos ajustam adequadamente o nível de dificuldade das questões aos examinandos, sem prejudicar a acurácia das estimativas.

Segurança do teste

O uso do CAT aumenta significativamente a segurança do teste. Se um banco de itens é suficientemente grande, um examinando que tenha acesso a ele terá pequena vantagem sobre os demais. E ainda, há a possibilidade de criptografar os dados, de forma que

somente o administrador do teste tenha a chave para decodificar as informações do banco.

Motivação e tecnologia

Os autores Sands e Waters (1997) citam estudos que mostram a preferência de alunos em fazer testes em computadores. Além disso, o CAT possibilita a utilização de itens com recursos multimídia, o que o torna mais atrativo do que os testes tradicionais.

O CAT oferece, ainda, a possibilidade de enriquecer o conjunto de informações registradas no teste, tal como o tempo de resposta a um item. Isso permite ao avaliador saber não só se o examinando acertou ou errou um item, mas quanto tempo ele dispensou em seu desenvolvimento.

1.2.2 Desvantagens do CAT

Embora possua boas vantagens sobre os testes papel-e-caneta, os adaptativos informatizados demandam consideráveis recursos financeiros e humanos para a sua organização. Em muitos casos, a implementação de técnicas complexas, mudanças econômicas e políticas são necessárias. Por exemplo, embora a segurança dos testes seja uma das vantagens do CAT, ela também pode-se tornar um de seus maiores problemas.

Bancos de itens devem ser continuamente atualizados para garantir a segurança dos testes. Itens que não atendem mais às especificações e objetivos da avaliação ou que foram utilizados constantemente em diferentes testes devem ser eliminados (definitiva ou temporariamente) do banco. Além disso, novos itens podem ser incluídos ao banco, o que aumenta o custo da implementação e operacionalização de um teste adaptativo informatizado. Embora a aplicação desses testes apresente esses tipos de problema, suas vantagens superam as desvantagens.

Apesar de ser uma tecnologia vastamente implementada no exterior, o CAT ainda se encontra em fase embrionária no Brasil. Pode-se destacar duas dissertações elaboradas por Oliveira (2002) e Gonçalves (2004) do Instituto de Ciências Matemáticas e de Computação, ICMC/USP. Embora direcionadas ao desenvolvimento de sistemas para montagem de CAT, ambas as dissertações salientaram a grande dificuldade inerente à teoria estatística nessa área.

Pretende-se, portanto, com esse trabalho apresentar os recentes métodos estatísticos sobre CAT com o intuito de esclarecer a teoria e impulsionar o desenvolvimento dessa técnica que certamente contribuirá para o progresso das avaliações brasileiras.

Dessa maneira, na Seção 1.3 será apresentada uma idéia geral sobre a Teoria de Resposta ao Item: o que é, quais são os principais modelos, como pode ser aplicada na implementação de testes adaptativos informatizados. Já na Seção 1.4 será apresentado o banco de dados que será analisado ao longo da dissertação. Por fim, a Seção 1.5 mostrará como a dissertação está organizada.

1.3 Teoria de Resposta ao Item

Teoria de Resposta ao Item (TRI) é uma família de modelos probabilísticos que descreve características dos indivíduos que não podem ser observadas diretamente (variáveis latentes), mas que podem ser inferidas a partir das respostas dadas aos itens de um teste.

Embora existam inúmeros modelos matemáticos que podem expressar a relação de probabilidade de acerto a um item e a proficiência medida pelo instrumento de avaliação, nesta dissertação será apresentado e discutido o modelo logístico unidimensional para dados dicotômicos de três parâmetros (ML3P). A probabilidade condicional de um indivíduo com traço latente θ_j responder corretamente o item i pelo modelo ML3P é dada por:

$$P_{ji}(\theta_j) = P(U_{ji} = 1|\theta_j) = c_i + \frac{(1 - c_i)}{1 + \exp\left[-Da_i\left(\theta_j - b_i\right)\right]}, \tag{1.1}$$

onde: $i = 1, 2, \ldots, I$ (itens); $j = 1, 2, \ldots, n$ (examinandos).

U_{ji} é a variável dicotômica que assume o valor 1, quando o examinando j responde corretamente o i-ésimo item, ou 0, caso contrário;

θ_j representa a proficiência (habilidade ou traço latente) do indivíduo j; $\theta_j \in (-\infty, \infty)$;

$P(U_{ji} = 1|\theta_j)$ é a probabilidade de um examinando j com proficiência θ_j responder corretamente o i-ésimo item;

a_i representa o parâmetro de discriminação (ou de inclinação) do i-ésimo item; $a_i \in [0, \infty)$;

b_i representa o parâmetro de dificuldade (ou de posição) do i-ésimo item; $b_i \in (-\infty, \infty)$;

c_i representa a probabilidade de acerto casual do i-ésimo item; $c_i \in [0, 1]$;

D representa o fator de escala. No modelo logístico, esse fator é constante e igual a 1;

Como destacam Andrade, Tavares e Valle (2000), o modelo 1.1 apresenta problema de *falta de identificabilidade*. Essa não-identificabilidade do modelo ocorre porque diferentes valores dos parâmetros dos itens e da proficiência produzem a mesma probabilidade de um indivíduo responder corretamente a um certo item. Ou seja, sejam $\theta_{kj}^* = \alpha\theta_{kj} + \delta$, $a_i^* = \frac{a_i}{\alpha}$, $b_i^* = \alpha b_i + \delta$ e $c_i^* = c_i$, onde α e δ são constantes reais e $\alpha > 0$, então:

$$\begin{aligned} P_{kji}(\theta_{kj}^*) &= c_i^* + \frac{(1-c_i^*)}{1+\exp\left[-Da_i^*\left(\theta_{kj}^* - b_i^*\right)\right]} \\ &= c_i + \frac{(1-c_i)}{1+\exp\left[-D\frac{a_i}{\alpha}\left(\alpha\theta_{kj} + \delta - (\alpha b_i + \delta)\right)\right]} = P_{kji}(\theta_{kj}). \end{aligned} \tag{1.2}$$

Destaca-se que essa não-identificabilidade está intimamente relacionada às características da população em estudo. Especificando uma medida de posição (a média, por exemplo) e uma medida de dispersão (desvio-padrão, por exemplo) para as proficiências esse problema da falta de identificabilidade do modelo pode ser eliminado. Dessa maneira, uma escala (unidade de medida) estará sendo definida para as proficiências e, conseqüentemente, os parâmetros dos itens serão estimados nessa mesma métrica.

A probabilidade condicional $P_{ji}(\theta)$ também pode ser especificada pelo modelo logístico de 1 ou 2 parâmetros (ML1P e ML2P, respectivamente). O modelo de 2 parâmetros é obtido quando se fixa $c = 0$ para todos os itens, enquanto o modelo de 1 parâmetro (modelo Rasch) pode ser obtido fixando-se $a = 1$ também para todos os itens. Para maiores detalhes sobre esses modelos, sugere-se a leitura de Andrade, Tavares e Valle (2000). Vale ainda dizer que nessa dissertação optou-se pelo modelo menos parcimonioso (ML3P) já que os três parâmetros desse modelo possuem fundamental interpretação para a construção de testes adaptativos e que serão descritos posteriormente.

Vale destacar que o modelo 1.1 teve sua forma explícita graças a outro modelo de resposta ao item (dois parâmetros) estabelecido por Lord (1952). A curva característica no modelo de Lord (1952) assume distribuição acumulada da Normal e pode ser descrita como:

$$P_{ji}(\theta_j) = P(U_{ji} = 1|\theta_j) = \int_{-\infty}^{a_i(\theta_j - b_i)} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \Phi[a_i(\theta_j - b_i)]. \tag{1.3}$$

em que: U_{ji}, θ_j, a_i e b_i são os mesmos do modelo logístico 1.1.

Como a função de distribuição logística pode ser representada por: $\Psi_L(z) = \frac{1}{1+e^{-z}}$, Birnbaum (1968) foi o primeiro a estabelecer relação entre as distribuições Logística e Normal o que tornou possível a modificação do modelo de Lord (1952) para o modelo logístico. Ele observou que, dado um fator de escala D igual a 1,702, a função de distribuição logística é uma boa aproximação para a função de distribuição Normal, ou

melhor:

$$|\Psi_L(1,702z) - \Phi(z)| < 0,01 \quad -\infty < z < \infty.$$

O modelo logístico é atualmente utilizado por ser matematicamente mais simples de se trabalhar, já que é uma função explícita dos parâmetros dos itens e não envolve a função de integração presente no modelo normal. Nesta dissertação será utilizado o modelo logístico de três parâmetros com o fator D igual a 1,702 para que a função logística forneça resultados semelhantes ao da função de distribuição acumulada da Normal. A seguir, serão descritas as características dos parâmetros desse modelo.

Para melhor descrever a representatividade dos parâmetros no modelo, a Figura 1.3 representa a curva característica de três itens, com seus respectivos parâmetros.

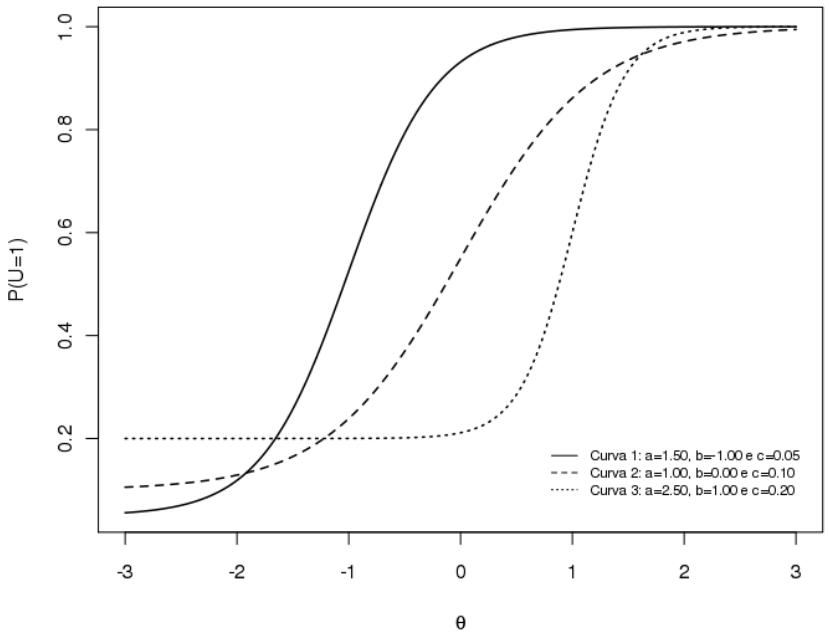


Figura 1.3: Curva característica de três itens.

A curva característica do item representa a relação existente entre a probabilidade $P(U_{ji} = 1|\theta)$ e os parâmetros dos itens. Observa-se que o modelo é acumulativo, ou seja, indivíduos com maior proficiência possuem maior probabilidade de acertar o item é que essa relação é não-linear.

Em relação às curvas, atribuiu-se maior parâmetro de dificuldade para a curva 3 ($b_3 = 1{,}00$). Dessa maneira, pode-se dizer que, entre os itens hipotéticos, o item 3 é o que exige maior habilidade (θ) para se obter uma alta probabilidade de acerto. Por outro lado, a curva 1 foi a que apresentou menor parâmetro b ($b_1 = -1{,}00$). Para esse item, examinandos de proficiências mais baixas possuem relativamente maior probabilidade de

acertar a esse item do que em relação ao item anterior.

Segundo o modelo 1.1, o ponto de maior inclinação da curva característica é quando $\theta = b_i$. O valor da inclinação no ponto b_i apresenta uma relação importante, pois é diretamente proporcional ao parâmetro de discriminação a_i. Quando se deriva a função 1.1 com relação a θ e se toma o máximo da função resultante, verifica-se que a inclinação máxima será igual a $0{,}425a_i(1 - c_i)$. Dessa maneira, pode-se dizer que quanto maior a inclinação da curva característica, maior é a discriminação do item (Gonçalves, 2006).

Por meio da Figura 1.3, verifica-se que atribuiu-se maior parâmetro de discriminação para o item representado pela curva 3. Isso significa que, próximo de $\theta = b_3$, a diferença entre as probabilidades de acerto de indivíduos com diferentes proficiências é maior que em relação aos outros itens do exemplo. Logo, o item 3 foi o item que melhor diferencia (ou discrimina) os examinandos entre os itens hipotéticos.

A representação gráfica do parâmetro c é expressa pela assíntota inferior da CCI. Ou seja, $\lim_{\theta \to -\infty} P(U_{ji} = 1|\theta, a_i, b_i) = c_i$.

Uma das maiores vantagem da TRI é que tanto os itens como os indivíduos são colocados em uma mesma métrica. Essa propriedade é muito importante no contexto de testes adaptativos, pois, embora cada estudante possa responder a diferentes itens de uma mesma prova no CAT, os resultados são comparáveis entre si. Isso se deve ao fato de que é possível a criação de uma escala de proficiência por meio da TRI. Um exemplo prático de uma escala de proficiência está no Sistema Nacional de Avaliação da Educação Básica (SAEB) (Andrade, Tavares e Valle, 2000). Essa avaliação é feita desde 1995 e tem como objetivo avaliar o desempenho dos estudantes da 4ª e 8ª séries do Ensino Fundamental e da 3ª série do Ensino Médio nas disciplinas de Língua Portuguesa e Matemática. Com a aplicação da TRI, foi possível criar uma escala única de conhecimento utilizando itens comuns entre essas provas, e, assim, tornando possível a formação de uma série histórica dos resultados da avaliação.

Nesta dissertação, será desenvolvida uma escala de proficiência para a prova de Proficiência em Inglês Instrumental I. A técnica que será utilizada para a construção dessa métrica envolve uma ferramenta estatística conhecida como *Ancoragem*. Fixados os níveis da escala (conhecidos como níveis-âncoras), essa técnica identifica itens que discriminam sucessivos pontos da escala utilizando itens com características próprias (itens-âncoras). É por meio da descoberta desses itens que se torna possível aos especialistas das áreas de conhecimento realizarem uma interpretação pedagógica dos níveis de habilidade da

escala.

A TRI está presente em todas as fases de um teste adaptativo informatizado. Desde a construção do banco de itens até o resultado final da avaliação. A primeira fase do CAT é caracterizada pela construção do banco de itens. Nessa fase, a TRI tem fundamental importância pois é a partir dela que será possível avaliar as características dos itens por meio da estimação dos parâmetros de cada item[2] (discriminação, dificuldade e acerto ao acaso). Na fase da aplicação do CAT, a TRI está intrisecamente ligada à estimação da proficiência do examinando. Os itens selecionados no CAT serão aqueles que produzem maior informação para a estimativa da proficiência e essa medida de "informação" do item só é possível graças a essa Teoria. Por último, o resultado fornecido ao examinando após a aplicação do CAT será interpretado na escala previamente construída pela TRI.

O estudo da aplicação do modelo da Teoria de Resposta ao Item no contexto dos testes adaptativos informatizados é a motivação desta dissertação. Deseja-se descrever todo o procedimento teórico da TRI em CAT para avaliação da estimação dos parâmetros dos itens e das proficiências. Este estudo também será estendido para a aplicação a dados de uma prova de Proficiência em Inglês.

1.4 Prova de Proficiência em Inglês Instrumental I

O banco de itens que será utilizado em toda a dissertação refere-se às questões da prova de proficiência em Inglês Instrumental I da Universidade de Brasília. Criada em 2004, essa avaliação é realizada duas vezes ao ano e tem como objetivo fundamental introduzir e levar o aluno a praticar as estratégias de compreensão leitora que favoreçam uma leitura mais eficiente de textos variados. Pode-se inscrever nesse exame qualquer aluno regular da Universidade de Brasília, sendo que os aprovados receberão os créditos para a disciplina Inglês Instrumental I.

Vale também dizer que, além da disciplina Inglês Instrumental I, os alunos dessa instituição também podem optar pela prova de proficiência em Língua Espanhola I. Contudo, os procedimentos descritos nesta dissertação delimitam-se à construção e avaliação do banco de itens de Proficiência em Inglês Instrumental I.

Os dados aqui analisados foram cedidos pelo Centro de Seleção e Promoção de Eventos (Cespe) que é a instituição responsável pelo exame. A prova de Proficiência em Inglês

[2]Esse processo também é conhecido como calibração dos itens.

Instrumental I contempla 50 questões objetivas de tipo múltipla escolha, com cinco alternativas (A, B, C, D e E).

Até o presente momento, conta-se com um histórico de 450 itens, aplicados a 2.969 alunos ao longo dos quatro anos em que esta avaliação está sendo feita. Em média, 330 estudantes realizam o exame a cada semestre.

Os alunos são submetidos a provas diferentes a cada período, mas com alguns itens comuns entre elas (geralmente oito) para a viabilidade da análise (equalização) pela Teoria de Resposta ao Item. Por meio da TRI, o cálculo da proficiência de cada aluno é medido numa escala de média 50 e desvio 16. O aluno aprovado é aquele que obteve proficiência igual ou superior a 50 nessa escala de proficiência.

Como essas provas apresentam itens em comum, o modelo da TRI para grupos múltiplos e respostas dicotômicas será aplicado. Após a calibração dos itens, pretende-se construir a escala de proficiência e ainda avaliar a eficiência dos métodos de seleção de itens em CAT para a estimação da proficiência utilizando esse banco de itens.

1.5 Organização da dissertação

O objetivo principal do presente trabalho consiste em apresentar os principais métodos estatísticos utilizados na montagem de um teste adaptativo informatizado. Os métodos aqui discutidos restringem-se a duas principais áreas do CAT: construção do banco de itens e construção do teste adaptativo.

Dessa maneira, esta dissertação foi divida em seis capítulos com a seguinte estrutura: no Capítulo 1 foi feita uma introdução aos conceitos de testes adaptativos informatizados e da Teoria de Resposta ao Item. No Capítulo 2 serão apresentados os métodos estatísticos relacionados à construção do banco de itens no CAT. Dar-se-á maior ênfase à abordagem de estimação dos parâmetros dos itens pela TRI. A calibração do banco de Proficiência em Inglês Instrumental I, bem como os passos para a construção da escala de proficiência, também serão apresentadas nesse capítulo.

No Capítulo 3, serão apresentados os métodos estatísticos relacionados à lógica dos testes adaptativos. O foco desse capítulo é de apresentar a estrutura de estimação da proficiência em CAT. Para isso, três métodos de seleção de itens que envolvem a TRI serão discutidos. O primeiro deles refere-se ao critério de Máxima Informação, um dos métodos mais populares para seleção de itens em CAT. O segundo método se baseia no

método da Informação Global definido por Chang e Ying (1996), utilizando a medida de Kullback-Leibler. Já o terceiro fundamenta-se na análise preditiva definida pelo critério de Máxima Informação Esperada proposto por van der Linden (1998).

O Capítulo 4 será reservado para a aplicação dos métodos de seleção adaptativa para o banco de Proficiência em Inglês Instrumental I.

Conclusões e sugestões pertinentes ao estudo serão ressaltadas no Capítulo 5.

Capítulo2

Construção do banco de itens

2.1 Introdução

O primeiro passo para a criação de um teste adaptativo informatizado consiste em organizar um banco de itens. Como os testes adaptativos se ajustam às capacidades de cada examinando, é possível criar testes individualizados. Para tanto, faz-se necessário que esse banco contenha uma ampla diversidade de itens. Para se aplicar uma prova adaptativa de Matemática, por exemplo, deve-se primeiramente criar itens de diferentes competências, tais como: Geometria, Álgebra e Trigonometria. É a partir dessa coleção de itens que será possível a aplicação do teste.

Um plano geral para o desenvolvimento de bancos de itens está ilustrado na Figura 2.1. Segundo Flaugher (2000), um banco de itens deve conter diversos elementos, dentre os quais pode-se destacar:

1. Criação de um número suficiente de itens para cada categoria de competências, baseando-se nas especificações do teste estabelecidas previamente;

2. Realização de revisões pedagógicas da qualidade dos itens. Observar, por exemplo, se os itens não apresentam funcionamento diferenciado (conhecidos na literatura por DIF) baseado em características específicas do examinado do que a habilidade mensurada pelo teste, tais como gênero ou etnia;

3. Pré-teste dos itens. Essa pré-testagem pode ser no formato papel-e-caneta e visa realizar uma análise psicométrica dos itens antes da aplicação do CAT. Objetiva-se com isso verificar o nível de dificuldade das questões, se os itens possuem bons

parâmetros de discriminação, entre outros quesitos.

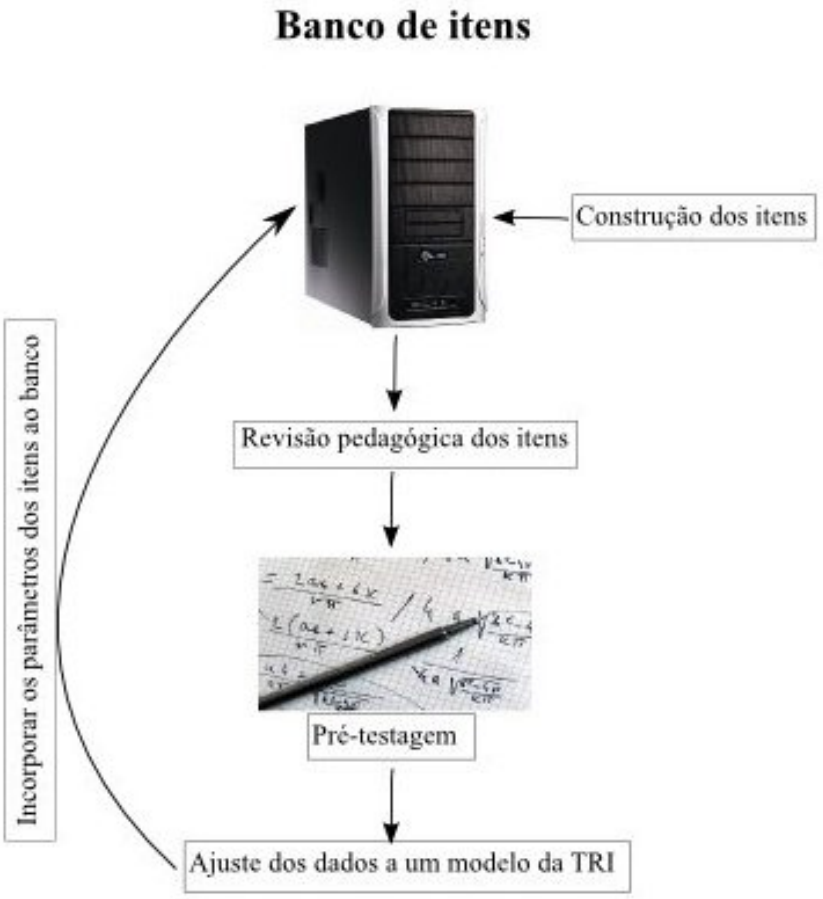


Figura 2.1: Construção do banco de itens.

A análise psicométrica dos itens citada no item 3 tem fundamental importância na construção de um banco de itens. Ela serve para avaliar a qualidade das questões. Um bom banco de itens deve conter uma variedade de itens com níveis de dificuldade bem diversificado. Flaugher (2000) registrou ainda que quanto melhor a qualidade do banco de itens, melhor será a tarefa que o algoritmo adaptativo poderá realizar. O melhor e mais sofisticado teste adaptativo não funcionará se esse estiver utilizando um banco de itens de baixa qualidade.

Duas principais teorias são empregadas para análise dos itens pré-testados, são elas: Teoria Clássica dos Testes (TCT) e Teoria de Resposta ao Item (TRI). A Teoria Clássica dos Testes preocupa-se em avaliar o indivíduo de acordo com os resultados obtidos em provas através de escores (brutos ou padronizados). Por exemplo, um indivíduo responde a um conjunto de questões, obtendo, no final, uma nota que é a soma das respostas corretas. Pela TCT, essa nota (denominada de escore) expressa a magnitude do que se desejava medir no sujeito, ou seja, indivíduos com maior escore são mais hábeis naquela

área de conhecimento do que aqueles com menor escore. Como a TCT apresenta algumas limitações, como por exemplo ser dependente do particular conjunto de itens que compõem a prova, a Teoria de Resposta ao Item (TRI) traz uma nova proposta estatística de análise, centrada nos itens e não na prova com um todo, e que não entra em contradição com os princípios da TCT (Pasquali, 2003).

A Teoria de Resposta ao Item, apresentada na Seção 1.1, parte da suposição de que existe no sujeito um traço latente (uma característica individual determinante de como responder aos itens de um teste: proficiência ou habilidade) que possui uma relação probabilística com cada um dos itens utilizados. Dessa maneira, a TRI modela a probabilidade de um indivíduo dar uma resposta correta a um item a partir das características dos itens e do traço latente desse indivíduo. Essa modelagem possibilita comparar os indivíduos entre si (avaliar quais são os respondentes mais - ou menos - hábeis na área de conhecimento avaliada), e da mesma forma, pode-se avaliar a qualidade das questões (verificar quais itens exigem maior - ou menor - nível de habilidade para se responder corretamente), já que os itens e os indivíduos são colocados em uma mesma métrica.

Na fase de construção do banco de itens, o principal interesse consiste na estimação dos parâmetros dos itens pela TRI. Dado um banco de itens pré-calibrado[1], a estimação das proficiências dos examinados em CAT depende essencialmente dos métodos de seleção dos itens que serão detalhados no Capítulo 3. Uma vez assumido como verdadeiro o modelo proposto, a partir das respostas dos examinados aos itens, faz-se possível a calibração dos itens. Neste capítulo será considerado o caso em que todos os itens do banco serão calibrados conjuntamente, ou seja, os itens podem ser considerados como itens "novos" (Andrade, Tavares e Valle, 2000).

Para as próximas seções, será considerado também que os respondentes dos itens do banco vieram de diferentes populações e realizaram provas parcialmente distintas. Assim, pode-se dizer que cada população em estudo foi submetida a diferentes provas, mas com alguns itens em comum. Esta estrutura de avaliação ilustra o maior avanço da TRI sobre a TCT e é um exemplo clássico da técnica de *Equalização via itens comuns*. Como citado por Andrade, Tavares e Valle (2000), "Equalização via itens comuns" significa colocar os parâmetros dos itens vindos de provas parcialmente distintas ou proficiências de respondentes de diferentes populações em uma escala comum, tornando os itens e/ou

[1]Um banco de itens pré-calibrado é aquele que possui estimativas para os parâmetros de seus itens. Utilizando a TRI, um teste adaptativo só funcionará se existir um banco de itens calibrado previamente.

as proficiências comparáveis. Essa metodologia possibilita a comparação e a construção de uma escala de proficiência interpretável. Uma vez estabelecida essa métrica, os valores dos parâmetros dos itens são invariantes a diferentes grupos de examinandos, desde que os indivíduos destes grupos tenham suas proficiências medidas na mesma escala.

Dessa maneira, na Seção 2.2 serão descritas as notações e definições referentes ao modelo unidimensional para grupos múltiplos; na Seção 2.3 será descrito o procedimento de estimação dos parâmetros dos itens e na Seção 2.4 serão apresentadas as estimativas dos parâmetros dos itens da prova de Proficiência em Inglês.

2.2 Notações e definições

Uma generalização da expressão 1.1 para a estimação dos parâmetros dos itens (ou das proficiências dos respondentes de diferentes populações que foram submetidos a diferentes provas com alguns itens em comum) foi proposta por Bock e Zimowski (1997). O modelo para grupos múltiplos apresentado pelos autores pode ser caracterizado da seguinte maneira:

$$P_{kji}(\theta_{kj}) = P(U_{kji} = 1|\theta_{kj}) = c_i + \frac{(1 - c_i)}{1 + \exp\left[-Da_i\left(\theta_{kj} - b_i\right)\right]}, \tag{2.1}$$

em que: $i = 1, 2, \ldots, I$ (itens); $j = 1, 2, \ldots, n_k$ (examinandos); $k = 1, \ldots, K$ (grupo).

U_{kji} é a variável dicotômica que assume o valor 1, quando o examinando j da população k responde corretamente o i-ésimo item; ou 0, caso contrário;

θ_{kj} representa a proficiência do indivíduo j da população k;

$P(U_{kji} = 1|\theta_{kj})$ é a probabilidade do examinando j da população k com proficiência θ_{jk} responder corretamente o i-ésimo item;

a_i, b_i, c_i **e** D possuem a mesma definição da Seção 1.3 da página 8;

No modelo de grupos múltiplos o problema da não-identificabilidade descrito na Seção 1.3 pode ser contornado padronizando-se as proficiências de cada grupo de forma que estas tenham distribuição com vetor de parâmetros $\eta_{\mathbf{k}} = (\mu_k, \sigma_k^2)'$ que representam, respectivamente, a média e a variância das proficiências da população k, $k = 1, \cdots, K$. Desta forma e na presença de várias populações, faz-se necessário estimar, além dos parâmetros dos itens, os chamados *Parâmetros Populacionais*: $\eta = (\eta_1, \eta_2, \cdots, \eta_K)$.

Para estabelecimento da métrica, os parâmetros populacionais de um dos grupos devem ser fixados, por exemplo, do primeiro grupo, e, portanto, resta apenas a estimação de $\eta_2, \eta_3, \cdots, \eta_k$.

Neste capítulo, as equações de estimação a serem desenvolvidas estão estritamente relacionadas ao conjunto de equações para estimação conjunta dos parâmetros dos itens no modelo de grupos múltiplos. Como o interesse deste capítulo consiste na construção do banco de itens, não serão aqui descritos os procedimentos de estimação dos parâmetros populacionais nem das proficiências dos examinandos. Recomenda-se a leitura de Andrade, Tavares e Valle (2000) a respeito dos métodos de estimação desses parâmetros.

A presença de diversas fontes independentes de variação que influenciam os atributos pessoais nas várias aplicações da TRI (áreas: médica, educacional, psicológica) justifica fortemente a hipótese de distribuição Normal dentro dos grupos. Por outro lado, pode-se pensar na atribuição de distribuições não-Normais para θ, como por exemplo, no caso em que as amostras delineadas para o estudo apresentam misturas de população com diferentes níveis médios de proficiência (Bock e Zimowski, 1997). Nesta parte da dissertação, a distribuição Normal(μ_k, σ_k^2) será adotada para a distribuição de θ_k. O grupo 1 será o de referência e seus parâmetros populacionais serão definidos como: $\eta_1 = (\mu_1 = 0, \sigma_1^2 = 1)$.

Embora se tenha K testes, vale destacar que os indivíduos pertencentes às diferentes populações não são submetidos aos mesmos itens mas, para permitir a comparação entre os grupos, alguns itens devem ser comuns a dois ou mais testes. Dessa forma, faz-se necessário fazer uma ordenação dos I distintos itens que compõem o conjunto dos K testes. O vetor de parâmetros de todos os itens pode ser definido como $\beta = (\beta_1, \cdots, \beta_I)$, onde $\beta_i = (a_i, b_i, c_i)$, onde a_i, b_i e c_i são os parâmetros do modelo ML3P definidos anteriormente. Denota-se I_k o conjunto dos índices dos itens do teste k, em que: $I \leq \sum_{k=1}^{K} I_k$.

Considere também as seguintes notações:

$\mathbf{U_{kj.}} = (U_{kj1}, U_{kj2}, \cdots, U_{kjI_k})$, o vetor aleatório de respostas do indivíduo j do grupo k aos I_k itens do teste k;

$\mathbf{U_{k..}} = (U_{1..}, U_{2..}, \ldots, U_{n_k..})$, o conjunto das respostas de todos os n_k indivíduos pertencentes ao grupo k;

$\mathbf{U_{...}} = (U_{1..}, U_{2..}, \ldots, U_{n..})$, o conjunto das respostas de todos os n indivíduos.

$\theta = (\theta_{11}, \theta_{12}, \cdots, \theta_{1n_1}, \cdots, \theta_{k1}, \cdots, \theta_{kn})$, o vetor de proficiências de todos os indivíduos.

De forma similar, pode-se representar as respostas observadas por $u_{kj.}, u_{k..}$ e $u_{...}$.

Ressalta-se ainda que os modelos de resposta ao item só podem ser considerados vantajosos quando o ajuste do modelo aos dados de interesse for satisfatório. Um modelo mal-ajustado não fornecerá estimativas invariantes para os itens e para as proficiências. Para tanto, faz-se necessário atender a duas principais suposições:

Independência entre indivíduos: as respostas advindas de diferentes examinandos são independentes.

Independência local: os itens são respondidos de forma independente por cada examinando, fixada sua proficiência.

A suposição de independência local é necessária para a validade do modelo, pois espera-se que, para cada valor de θ, a correlação entre as respostas do indivíduo dada a dois diferentes itens do teste seja nula (Andrade, Tavares e Valle, 2000). Ou melhor, para uma dada proficiência, as respostas de um indivíduo aos diferentes itens do teste devem ser independentes.

Como procurou-se direcionar este estudo para o modelo ML3P unidimensional, outra suposição necessária para o desenvolvimento do modelo consiste em verificar se o banco de itens avalia apenas um traço latente ou habilidade. Uma das técnicas mais utilizadas para avaliação da dimensionalidade das provas é a Análise Fatorial de Informação Plena (Nojosa, 2001).

Como a independência local afirma que, se houver correlação, esta se deve à influência de outros fatores que não o dominante (traço latente), quando a suposição unidimensionalidade é válida, tem-se que para o banco de itens o fator dominante é a única fonte de variação e, portanto, as respostas aos itens são independentes. Dessa maneira, se o banco for unidimensional, a condição de independência local do teste é válida, isto pois a unidimensionalidade implica em independência local do teste (Pasquali, 2003).

A seção subseqüente apresentará os principais passos para estimação dos parâmetros dos itens do banco para o modelo de grupos múltiplos.

2.3 Estimação dos parâmetros dos itens

Existe uma vasta literatura a respeito dos métodos para estimação dos parâmetros dos itens na TRI. Geralmente a estimação é feita pelo método de *Máxima Verossimilhança (MV)* ou por métodos bayesianos tais como: a *Moda a Posteriori (MAP)* ou a *Média a Posteriori (EAP)*. Nesta dissertação, será abordada a metodologia bayesiana para o modelo de grupos múltiplos da TRI.

Os métodos Bayesianos em TRI originaram-se com o propósito de contornar problemas surgidos pelos procedimentos de Máxima Verossimilhança, tais como: problemas de estimação dos parâmetros de itens respondidos corretamente (ou incorretamente) por todos os examinandos; problemas de estimação das proficiências de examinandos que responderam corretamente (ou incorretamente) a todos os itens; estimativas fora do esperado ($a_i < 0$ ou $c_i \notin [0, 1]$).

Como os métodos Bayesianos tentam aprimorar a acurácia das estimativas dos parâmetros dos modelos propostos e procuram contornar os problemas existentes nos MV, sua aplicação na TRI tem sido extremamente importante para os avanços dessa Teoria. Dessa forma, diversos programas computacionais têm implementado esses métodos de estimação, tais como o Bilog-MG (Zimowski et al., 1996). Entretanto, para se desfrutar ao máximo do potencial que o software pode oferecer, faz-se necessário conhecer os procedimentos básicos de estimação na TRI e isso será explorado nesta parte da dissertação.

A abordagem Bayesiana para estimação dos parâmetros dos itens na TRI baseia-se no Teorema de Bayes, que fornece uma forma de expressar probabilidade condicional. O objetivo é combinar probabilidades obtidas de uma função de Verossimilhança oriunda dos dados amostrais com probabilidades obtidas de uma informação *a priori* sobre a distribuição de um conjunto de parâmetros desconhecidos. Uma aplicação do Teorema de Bayes produz uma distribuição de probabilidade *a posteriori* que é proporcional ao produto da função de Verossimilhança e da distribuição de probabilidade *a priori*. A distribuição *a posteriori* é usada para se fazer inferências sobre os parâmetros desconhecidos.

Para ilustrar o uso do Teorema de Bayes na TRI, suponha que se deseje estimar os parâmetros dos itens quando as proficiências são conhecidas. Seja $\beta_i,\ i = 1, \ldots, I$ o conjunto de parâmetros dos itens (a_i, b_i, c_i) que é função do vetor de hiperparâmetros τ, com densidade $f(\beta|\tau)$ que reflete a crença *a priori* sobre a distribuição dos possíveis

valores dos parâmetros do item. Assume-se que os I itens têm a mesma distribuição *a priori*. Sejam $u_{...}$ uma matriz de respostas de todos examinandos aos I itens, sendo que cada população k responde aos n_k itens do teste, e $L(U;\theta,\beta)$ a função de Verossimilhança associada às respostas dos examinados e condicionada aos parâmetros θ e β. Pode-se representar a Verossimilhança da seguinte maneira:

$$L(u_{...};\theta,\beta) = \prod_{k=1}^{k}\prod_{j=1}^{n_k}\prod_{i=1}^{I} P_{kij}^{u_{kij}}(\theta_{kj})[1-P_{kij}(\theta_{kj})]^{1-u_{kij}}.$$

Assim, tem-se a seguinte expressão para a distribuição *a posteriori*:

$$g(\beta|u_{...},\theta,\tau) \propto L(u_{...};\theta,\beta)f(\beta|\tau).$$

Pode-se escolher qualquer característica da $g(\beta|u_{...},\theta,\tau)$ para se fazer inferência sobre os parâmetros dos itens, β, sendo que as mais adotadas são a *Média A Posteriori* (EAP) ou a *Moda A Posteriori* (MAP). Será aqui considerada a metodologia do MAP cujo objetivo é encontrar estimativas pontuais que maximizam a distribuição *a posteriori* com respeito aos parâmetros dos itens.

Como ocorre na prática, muita das vezes não se conhece as proficiências dos examinandos quando se deseja estimar os parâmetros associados aos itens do banco. Como as proficiências não são conhecidas, faz-se necessário utilizar algum artifício de forma que a função de Verossimilhança não seja mais função das proficiências. Para tanto, Bock e Aitkin (1981) desenvolveram um método em que ao se marginalizar a função de Verossimilhança (integrando-a em relação à distribuição das proficiências), esta não dependeria mais de θ. Esse artifício, denominado de Máxima Verossimilhança Marginal (MVM), é ainda hoje um dos métodos mais utilizados na prática devido sua facilidade computacional.

Mislevy (1986), por sua vez, estendeu a proposta de Bock e Aitkin (1981) para a estimação Bayesiana Marginal. A abordagem de Mislevy (1986) inclui as propriedades inerentes aos estimadores via MVM, porém restringe as estimativas dos parâmetros dos itens para que não se obtenha valores discrepantes ($a_i < 0$, por exemplo). Esse procedimento será explorado nas próximas seções.

Vale ainda acrescentar que os métodos de estimação dos parâmetros dos itens do banco que serão explicitados nesta dissertação aplicam-se a conjuntos de dados completos. Qualquer resposta nula (casos em que o examinando escolheu mais de uma alternativa) ou dados omissos (falta de resposta ao item pelo examinando) serão considerados como respostas incorretas.

2.3.1 Estimação Bayesiana Marginal

Mislevy (1986) empregou o modelo bayesiano de dois-estágios proposto por Lindley e Smith (1972) no qual a informação a priori é especificada de maneira hierárquica. Começa-se com a distribuição conjunta de todos os parâmetros antes da coleta dos dados. Assume-se que esses parâmetros são variáveis aleatórias contínuas e independentes com distribuição conjunta dada por:

$$g(\theta, \beta, \eta_k, \tau) = g(\theta|\eta)f(\beta|\tau)g(\eta)f(\tau) = \left[\prod_{k=1}^{K}\prod_{j=1}^{n_k} g(\theta_{kj}|\eta_k)g(\eta_k)\right]\left[\prod_{i=1}^{I} f(\beta_i|\tau)f(\tau)\right].$$

O termo $g(\theta|\eta)$ é a distribuição das proficiências do grupo k em função do vetor de parâmetros η. Como se assume que as proficiências são independentes e identicamente distribuídas, η_k tem média (μ_k) e variância (σ_k^2) comum à distribuição *priori* da população k. No modelo de Lindley e Smith (1972), θ é o parâmetro do modelo e os parâmetros populacionais μ_k e σ_k^2 são conhecidos como hiperparâmetros. Os hiperparâmetros podem ser considerados como variáveis aleatórias com distribuição de probabilidade denotada por $g(\eta)$.

Como já definido, a densidade $f(\beta|\tau)$ representa a função densidade para os parâmetros dos itens, condicional aos parâmetros populacionais no vetor τ. Como no caso das proficiências, os parâmetros a_i, b_i e c_i contidos no vetor β_i são os parâmetros do modelo e o vetor τ é o hiperparâmetro para o item com distribuição de probabilidade denotada por $f(\tau)$.

Após a observação das respostas dos examinandos, a distribuição *a posteriori* para todos os itens e examinados pode ser obtida através do Teorema de Bayes e expressa por:

$$g(\theta, \beta, \eta_k, \tau|u_{...}) \propto L(u_{...}; \theta, \beta)g(\theta|\eta)g(\eta)f(\beta|\tau)f(\tau). \tag{2.2}$$

Mislevy (1986) apontou que a equação 2.2 envolve a informação de todos os parâmetros do modelo. Entretanto, deseja-se fazer inferências com relação aos parâmetros dos itens. Logo, deve-se "marginalizar" a distribuição *a posteriori* dada em 2.2 para que esta seja função apenas do parâmetro de interesse: β. Esse procedimento foi apresentado primeiramente por Bock e Lieberman (1970).

Baker e Kim (2004) destacam que se o modelo de resposta ao item e a distribuição *a priori* das proficiências forem corretamente especificadas, as estimativas dos parâmetros (para testes de comprimento finito) serão aproximadamente próximas de seus valores verdadeiros à medida que o número de examinandos cresce. Em geral, a escolha das variáveis

que serão marginalizadas da função é especificada pela distinção de quais parâmetros são de interesse ao estudo e quais não são (parâmetros incidentais ou de perturbação). Mislevy (1986) sugeriu que, em muitas pesquisas educacionais, o vetor τ não é de interesse, podendo ser tratado também como parâmetro de perturbação.

Integrando sobre a distribuição de probabilidade das proficiências $g(\theta|\eta)$ com respeito a θ e os parâmetros populacionais dos itens $f(\tau)$ com respeito a τ, obtém-se a seguinte distribuição *a posteriori* marginalizada:

$$\begin{aligned} g(\beta,\eta|u_{\cdot\cdot\cdot}) &\propto \int\int L(u_{\cdot\cdot\cdot};\theta,\beta)f(\beta|\tau)g(\theta|\eta)f(\tau)g(\eta)d\theta d\tau \\ &\propto g(\eta)\left[\int L(u_{\cdot\cdot\cdot};\theta,\beta)g(\theta|\eta)d\theta\right]\left[\int f(\beta|\tau)f(\tau)d\tau\right] \\ &\propto g(\eta)L(\beta,\eta)f(\beta). \end{aligned} \tag{2.3}$$

Apesar de não ser necessário a presença do vetor η na estimação dos parâmetros dos itens para construção do banco, optou-se por deixá-lo na equação 2.3. Isto porque, embora se tenha eliminado a dependência das estimativas das proficiências de cada examinando individualmente na estimação dos parâmetros dos itens, a função de Verossimilhança marginal ainda é condicionada aos hiperparâmetros μ_k e σ_k^2 (garantindo, assim, que os parâmetros dos itens sejam estimados na mesma métrica das proficiências). Em relação ao vetor τ, embora este tenha sido integrado em 2.3, mesmo assim faz-se necessário sua especificação em $f(\beta)$ (Baker e Kim, 2004).

Sob a suposição de independência entre os itens, a estimação pode ser feita um item por vez. Dessa maneira, para estimar os parâmetros de um item, por meio da metodologia da *Moda a Posteriori* (MAP), deve-se derivar a equação 2.3 com respeito aos parâmetros dos itens e igualar o conjunto de equações a zero. Por conveniência, pode-se trabalhar com o logaritmo da distribuição posteriori marginal. Logo, o sistema de equações para estimação Bayesiana dos parâmetros dos itens é dado por:

$$\frac{\partial \log[g(\eta)]}{\partial \beta_i} + \frac{\partial \log[L(\beta,\eta)]}{\partial \beta_i} + \frac{\partial \log[f(\beta)]}{\partial \beta_i} = 0. \tag{2.4}$$

Como a distribuição $g(\eta)$ não contém os parâmetros dos itens, sua derivada com respeito a β_i é nula. Assim, pode-se eliminar $g(\eta)$ da equação 2.4, resultando no seguinte sistema de equações para se obter a estimativa MAP para os parâmetros dos itens:

$$\underbrace{\frac{\partial \log[L(\beta,\eta)]}{\partial \beta_i}}_{\text{Verossimilhança}} + \underbrace{\frac{\partial \log[f(\beta)]}{\partial \beta_i}}_{\text{Distribuição } a \text{ } priori} = 0. \tag{2.5}$$

As parcelas envolvidas na equação 2.5 merecem atenção especial. Enquanto a primeira refere-se à função de Verossimilhança, a segunda está associada às distribuições *a priori* dos parâmetros dos itens. Como cada componente dessa equação envolve elementos distintos, eles serão descritos separadamente. Dessa forma, a Subseção 2.3.1 contemplará os aspectos da função de Verossimilhança; a Subseção 2.3.1, os conceitos relacionados às distribuições *a priori* e a Subseção 2.3.1, os procedimentos associados à estimação via o Algoritmo EM.

Componente de Verossimilhança

Para se resolver o termo que envolve a Verossimilhança na equação 2.5, faz-se necessário descrever a metodologia da Máxima Verossimilhança Marginal (MVM). Seja:

$$L(\beta,\eta) = P(U_{...}|\beta,\eta) = \prod_{k=1}^{K}\prod_{j=1}^{n_k} P(U_{kj.}|\beta,\eta_k), \tag{2.6}$$

onde: $P(U_{kj.}|\beta,\eta_k) = \int_{\mathbb{R}} P(U_{kj.} = u_{kj.}|\theta,\beta)g(\theta|\eta_k)d\theta$ representa a probabilidade marginal do vetor de respostas $U_{kj.}$ com respeito aos parâmetros β e η_k.

Uma alternativa à apresentada pela Verossimilhança 2.6 está na análise dos *Padrões de Resposta* dos examinados (Andrade, Tavares e Valle, 2000). Quando o número de examinandos é grande com relação à quantidade de itens, pode haver vantagens computacionais em se trabalhar com o número de ocorrência dos diferentes padrões de resposta em vez da análise individual dos examinados. Com I_k itens no teste k, há 2^{I_k} possíveis respostas (padrões de resposta). Seja r_{kl} o número de ocorrências distintas do padrão de resposta l no grupo k e $s_k \leq min(n, 2^{I_k})$, onde s_k representa o número de padrões de resposta com $r_{kl} > 0$. Segue disso que: $\sum_{l=1}^{s_k} r_{kl} = n_k$.

Pela independência entre as respostas dos diferentes examinandos, tem-se que os dados seguem uma distribuição *Produto-Multinomial*, isto é:

$$L(\beta,\eta) = \prod_{k=1}^{K}\left\{\frac{n_k!}{\prod_{l=1}^{s_k} r_{kl}!}\prod_{l=1}^{s_k}[P(U_{kj.}|\beta,\eta_k)]^{r_{lk}}\right\}. \tag{2.7}$$

Após se tomar o logaritmo de $L(\beta,\eta)$, a primeira componente da equação 2.5 pode ser reescrita como:

$$\frac{\partial}{\partial\beta_i}\left[\sum_{k=1}^{K}\sum_{j=1}^{n_k} r_{kl}P(U_{kj.}|\beta,\eta_k)\right] = 0, \qquad i = 1,\ldots,I.$$

Com os desenvolvimentos descritos em Andrade, Tavares e Valle (2000), obtêm-se as

seguintes equações de estimações para os parâmetros a_i, b_i e c_i:

$$a_i : D(1-c_i)\sum_{k=1}^{K}\sum_{l=1}^{s_k} r_{kl}\int_{\mathbb{R}}\left[(u_{kli}-P_i)(\theta-b_i)W_i\right]g^*_{kl}(\theta)=0; \tag{2.8}$$

$$b_i : -Da_i(1-c_i)\sum_{k=1}^{K}\sum_{l=1}^{s_k} r_{kl}\int_{\mathbb{R}}\left[(u_{kli}-P_i)W_i\right]g^*_{kl}(\theta)=0; \tag{2.9}$$

$$c_i : \sum_{k=1}^{K}\sum_{l=1}^{s_k} r_{kl}\int_{\mathbb{R}}\left[(u_{kli}-P_i)\frac{W_i}{P^*_i}\right]g^*_{kl}(\theta)=0. \tag{2.10}$$

Em que: $W_i = \frac{P^*_i Q^*_i}{P_i Q_i}$; $P_i = c_i + \frac{(1-c_i)}{1+\exp\left[-Da_i\left(\theta_{kj}-b_i\right)\right]}$; $Q_i = 1-P_i$; $P^*_i = \frac{1}{1+\exp\left[-Da_i\left(\theta_{kj}-b_i\right)\right]}$; $Q^*_i = 1-P^*_i$ e $g^*_{kl}(\theta) = \frac{P(U_{kj.}|\beta,\theta)g(\theta|\eta_k)}{\int P(U_{kj.}|\beta,\theta)g(\theta|\eta_k)d\theta}$.

Por causa da marginalização, todas as três equações 2.8 a 2.10 envolvem uma integração com respeito às proficiências. Desta maneira, faz-se necessário encontrar alguma solução numérica para a integral. Embora existam muitos métodos de aproximação de integrais, neste trabalho será apresentado o procedimento *Hermite-Gauss*, usualmente denominado de *Método de Quadratura Gaussiana*, que é o método usado no software Bilog-MG (utilizado para estimação dos parâmetros dos itens de proficiência em Inglês Instrumental I).

Estimação em forma de quadratura

Se $g(\theta|\eta)$ é uma função contínua com integral finita, ela pode ser aproximada, para qualquer grau de precisão desejado, por uma distribuição discreta com um número finito de pontos (por exemplo, um histograma). O procedimento de quadratura torna o problema de se encontrar a soma da área sob uma curva contínua mais simples: encontra-se a soma das áreas de um número finito de retângulos que aproxima a área sob a curva (Baker e Kim, 2004).

Seja q o número de retângulos do histograma. Os pontos médios desses retângulos são conhecidos como pontos de quadraturas (ou nós) e podem ser denotados por X_t, $t = 1, \ldots, q$. Cada um desses pontos de quadratura tem um peso associado $A_t = A(X_t)$ expresso pelo valor da altura da função ($g(X_t|\eta)$) e o respectivo comprimento do intervalo Δ_t. Os valores de X_t e A_t são obtidos resolvendo-se um conjunto de equações que envolvem a função $g(\theta|\eta)$ e a quantidade de pontos de quadratura (Hildebrand, 1956).

Ressalta-se ainda que não há necessidade de que $g(\theta|\eta)$ tenha distribuição Normal, nem seja necessariamente contínua. Geralmente, essa distribuição pode ser definida empiricamente (Azevedo, 2003).

Substituindo as aproximações por pontos de quadratura nas equações 2.8 a 2.10, tem-se que:

$$a_i : D(1-c_i)\sum_{k=1}^{K}\sum_{l=1}^{s_k}\sum_{t=1}^{q} r_{kl}\left[(u_{kli}-P_{ti})(X_t-b_i)W_{ti}\right]g^*_{kl}(X_t)=0; \tag{2.11}$$

$$b_i : -Da_i(1-c_i)\sum_{k=1}^{K}\sum_{l=1}^{s_k}\sum_{t=1}^{q} r_{kl}\left[(u_{kli}-P_{ti})W_{ti}\right]g^*_{kl}(X_t)=0; \tag{2.12}$$

$$c_i : \sum_{k=1}^{K}\sum_{l=1}^{s_k}\sum_{t=1}^{q} r_{kl}\left[(u_{kli}-P_{ti})\frac{W_{ti}}{P^*_{ti}}\right]g^*_{kl}(X_t)=0. \tag{2.13}$$

com:

$$g^*_{kl}(X_t)=\frac{P(U_{kj.}|\beta,X_t)A_t}{\sum_{t=1}^{q}P(U_{kj.}|\beta,X_t)A_t}, \tag{2.14}$$

quando $\Delta_t = 1$ e $P(U_{kj.}|\beta,X_t)=\prod_{i=1}^{I}P_{ti}^{u_{kji}}Q_{ti}^{1-u_{kji}}$.

A expressão P_{ki} pode ser definida exatamente como P_i apenas substituindo θ por X_k, $Q_{ki}=1-P_{ki}$ e $W_i=\frac{P^*_{ki}Q^*_{ki}}{P_{ki}Q_{ki}}$.

Como as equações 2.11 a 2.13 não apresentam soluções explícitas, faz-se necessário a aplicação de algum método iterativo. Bock e Aitkin (1981) propuseram uma adaptação do algoritmo EM de Dempster, Laird e Rubin (1977) para se obter estimativas para os parâmetros dos itens. Essa metodologia envolve o uso de dados "artificiais" para os q pontos de quadraturas. Esses dados artificiais consistem em: o número esperado de indivíduos que respondem ao item i com habilidade em torno de X_t ($\bar{f}_{kli}$) e o número esperado daqueles que respondem corretamente ao i-ésimo item ($\bar{r}_{kli}$). Ou melhor, as quantidades $\bar{f}_{kli}$ e $\bar{r}_{kli}$ podem ser definidas como:

$$\bar{f}_{kli}=\sum_{l=1}^{s_k} r_{kl}g^*_{kl}(X_t) \qquad \text{e} \qquad \bar{r}_{kli}=\sum_{l=1}^{s_k} r_{kl}u_{kli}g^*_{kl}(X_t).$$

Reescrevendo as equações 2.11-2.13, utilizando os dados artificiais para solução da componente da Verossimilhança para estimação dos parâmetros dos itens, tem-se:

$$a_i : D(1-c_i)\sum_{k=1}^{K}\sum_{t=1}^{q}\left[\left(\bar{r}_{kli}-P_{ti}\bar{f}_{kli}\right)(X_t-b_i)W_{ti}\right]=0; \tag{2.15}$$

$$b_i : -Da_i(1-c_i)\sum_{k=1}^{K}\sum_{t=1}^{q}\left[\left(\bar{r}_{kli}-P_{ti}\bar{f}_{kli}\right)W_{ti}\right]=0; \tag{2.16}$$

$$c_i : \sum_{k=1}^{K}\sum_{t=1}^{q}\left[\left(\bar{r}_{kli}-P_{ti}\bar{f}_{kli}\right)\frac{W_{ti}}{P^*_{ti}}\right]=0. \tag{2.17}$$

Para complementar a solução das equações de estimação Bayesiana dos parâmetros dos itens, a componente relacionada às distribuições *a priori* (segundo termo da equação 2.5) precisa ser detalhada.

Componente de distribuição *a priori*

Uma das principais características da Estatística Bayesiana consiste na idéia de que os parâmetros possuem distribuições, ao contrário da análise freqüentista em que os parâmetros são constantes. Como o parâmetro a_i deve ser positivo, b_i pode assumir qualquer valor real e c_i deve estar no intervalo $[0, 1]$, assume-se distribuições para esses parâmetros levando em conta cada limitação. A escolha da distribuição *a priori* dos parâmetros dos itens é arbitrária, embora um número limitado de funções seja empregado na prática. Além disso, faz-se necessário especificar valores numéricos para os hiperparâmetros das distribuições o que, em geral, é realizado de maneira subjetiva (Baker e Kim, 2004). A seguir, serão consideradas as suposições mais freqüentes na literatura.

Distribuição *a priori* para a_i

As distribuições mais adotadas para o parâmetro a_i são: *Log-Normal* e *Qui-Quadrado*. A justifica 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). Nesta dissertação, será assumida a distribuição Log-Normal com parâmetro $\tau_a = (\mu_a, \sigma_a^2)$.

A densidade do parâmetro a_i, sob suposição de que esse pertence à distribuição Log-Normal, é expressa por:

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

Assim, a derivada do logaritmo da função associada ao parâmetro a será igual a:

$$\frac{\partial \log f(a_i|\mu_a, \sigma_a^2)}{\partial a_i} = -\frac{1}{a_i}\left[1 + \frac{(\ln a_i - \mu_a)}{\sigma_a^2}\right]. \tag{2.18}$$

Distribuição *a priori* para 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 proficiências, pode-se adotar a distribuição Normal com vetor de parâmetros $\tau_b = (\mu_b, \sigma_b^2)$.

Sabe-se que a densidade da Normal para b_i é expressa por:

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

Logo, a derivada do logaritmo da função associada ao parâmetro b será igual a:

$$\frac{\partial \log f(b_i|\mu_b, \sigma_b^2)}{\partial b_i} = -\frac{(b_i - \mu_b)}{\sigma_b^2}. \tag{2.19}$$

Distribuição *a priori* para c_i

Como c_i representa a probabilidade de acerto ao acaso, seu valor só pode pertencer ao intervalo $[0, 1]$. Dessa maneira, Swaminathan e Gifford (1986) sugeriram uma priori *Beta* para esse parâmetro. A função densidade da Beta com parâmetros $\tau_c = (\alpha_c - 1, \beta_c - 1)$ é dada por:

$$f(c_i|\alpha_c, \beta_c) = \frac{\Gamma(\alpha_c + \beta_c - 2)}{\Gamma(\alpha_c - 1)\Gamma(\beta_c - 1)} c_i^{\alpha_c - 2}(1 - c_i)^{\beta_c - 2}.$$

Logo, a derivada do logaritmo da função associada ao parâmetro c será igual a:

$$\frac{\partial \log f(c_i|\alpha_c, \beta_c)}{\partial c_i} = \frac{(\alpha_c - 2)}{c_i} - \frac{(\beta_c - 2)}{1 - c_i}. \tag{2.20}$$

Estimação Bayesiana via Algoritmo EM

Com as componentes derivadas das equações de Verossimilhança em forma de quadratura (equações 2.15 a 2.17) e das distribuições *a priori* dos parâmetros (equações 2.18 a 2.20), as equações de estimação da *Moda a Posteriori* são definidas por:

$$a_i : D(1 - c_i) \sum_{k=1}^{K} \sum_{t=1}^{q} \left[\left(\bar{r}_{kli} - P_{ti} \bar{f}_{kli} \right) (X_t - b_i) W_{ti} \right] - \frac{1}{a_i} \left[1 + \frac{(\ln a_i - \mu_a)}{\sigma_a^2} \right] = 0; \tag{2.21}$$

$$b_i : -Da_i(1 - c_i) \sum_{k=1}^{K} \sum_{t=1}^{q} \left[\left(\bar{r}_{kli} - P_{ti} \bar{f}_{kli} \right) W_{ti} \right] - \frac{(b_i - \mu_b)}{\sigma_b^2} = 0; \tag{2.22}$$

$$c_i : \sum_{k=1}^{K} \sum_{t=1}^{q} \left[\left(\bar{r}_{kli} - P_{ti} \bar{f}_{kli} \right) \frac{W_{ti}}{P_{ti}^*} \right] + \frac{(\alpha_c - 2)}{c_i} - \frac{(\beta_c - 2)}{1 - c_i} = 0. \tag{2.23}$$

Como os dados artificiais ($\bar{f}_{kli}$ e $\bar{r}_{kli}$) são desconhecidos, faz-se necessário estimar essas quantidades de uma maneira iterativa, juntamente com os parâmetros dos itens. Dessa maneira Bock e Aitkin (1981) propuseram uma adaptação do algoritmo EM de Dempster, Laird e Rubin (1977).

Algoritmo EM

O algoritmo EM aplicado à TRI consiste em um processo iterativo para se obter estimativas dos parâmetros dos itens na presença de variáveis não-observáveis (proficiências

dos examinandos). Para melhor entendimento sobre o algoritmo EM, esse procedimento será descrito brevemente.

Suponha que as proficiências estejam restritas a um conjunto de $t = 1, \ldots, q$ valores $\bar{\theta}_t$, com probabilidades $\pi_1, \ldots, \pi_q$. Essa suposição pode ser feita porque as aproximações de integrais são realizadas por métodos de quadratura e os valores $\bar{\theta}_t$ corresponderão aos pontos de quadraturas (anteriormente representados por X_t). Defina também $f_k = (f_{k1i}, \ldots, f_{kqi})'$ como o número de examinandos do grupo k de cada nível q de proficiência para o i-ésimo item e $r_k = (r_{k1i}, \ldots, r_{kqi})'$, o número de respostas corretas do grupo k para cada valor q de proficiência para o item i. Se as proficiências dos n_k examinandos da população k que responderam ao item i são selecionadas aleatoriamente da população, a probabilidade conjunta que os f_{kti} indivíduos tenham proficiência $\bar{\theta}_t$, segue uma distribuição *Produto-Multinomial* dada por:

$$P(F_i = f_i|\pi) \equiv P(f_i|\pi) = \prod_{k=1}^{K} \left\{ \frac{n_k!}{\prod_{t=1}^{q} f_{kti}!} \prod_{t=1}^{q} \pi_t^{f_{kti}} \right\}, \ i = 1, \ldots, I_k$$

Dados f_{kit} e $\bar{\theta}_t$, a probabilidade de se obter r_{kti} acertos ao item i dentre as f_{kti} respostas dos examinandos do grupo k com habilidade $\bar{\theta}_t$ será:

$$P(R_{kti} = r_{kti}|f_{kti}, \bar{\theta}_t) \equiv P(r_{ti}|f_{kti}, \bar{\theta}_t) = \prod_{k=1}^{K} \left\{ \binom{f_{kti}}{r_{kti}} P_{kti}^{r_{kti}} Q_{kti}^{f_{kti}-r_{kti}} \right\},$$

onde: $P_{kti} = c_i + \frac{(1-c_i)}{1+\exp\left[-Da_i\left(\bar{\theta}_{kt}-b_i\right)\right]}$ e $Q_{kti} = 1 - P_{kti}$.

A probabilidade conjunta dos vetores f e r, dados $\bar{\theta} = (\bar{\theta}_1, \ldots, \bar{\theta}_q)'$ e $\pi = \pi_1, \ldots, \pi_q$, é:

$$\begin{aligned} P(f, r|\bar{\theta}, \pi) &= P(r|f, \bar{\theta}, \pi)P(f|\bar{\theta}, \pi) = P(r|f, \bar{\theta})P(f|\pi) = \\ &= \prod_{k=1}^{K} \left\{ \left[\prod_{t=1}^{q} \prod_{i=1}^{I_k} \binom{f_{kti}}{r_{kti}} P_{kti}^{r_{kti}} Q_{kti}^{f_{kti}-r_{kti}} \right] \times \left[\frac{n_k!}{\prod_{t=1}^{q} f_{kti}!} \prod_{t=1}^{q} \pi_t^{f_{tki}} \right] \right\}. \end{aligned}$$

Pelo critério da fatoração, tem-se que (f, r) é uma estatística suficiente para os dados completos (U, θ) (Baker e Kim, 2004).

Ignorando alguns termos constantes em relação aos parâmetros dos itens, β, a função logaritmo da Verossimilhança para os dados completos pode ser escrita da seguinte forma:

$$\log L(\beta) \propto \sum_{k=1}^{K} \sum_{t=1}^{q} \sum_{i=1}^{I_k} \left[r_{kti} \log P_{kti} + (f_{kti} - r_{kti}) \log Q_{kti} \right].$$

Embora (f, r) sejam quantidades não-observáveis, pode-se estimar os parâmetros dos itens por meio da esperança do logaritmo da Verossimilhança condicionada a $u_{\cdot\cdot}$ e β e

das seguintes definições (Baker e Kim, 2004):

$$\bar{r}_{kti} = E[r_{kti}|u_{\cdots}, \beta] = \sum_{j=1}^{n_k} \left[\frac{P(U_{kj.}|\beta, X_t)A_t}{\sum_{t=1}^{q} P(U_{kj.}|\beta, X_t)A_t} \right]. \tag{2.24}$$

$$\bar{f}_{kti} = E[f_{kti}|u_{\cdots}, \beta] = \sum_{j=1}^{n_k} \left[u_{kji} \frac{P(U_{kj.}|\beta, X_t)A_t}{\sum_{t=1}^{q} P(U_{kj.}|\beta, X_t)A_t} \right]. \tag{2.25}$$

Assim, obtém-se:

$$E[\log L(\beta)] \propto \sum_{k=1}^{K} \sum_{t=1}^{q} \sum_{i=1}^{I_k} \left[\bar{r}_{kti} \log P_{kti} + (\bar{f}_{kti} - \bar{r}_{kti}) \log Q_{kti} \right]. \tag{2.26}$$

Maximizar a equação 2.26 é equivalente a resolução das equações de Máxima Verossimilhança Marginal para os parâmetros dos itens apresentadas em 2.15 a 2.17. Contudo, deseja-se estimar a moda da distribuição *a posteriori* marginalizada. Logo, basta acrescentar as quantidades referentes às distribuições *priori* dos parâmetros dos itens que estão apresentadas nas equações 2.21 a 2.23.

O primeiro passo do Algoritmo EM consiste no cálculo dos dados "artificiais" utilizando os valores *a priori* dos parâmetros dos itens. Esse passo é conhecido como Esperança (E). A próxima etapa consiste no cálculo dos parâmetros dos itens (passo de Maximização, M). Esses parâmetros são calculados um item por vez tendo por base os dados artificiais do passo E. Como as equações 2.21, 2.22 e 2.23 são não-lineares nos parâmetros, algum método iterativo como *Newton-Raphson* ou *Scoring de Fisher* deve ser usado no passo M. Maiores detalhes sobre os métodos iterativos recomenda-se Andrade, Tavares e Valle (2000) ou Azevedo (2003).

Pode-se resumir os procedimentos da estimação Bayesiana pela *Moda a posteriori* para a adaptação do algoritmo EM de Bock e Aitkin (1981) da seguinte maneira (Baker e Kim, 2004):

1. **Passo da Esperança (E):**

   (a) Utilize a forma da quadratura $\bar{\theta}_t$ e estimativas iniciais dos parâmetros dos itens para calcular a verossimilhança do vetor de respostas dos itens de cada examinando do grupo k para os q pontos de quadratura.

   (b) Utilize os pesos $A(\bar{\theta}_t)$ e os q pontos de quadraturas para o cálculo de $g^*_{kl}(X_t)$ dado pela expressão 2.14.

   (c) Utilize as equações 2.24 e 2.25 para gerar r_{kti} e f_{kti}, o número esperado de examinandos do grupo k submetidos ao item i e o número de respostas corretas (desse mesmo grupo) para i-item a cada ponto q de quadratura.

2. **Passo da Maximização (M):**

   (a) Resolva as equações 2.21 a 2.23 utilizando as quantidades r_{kti} e f_{kti}, que são os dados artificiais do passo E. Como esses valores dependem dos valores desconhecidos dos parâmetros dos itens, essas equações devem ser resolvidas iterativamente (por exemplo, métodos de Newton-Raphson/Fisher).

Os passos que compõem cada iteração do algoritmo EM são repetidos até que algum critério de parada seja alcançado.

Com base na metodologia descrita para estimação dos parâmetros dos itens pela TRI, uma aplicação aos dados da prova de Proficiência em Inglês Instrumental I será apresentada.

2.4 Proficiência em Inglês Instrumental I

O programa de aproveitamento de créditos em Inglês Instrumental I da Universidade de Brasília (UnB) foi criado em 2004 com o objetivo de aferir as habilidades de compreensão leitoras dos estudantes em textos variados. Essa prova é aplicada a cada semestre a estudantes dessa instituição que visam obter os créditos na disciplina de Inglês Instrumental I. A cada aplicação, 50 itens são apresentados aos examinandos, com cinco alternativas, e o resultado final do exame é apresentado aos alunos por meio da utilização das técnicas da Teoria de Resposta ao Item. Apesar de ser um teste convencional (papel-e-caneta), pretende-se avaliar a adequabilidade desse conjunto de itens para a implementação de um teste adaptativo informatizado.

Os dados dessa dissertação referem-se a nove aplicações da prova de Proficiência em Inglês Instrumental I realizadas entre o segundo semestre de 2004 até o segundo semestre de 2008. A Tabela 2.1 apresenta a quantidade de examinandos e de questões a cada período de aplicação.

Vale ressaltar que o código apresentado se refere a: identificação da prova de Inglês (ING), semestre e o ano em que os estudantes ganharam os créditos da disciplina. Desta maneira, o código ING105 refere-se a prova de Inglês em que os estudantes aproveitaram créditos nesta disciplina para o período do 1º semestre de 2005, enquanto o código ING205, prova de Inglês para o aproveitamento de crédito no período do 2º semestre de 2005.

Destaca-se ainda que o período de aproveitamento de créditos difere do período em

Tabela 2.1: Distribuição dos itens e examinandos nas provas

Prova	ING105	ING205	ING106	ING206	ING107	ING207	ING108	ING208	ING109	Total
ING105	50	-		-	-	-	-	-	-	50
ING205	8	42	-	-	-	-	-	-	-	50
ING106	4	4	42	-	-	-	-	-	-	50
ING206	2	2	4	42	-	-	-	-	-	50
ING107	2	-	-	6	42	-	-	-	-	50
ING207	8	-	-	-	-	42	-	-	-	50
ING108	-	-	-	-	-	8	42	-	-	50
ING208	1	-	-	-	-	8	-	41	-	50
ING109	1	-	-	-	-	6	-	2	41	50
Prova	**ING105**	**ING205**	**ING106**	**ING206**	**ING107**	**ING207**	**ING108**	**ING208**	**ING109**	**Total**
Examinandos	428	199	290	257	329	306	389	309	462	2969

que a prova foi aplicada, ou seja, alunos que ganharam os créditos no 1º semestre de 2005 fizeram a prova no período anterior (2º semestre de 2004).

Por meio do Tabela 2.1 pode-se observar como os itens estão distribuídos nas diferentes provas. Para que seja feita a equalização, os testes foram montados de maneira que existissem itens comuns entre as provas com a finalidade de que, ao final do processo, todos os itens que formarão a escala estejam numa mesma métrica. Dessa maneira, observou-se que, dos 450 itens aplicados, 384 é a quantidade total de itens diferentes entre as provas (valor este equivalente à soma da quantidade de itens na diagonal da Tabela 2.1) . Como os itens comuns podem ser aplicados a mais de duas avaliações, o número efetivo de itens comuns apresentados até a coleta desses dados foi de 37.

Vale destacar que, apesar de a TRI ser utilizada desde o primeiro ano de aplicação dessa prova de Proficiência, até o presente momento nenhum procedimento de estimação conjunta entre os itens dos diferentes períodos tinha sido feito. Embora se tenha adotado um delineamento para a montagem das provas contendo itens comuns entre elas, as análises e divulgação dos resultados pela TRI foram realizados separadamente em cada semestre.

Como este trabalho tem por objetivo a construção do banco de itens da prova de Proficiência em Inglês Instrumental I, todas as questões serão calibradas conjuntamente. Dessa maneira, faz-se possível também construir uma escala de proficiência interpretável e que, certamente, será de grande contribuição para a melhoria do programa de Proficiência da UnB.

Os parâmetros dos itens serão estimados por meio do Software Bilog-MG, criado pela *Scientific Software* (Zimowski et al., 1996). O programa Bilog-MG é um dos softwares

comerciais mais populares para análise de itens binários pela Teoria de Resposta ao Item. Esse programa permite analisar respondentes provenientes de várias populações e não possui número limitado de itens nem de respondentes.

O primeiro passo para a estimação conjunta dos itens consiste na identificação das questões comuns entre as provas. Analisar a freqüência em que cada item comum foi apresentado na versão papel-e-caneta também é importante. Para essa prova, verificou-se que dois itens do período ING105 foram apresentados seis vezes nas nove aplicações do exame. Dessa maneira, sugere-se que esses itens não sejam apresentados nas avaliações futuras pois itens muito expostos podem prejudicar a validade do teste, já que o aluno pode repetir a prova quantas vezes julgar conveniente.

Andrade (2001) destaca que quanto menor o número de itens comuns, maior é o erro cometido no processo de calibração dos itens. Portanto, faz-se necessário fazer uma ressalva em relação à quantidade de itens comuns nesse Exame. Como nesta dissertação optou-se pela equalização das diferentes populações durante a calibração (procedimento descrito na Seção 2.3.1), alguns autores sugerem que o número mínimo de itens comuns entre os testes deve ser de pelo menos 20%, equivalente a 10 itens para a prova de Proficiência em Inglês Instrumental I. Para assegurar um processo de calibração mais preciso, vale aqui registrar que um aumento na quantidade de itens comuns entre as provas de Proficiência em Inglês Instrumental I deve ser feito.

Além disso, deve-se avaliar também a qualidade desses itens. Como apontam Andrade, Tavares e Valle (2000), se os itens comuns utilizados na equalização tiverem níveis de dificuldade muito baixos ou altos em relação às populações em estudo, ou se apresentarem baixo poder de discriminação, haverá necessidade de um número maior de itens em cada teste.

Outro ponto importante a ser destacado envolve a quantidade de examinandos do teste. Andrade (2001) cita ainda que, para se obter estimativas com erros-padrão pequenos, cada item deve ser submetido a pelo menos 300 respondentes de diferentes níveis de proficiências. Por meio da Tabela 2.1, verifica-se que essa quantidade não é alcançada para três testes: ING205, ING106 e ING206. Para contornar tal situação, sugere-se uma nova testagem desses itens a novos indivíduos para que, assim, se obtenham estimativas mais precisas.

A estimação dos parâmetros dos itens pelo software Bilog-MG pode ser feita por dois métodos: Máxima Verossimilhança Marginal ou Bayesiana Marginal pela Moda a

Posteriori. Neste trabalho, foram especificadas as opções necessárias para que o programa executasse a estimação pelo método Bayesiano (MAP), descrita matematicamente nas seções anteriores.

Para solucionar o problema de falta de identificabilidade do modelo, assumiu-se que os respondentes representam uma amostra aleatória de uma população de proficiências com distribuição empírica, a ser estimada conjuntamente com os parâmetros dos itens pelo Bilog-MG (opção padrão do programa). Esta distribuição empírica é representada na forma de uma distribuição discreta, através dos pontos de quadraturas. O grupo de referência foi o ING105.

Sob o modelo Normal, as distribuições *a priori* associadas aos parâmetros dos itens foram: $a_i \sim LogNormal(\mu_a = 1, \sigma_a^2 = 2.72)$, $b_i \sim Normal(\mu_b = 0, \sigma_b^2 = 4)$ e $c_i \sim Beta(\alpha_c = 5, \beta_c = 17)$. Apesar de serem valores padrão do programa, optou-se por empregá-las também neste estudo por apresentarem valores para os parâmetros próximas ao esperado na prática (valor do parâmetro a situando-se entre 0,05 e 4,90; o parâmetro b entre -3,92 a 3,92 e um valor esperado para o parâmetro c de 0,20).

Além da estimação dos parâmetros dos itens, o Bilog-MG também estima os parâmetros populacionais e as proficiências dos examinados para os nove grupos. Contudo, para a construção do banco de itens faz-se necessário somente a análise das estimativas dos itens.

2.4.1 Análise dos itens

Primeiramente, foi feita uma análise de cada período em separado (ING105 a ING109). Nesta fase, foram analisadas as estatísticas clássicas e as estimativas dos parâmetros a, b e c dos itens de cada grupo separadamente. Em relação às estatísticas clássicas, foram observados o percentual de acerto ao item e a Correlação Bisserial que o Bilog-MG estima na fase 1.

Enquanto o percentual de acerto ao item representa o índice de dificuldade pela TCT, a Correlação Bisserial representa o índice de discriminação e pode ser calculada da seguinte forma:

$$\rho_{bis} = \frac{\bar{X}_A - \bar{X}_T}{s_T} \frac{pq}{h(Z_p)},$$

em que: $\bar{X}_A$ é o escore bruto médio dos alunos que acertam o item; $\bar{X}_T$, o escore bruto médio de todos os examinandos; s_T, o desvio-padrão dos escores brutos obtidos no testes; p, a proporção de indivíduos que acertaram o item no teste; $q = 1 - p$ e $h(Z_p)$, o valor

da função de densidade Normal Padrão em z_p (Pasquali, 2003).

Esses índices merecem atenção especial, pois são usados como valores iniciais para a estimação dos parâmetro da TRI pelo software Bilog-MG.

Em relação ao modelo da TRI, salienta-se que o fator de escala D definido no modelo 1.3 foi igual a 1,702, com o propósito de se aproximar a distribuição Logística da distribuição Normal (opção padrão do Bilog-MG).

Para o banco de Proficiência em Inglês Instrumental I, verificou-se que alguns itens apresentavam baixa discriminação (tanto na análise clássica quanto na análise da TRI). Dessa forma, algumas regras foram estabelecidas para a retirada de itens. Itens que apresentavam índice de correlação bisserial abaixo de 0,20 ou parâmetro a abaixo de 0,50 foram retirados da análise. Esses valores foram estabelecidos com base em Baker (2001) e por serem os mais utilizados na prática. Essa exclusão de itens deve ser feita pois espera-se que os itens do banco apresentem bons índices psicométricos para a aplicação do CAT. A Tabela 2.2 apresenta os critérios utilizados para retirada dos itens a cada rodada do programa Bilog-MG.

Tabela 2.2: Critérios de retirada de itens na análise individual

Compilação	Critério
1ª Rodada	Correlação Bisserial Negativa
2ª Rodada	Correlação Bisserial $<$ 0,10
3ª Rodada	Correlação Bisserial $<$ 0,20
4ª Rodada	Parâmetro $a <$ 0,40
5ª Rodada	Parâmetro $a <$ 0,50

Destaca-se que nenhum item da prova ING206 precisou ser retirado, dados os critérios estabelecidos pela análise individual. Por outro lado, foram necessários pelo menos três compilações da sintaxe do Bilog-MG para as outros testes. A cada rodada, os itens que se enquadraram nos critérios foram retirados da análise e não farão parte do banco de itens do CAT. A Tabela 2.3 apresenta a quantidade de itens retirados em cada prova.

Tabela 2.3: Quantidade de itens retirados da análise

Análise	**ING105**	**ING205**	**ING106**	**ING107**	**ING207**	**ING108**	**ING208**	**ING109**
Análise Clássica	8	5	3	9	7	5	3	6
Análise pela TRI	7	4	3	4	3	5	5	3
Total	15	9	6	13	10	10	8	9

Vale ressaltar que o critério de seleção dos itens baseou-se principalmente no índice de discriminação, visto que, ao se retirar os itens por esse critério, os outros parâmetros (dificuldade e acerto ao acaso) apresentavam estimativas e erro-padrão aceitáveis.

Após a análise individual, os itens foram calibrados conjuntamente, tomando-se o devido cuidado na definição dos itens comuns entre as provas e na retirada dos itens que não tiveram bom desempenho na análise individual.

Na análise conjunta, verificou-se que alguns itens apresentaram parâmetro a abaixo de 0,50. Da mesma forma que na análise individual, o programa foi rodado várias vezes até que os índices de discriminação dos itens pela TRI fosse maior ou igual a 0,49. É desejável que todos os itens do banco possuam parâmetro a de pelo menos 0,49 porque, segundo Baker (2001), valores acima desse limite possuem discriminações moderadas a altas que são bons índices psicométricos para um banco de itens.

Dos 384 itens, 36% foram retirados do banco de itens, seja pela análise individual da prova, seja pela análise conjunta. Dessa forma, o banco final que será utilizado para as análises de CAT nesta dissertação contém um total de 246 itens.

A Figura 2.2 apresenta como os itens estão distribuídos segundo as estimativas dos parâmetros de discriminação, dificuldade e acerto ao acaso pela TRI.

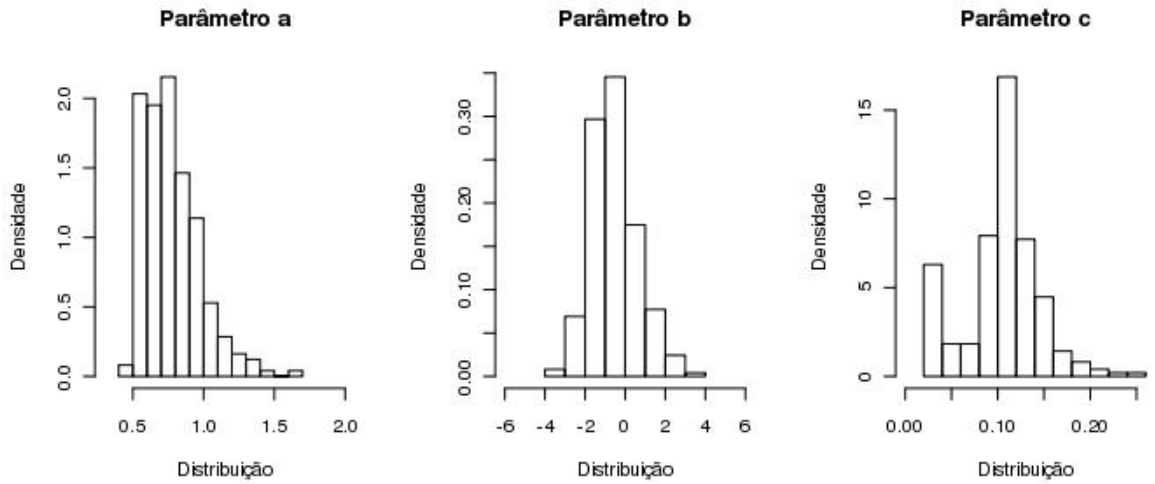


Figura 2.2: Histogramas com as estimativas dos parâmetros dos itens pela TRI.

Como era esperado, verifica-se uma distribuição assimétrica à direita para os índices de discriminação dos itens pela TRI. O valor médio do parâmetro a foi de 0,77 com desvio-padrão igual a 0,20, variando de 0,49 a 1,67. Em relação ao parâmetro b, observa-se que há itens de diferentes níveis de dificuldade ao longo da escala (variando de -3,56

a 3,23). Destaca-se ainda que 75% das estimativas do parâmetro b estão abaixo de 0,10. Sobre o parâmetro de acerto ao acaso (parâmetro c), o valor médio foi de 0,11, com desvio-padrão igual a 0,04 e amplitude de 0,03 a 0,24.

Pasquali (2003) menciona que o nível ideal de dificuldade para os itens de um teste depende da finalidade do teste. Se for desejado um teste, por exemplo, para se selecionar somente os 30% dos melhores candidatos, os níveis de dificuldade dos itens devem ser altos de modo que somente 30% (ou menos) dos examinandos tenham elevada probabilidade de acerto aos itens. O maior interesse, neste caso, consiste em discriminar entre sujeitos de alta aptidão, sendo sem interesse itens que apenas discriminariam examinandos de menor aptidão. Se, entretanto, o principal interesse for na avaliação da proficiência dos examinandos na área de conhecimento e não na seleção desses indivíduos (como é o caso do banco avaliado neste estudo), uma distribuição mais equilibrada dos itens em termos do parâmetro de dificuldade é requerida.

Flaugher (2000) destaca que um satisfatório banco de itens para a testagem adaptativa deve conter itens com alta discriminação (parâmetro a maior do que 0,60), uma distribuição retangular para o parâmetro de dificuldade e baixa probabilidade de acerto ao acaso (parâmetro c menor do que 0,20).

Segundo Segall, Moreno e Hetter (1997), avaliar a medida de precisão do banco de itens é importante no CAT para se julgar a adequabilidade dos itens. O impacto da seleção adaptativa dos itens e da estimação da proficiência na precisão das estimativas depende tanto da qualidade do banco de itens quanto dos procedimentos utilizados no teste adaptativo. As especificidades relativas à seleção de itens e estimação das proficiências serão abordadas no Capítulo 3. Nesta seção, a precisão dos itens do banco para o CAT será avaliada por meio da *Função de Informação do Item*.

A Função de Informação do item é calculada pela medida de *Informação de Fisher* a partir dos parâmetros dos itens. Ela permite analisar quanto um item contém de informação para a medida de habilidade (Andrade, Tavares e Valle, 2000). A Informação de Fisher do i-ésimo item no teste é representada por:

$$I_{F,i}(\theta) = \frac{\left[\frac{dP_i(\theta)}{d\theta}\right]^2}{P_i(\theta)[1-P_i(\theta)]} = D^2 a_k^2 \left[\frac{1-P_k(\theta)}{P_k(\theta)}\right]\left[\frac{P_k(\theta)-c_k}{1-c_k}\right]^2. \qquad (2.27)$$

Dado que a contribuição individual dos itens de um teste pode ser determinada sem conhecimento de outros itens, pode se obter uma medida informação para o teste. Essa medida consiste na soma de todas as funções de informação dos itens e denota-se por:

$I_F(\theta) = \sum_{k=1}^{n} I_{F,i}(\theta)$.

Pela teoria assintótica do estimador MV, o montante de informação fornecida pelo teste na estimação de θ é inversamente proporcional à precisão com a qual a medida de proficiência é estimada naquele ponto. Matematicamente, pode-se denotar essa precisão por: $EP(\hat{\theta}) = \frac{1}{\sqrt{I_F(\theta)}}$, onde $EP(\hat{\theta})$ é conhecido como *erro-padrão* da estimativa MV.

Hambleton, Swaminathan e Rogers (1991) indicam que a magnitude do erro-padrão depende, em geral, de: (a) o número de itens do teste (menor erro-padrão está associado a testes mais longos); (b) a qualidade dos itens que compõem o teste (em geral, menor erro-padrão está associado a itens de maior discriminação); (c) itens com dificuldade adequada à habilidade do examinando (menor erro-padrão está associado a testes que são compostos por itens com parâmetros de dificuldade aproximadamente iguais à habilidade do examinando).

Dessa maneira, as medidas de informação e do erro-padrão associado a um teste com 15 itens do banco de Proficiência em Inglês Instrumental I foram calculadas para diferentes valores de θ. Além do cálculo dessas medidas, procurou-se apresentar também uma listagem de itens, ordenados pelo itens de maior valor da informação para cada proficiência. A Tabela 2.4 mostra esses valores. Vale ainda destacar que a numeração dos itens foi codificada da seguinte maneira: o item 10833 é o item 33 da prova ING108, o item 10704 é o item 04 de ING107 e assim por diante.

Com os dados da Tabela 2.4, verifica-se que, para um teste com 15 itens, as proficiências -1,00 e -0,67 seriam as que melhor seriam estimadas por esse banco. Isto porque o valor da Informação para esse teste foi o maior entre os valores das outras proficiência e, conseqüentemente, os valores associados ao erro-padrão foi o menor encontrado (0,31). Como esperado, proficiências situadas nos extremos da distribuição possuem menor medida de Informação com os itens desse banco.

Se um teste adaptativo fosse feito com esse banco e considerasse como seleção de itens as questões que possuíssem maior valor da Função de Informação, poder-se-ia perceber que muitos itens seriam superexpostos. Por essa razão, faz-se necessário algum tipo de restrição ao algoritmo de seleção de itens. Essa abordagem será analisada detalhadamente no capítulo seguinte.

Uma vez que todos os parâmetros dos itens e todas as proficiências dos examinandos de todos os grupos avaliados estão numa mesma métrica, pode-se então construir uma escala de conhecimento interpretável para esse conjunto de dados.

Tabela 2.4: Informação e erro associado a diferentes proficiências num teste com 15 itens.

θ	Itens na ordem de preferência da esquerda para direita para cada θ															I(θ)	$EP(\theta)$
-4,00	10833	10704	20704	20835	10907	10810	10834	10707	10711	10821	20505	20708	20531	20750	20732	1,22	0,91
-3,67	10833	20704	10836	10704	20835	10810	10718	10522	10821	20626	20834	20708	20531	20732	20750	1,72	0,76
-3,33	20704	10907	10704	10810	20835	10718	10821	20834	10622	10926	20728	10617	20732	20505	20708	1,84	0,74
-3,00	10836	20704	20626	10907	10734	10810	10704	10702	10821	10519	20835	20611	10926	20728	20605	2,99	0,58
-2,67	20626	20704	10833	10707	10907	10734	20611	20834	10831	10821	10926	20605	10718	20728	20835	3,92	0,5
-2,33	20626	10834	10701	20611	10836	10614	10831	20704	20834	10907	20605	10734	10810	10711	20728	5,56	0,42
-2,00	20626	20611	10834	10831	20625	20605	20834	20705	20807	10937	20841	20728	20719	20704	10711	6,00	0,41
-1,67	20627	20626	20611	10624	20705	20807	20605	10839	20701	10937	10834	20634	20614	20834	10634	7,59	0,36
-1,33	20627	20625	20622	20705	20807	20701	20611	20626	20614	10937	20634	20840	20746	20612	20719	9,83	0,32
-1,00	20627	20625	10634	20701	20807	20705	10936	20840	20632	20746	20605	10637	20829	20636	10816	10,18	0,31
-0,67	20627	20622	10634	20625	20614	10632	10936	10719	20701	20840	10518	20634	20612	20630	20746	10,64	0,31
-0,33	10633	20627	20622	10719	10634	10625	20614	20519	10941	20625	20632	10816	20630	20612	20701	9,47	0,32
0,00	20618	10632	10625	20610	10941	10639	20622	20614	10816	20632	20630	10930	20832	20816	20633	8,85	0,34
0,33	20618	20813	10719	20519	10941	20816	10640	20545	10639	20832	20804	20633	20630	20527	20632	7,09	0,38
0,67	20618	20813	20610	20725	10632	10640	20635	20510	20816	20804	20511	20519	20544	20501	20545	7,35	0,37
1,00	20813	20725	20618	20610	10625	20635	10940	20804	20816	20510	20512	20642	20812	10636	20817	6,52	0,39
1,33	20725	20813	20642	20610	20544	10640	20635	10806	10635	20804	20618	20525	20812	20817	20831	5,71	0,42
1,67	20725	20642	20512	20511	20544	20813	20831	20610	10806	20529	20817	20812	20703	20804	20747	5,24	0,44
2,00	20725	20532	20831	20511	20736	20544	20529	10940	10822	20817	20522	20703	20812	20503	20813	4,29	0,48
2,33	20831	20725	20512	10527	20529	20544	20522	10722	10644	20503	20817	10640	20703	20812	20806	3,36	0,55
2,67	20736	20831	20512	20725	20529	10646	20511	20544	20522	20817	10940	20806	20703	10806	20812	3,08	0,57
3,00	20736	20532	20642	20831	20512	10646	20529	20725	20522	20615	20511	20503	20544	10527	20817	3,38	0,54
3,33	20736	20532	10835	20831	10822	20529	20512	10722	20615	20522	20503	20511	20806	20817	20725	2,84	0,59
3,67	20736	20532	20831	20642	10822	20529	10722	10644	20615	10541	20522	20512	20503	10909	20806	2,00	0,71
4,00	20736	20532	20831	20642	20529	20615	10644	10541	10516	20522	20503	20806	20817	20703	20525	1,28	0,88

2.4.2 Construção da escala de proficiência

Constrói-se uma escala de proficiência com a finalidade de se buscar uma interpretação qualitativa dos valores das proficiências obtidos. Na TRI, como os parâmetros dos itens vindos de provas distintas estão em uma escala comum, os itens e/ou as proficiência são comparáveis. Dessa forma, torna-se possível atribuir um significado pedagógico aos valores obtidos.

Após a calibração dos itens, o *software* utilizado apresenta tanto os parâmetros dos itens quanto as proficiências dos examinandos na escala $(0, 1)$, em que a média dos valores obtidos é igual a 0 e o desvio-padrão é 1. Como indica Valle (2001), muitas das vezes os valores das proficiências nessa escala são interpretados inadequadamente. Poderia existir questionamentos tais como: "Um examinando que tenha a proficiência igual a zero não possui nenhum tipo de conhecimento em Inglês Instrumental I?" ou "Como o examinando X pode possuir uma proficiência negativa?". Por praticidade e para evitar equívocos como esses, uma escala mais conveniente foi definida ao estudo.

A escala que tem sido utilizada pelo programa de Proficiência em Inglês Instrumental I da UnB tem média 50 e desvio 16. Dessa forma, sugere-se que os valores das proficiências estão localizados aproximadamente no intervalo 0 e 100. Contudo, verificou-se que existe bastante confusão na interpretação desses valores entre os examinandos bem como entre os especialistas dessa área de conhecimento. Isto ocorre porque os valores das proficiências são, muitas vezes, associados a "percentual de acerto", confusão bastante comum em escalas com esse tipo de variação.

A fim de contornar tais equívocos, uma nova definição de escala foi proposta. Essa escala tem média 100 e desvio 25. Uma vez definida a escala, uma transformação linear foi feita em todos os valores dos parâmetros da TRI. Assim, tem-se que:

- $\theta^* = 25 \times \theta + 100$;
- $a^* = a/25$;
- $b^* = 25 \times b + 100$;
- $c^* = c$.

A técnica empregada nesta dissertação foi a de *Ancoragem*. A escala construída por esse método é definida por *níveis-âncora*, que por sua vez são caracterizados por conjuntos de itens denominados de *itens-âncora*. Os níveis-âncora são os pontos da escala selecionados pelo analista na escala de proficiência para serem interpretados pedagogicamente. Depois de estabelecidos os níveis-âncora, faz-se necessário identificar os itens que caracterizam cada um desses níveis, ou seja, identificar os itens-âncora. Os itens-âncora são selecionados seguindo a definição subseqüente (Andrade, Tavares e Valle, 2000):

Definição 2.4.1 (Item-âncora) *Considere dois níveis-âncora da escala consecutivos Y e Z, com Y<Z. Um determinado item é âncora para o nível Z se e somente se obedecer simultaneamente às três condições abaixo:*

1. $P(U = 1|\theta = Z) \geq 0,65$;
2. $P(U = 1|\theta = Y) < 0,50$;
3. $P(U = 1|\theta = Z) - P(U = 1|\theta = Y) \geq 0,30$.

Em outras palavras, um item-âncora é aquele "típico" do nível, ou seja, bastante acertado por indivíduos com aquele nível de proficiência (respondido corretamente por um percentual de examinandos de pelo menos 65%), pouco acertado por examinandos com

nível de proficiência imediatamente inferior (respondido corretamente por um percentual de examinandos do nível anterior de no máximo 50%) e, além disso, a diferença entre a proporção de indivíduos com esses níveis de proficiência que acertam a esse item deve ser de pelo menos 30 pontos percentuais.

Uma das desvantagens dessa metodologia é que não há garantia de que existirão no teste aplicado itens-âncora para todos os níveis-âncora desejados. Para tanto, é fundamental que o número de itens aplicados seja bastante grande de modo a não comprometer a representatividade dos níveis, pois prejudicaria a construção e a interpretação da escala de proficiência.

Na prática, existe uma grande dificuldade de se encontrar itens que obedeçam a todas as características de um item-âncora. Isto se deve ao fato que nem todos os itens apresentam bons índices de discriminação. Uma boa alternativa é dividir os itens em categorias segundo o atendimento aos critérios da definição 2.4.1. Neste trabalho, os itens foram classificados por cores: azul, verde, vermelho e rosa. A divisão de cores procedeu-se da seguinte maneira:

Item azul é aquele mais "típicos" de um determinado nível de proficiência e que melhor descreve o nível-âncora. Esse é acertado por pelo menos 65% dos estudantes num nível de proficiência específico e acertado por menos de 50% dos estudantes com um nível de proficiência imediatamente anterior, além da diferença entre as probabilidades entre esses dois níveis-âncoras ser maior que 0,30. Este é o item que obedece aos três critérios.

Item verde é aquele que atende ao primeiro e ao segundo critérios: é acertado por pelo menos 65% dos indivíduos num nível de proficiência específico e acertado por menos de 50% dos indivíduos com um nível de proficiência imediatamente anterior, mas a diferença entre essas proporções é menor que 30 pontos percentuais.

Item vermelho só obedece ao primeiro critério: acertado por pelo menos 65% dos indivíduos num nível de proficiência específico e acertado por mais de 50% dos indivíduos com um nível de proficiência imediatamente anterior.

Item rosa também só obedece ao primeiro critério, porém poderia ser classificado no nível de proficiência imediatamente inferior já que possui uma alta probabilidade de acerto dos indivíduos desse nível (acima de 59%).

Salienta-se ainda que, apesar dos itens vermelho e rosa serem os piores itens quanto

ao atendimento aos critérios de itens-âncora, estes foram deixados na análise somente para auxiliar a interpretação dos níveis da escala de proficiência. A Figura 2.4 sintetiza essas relações.

Níveis da escala											
Item	0	25	50	75	100	125	150	175	200	225	Diferença
ING10833	0,41	0,73	0,92	0,98	1,00	1,00	1,00	1,00	1,00	1,00	0,32
ING10834	0,16	0,35	0,70	0,92	0,98	1,00	1,00	1,00	1,00	1,00	0,35
ING10701	0,18	0,35	0,67	0,90	0,98	1,00	1,00	1,00	1,00	1,00	0,32
ING20704	0,23	0,43	0,70	0,89	0,97	0,99	1,00	1,00	1,00	1,00	0,28
ING20835	0,29	0,46	0,67	0,84	0,93	0,97	0,99	1,00	1,00	1,00	0,21
ING10836	0,21	0,41	0,70	0,90	0,97	0,99	1,00	1,00	1,00	1,00	0,29
ING10704	0,35	0,54	0,74	0,88	0,95	0,98	0,99	1,00	1,00	1,00	0,20
ING10702	0,35	0,54	0,74	0,88	0,95	0,98	0,99	1,00	1,00	1,00	0,20
ING10711	0,16	0,29	0,54	0,78	0,92	0,97	0,99	1,00	1,00	1,00	0,25
ING10810	0,23	0,39	0,64	0,85	0,95	0,98	1,00	1,00	1,00	1,00	0,21
ING10718	0,25	0,41	0,63	0,83	0,93	0,98	0,99	1,00	1,00	1,00	0,19
ING10542	0,28	0,42	0,61	0,78	0,89	0,95	0,98	0,99	1,00	1,00	0,17

Figura 2.3: Probabilidades de acerto segundo os critérios definidos para um item-âncora.

O item 33 da prova ING108 (codificado como: ING10833), por exemplo, atendeu ao primeiro critério no nível 25 da escala, isto porque a probabilidade de acerto desse item só alcança a proporção maior que 65% de acertos dos estudantes no nível-âncora igual a 25 ($P(U = 1|\theta = 25) = 0,73$). Observa-se também que esse item atendeu ao segundo critério, pois obteve uma proporção de acerto no nível-âncora imediatamente anterior (nível 0) menor que 50% ($P(U = 1|\theta = 0) = 0,41$). A diferença entre a probabilidade de acerto no nível 0 e no nível 25 foi maior do que 30% ($P(U = 1|\theta = 25) - P(U = 1|\theta = 0) = 0,32$), o que o torna da cor azul. Ou seja, o item ING10833 é âncora para o nível de proficiência 25. Essa análise foi estendida a todos os outros itens do banco.

Após a identificação do conjunto de itens em cada nível da escala, o passo seguinte deve ser realizado por especialistas da área de conhecimento. Cabe a esses especialistas a interpretação dos níveis da escala, a partir do estudo do conteúdo abordado no conjunto de itens que definem cada nível-âncora. A Figura 2.4 apresenta uma parte da distribuição de itens nos níveis da escala montada para a prova de Proficiência em Inglês Instrumental I.

Uma oficina foi montada para a interpretação dos níveis da escala desse banco de itens. Diante dos resultados descritos anteriormente, cinco especialistas de cada área do conhecimento foram convidados pelo Cespe/UnB para interpretação dos níveis de proficiência. A oficina foi realizada entre os dias 19 a 22 de dezembro de 2008, nas

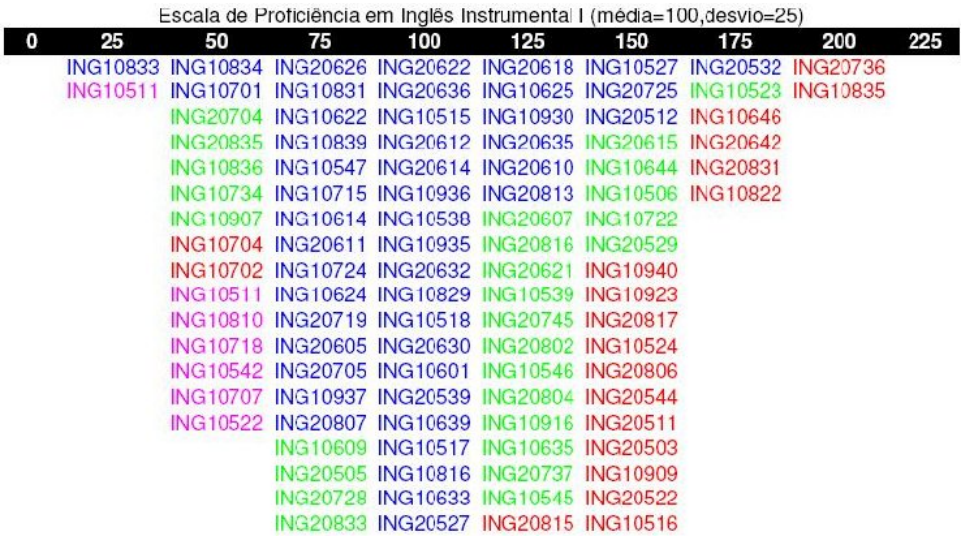


Figura 2.4: Parte da escala de proficiência da prova de Inglês Instrumental I.

instalações do Cespe/UnB. Primeiramente, uma palestra foi conduzida pela mestranda para esclarecimento dos objetivos da oficina e dos procedimentos estatísticos utilizados para a montagem da escala. Após a palestra, foram entregues aos especialistas folhas com os itens, índices de percentual de acerto pela TCT, estimativas dos parâmetros pela TRI e uma representação gráfica pela CCI. Esse material está anexo a essa dissertação.

Nesta fase de construção da escala, os especialistas realizaram somente a análise interpretativa dos itens para fornecer subsídios necessários à consolidação e à categorização da escala de Proficiência em Inglês Instrumental I. A interpretação feita pelo especialistas está sintetizada a seguir.

Interpretação dos níveis da escala

Nível 25 O aluno é capaz de fazer inferências e deduções básicas a partir de idéias contidas no texto, por meio de cognatos.

Nível 50 Além de compreender as habilidades do nível anterior, o aluno nesse nível entende vocabulário básico de textos acadêmicos, mas sofre limitações no que diz respeito ao vocabulário cotidiano. Além disso, ele reconhece idéias principais e conecta informações entre itens e texto. Ele também analisa paráfrases de informações textuais e estabelece relações entre as expressões contidas nos itens e no texto. O candidato faz inferências/deduções e identifica o tema do texto. Ainda, consegue reconhecer partes principais e secundárias do mesmo.

Nível 75 O aluno nesse nível é capaz de entender vocabulário básico/intermediário de

textos acadêmicos, mas ainda sofre limitações no que diz respeito ao vocabulário cotidiano. Ele reconhece palavras ou expressões sinônimas/antônimas e consegue relacioná-las e/ou substituí-las adequadamente no contexto. Além disso ele identifica estruturas gramaticais e localiza idéias principais e secundárias do texto. Há, também nesse nível, a diferenciação de fatos e opiniões e o estabelecimento de relação entre tese e argumentos oferecidos para sustentá-la. As inferências/deduções estão presentes e são mais apuradas. O candidato sabe reconhecer paráfrase.

Nível 100 O aluno nesse nível entende vocabulário intermediário de textos acadêmicos, mas sofre algumas limitações no que diz respeito ao vocabulário cotidiano. O aluno estabelece idéia de causa e conseqüência entre as partes e elementos do texto. Além disso, consegue definir sinonímia entre partes de palavras, como sufixos e/ou prefixos. Ainda, ele é capaz de selecionar a opção que melhor resume a(s) idéia(s) do texto e o título mais adequado dentre as alternativas dadas.

Nível 125 O aluno nesse nível entende vocabulário intermediário de textos acadêmicos e não sofre tantas limitações no que diz respeito ao vocabulário cotidiano. Ele está apto a reconhecer o efeito de sentido decorrente da exploração de recursos morfossintáticos. Ele faz correspondência gramatical entre trechos do texto.

Nível 150 O aluno nesse nível entende vocabulário intermediário/avançado de textos acadêmicos e não sofre limitações no que diz respeito ao vocabulário cotidiano. A correspondência gramatical é mais apurada e ainda as deduções/inferências têm um maior grau de complexidade.

Nível 175 O aluno nesse nível entende vocabulário avançado de textos acadêmicos e não sofre limitações no que diz respeito ao vocabulário cotidiano. A correspondência gramatical é apurada. Além disso, as deduções/inferências têm um grau de complexidade superior.

Nível 200 Além de todas as habilidades descritas nos níveis anteriores, o aluno nesse nível entende vocabulário avançado de textos acadêmicos e cotidianos.

Capítulo3

Construção do teste adaptativo

3.1 Introdução

A lógica para a construção de testes adaptativos está ilustrada na Figura 3.1

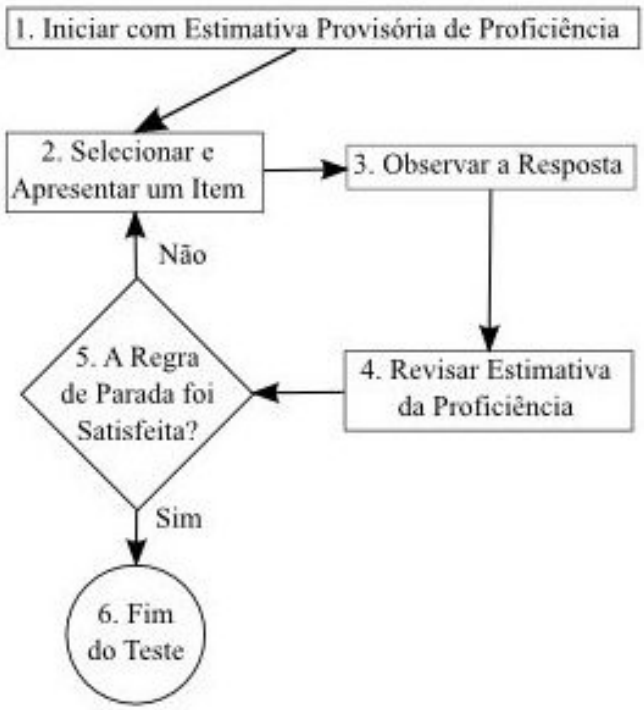


Figura 3.1: Base para a montagem do teste adaptativo.

De maneira geral, um teste adaptativo contempla a seguinte estrutura: um item é administrado e respondido. Com base nesta resposta, o algoritmo adaptativo irá (re)estimar a proficiência do examinando. Após o cálculo dessa estimativa, um novo item será selecionado. A incorporação da Teoria de Resposta ao Item à testagem adaptativa tornou

possível estimar as proficiências dos candidados com base na resposta a um único item ou a uma série de itens. O foco deste capítulo é apresentar toda a estrutura dos testes adaptativos para uma estimação eficiente das proficiências dos examinandos.

Na prática, a estimação das proficiências dos examinandos em um teste convencional (papel-e-caneta) é feita por meio de algum *software* comercial, tal como o Bilog-MG. Os programas no CAT, por sua vez, são desenvolvidos especificamente para um teste particular, utilizando ferramentas próprias, que na maioria das vezes não são aplicáveis a outros testes. A razão para essa prática é que, diferentemente dos testes tradicionais, a testagem adaptativa pode contemplar diferentes métodos para estimação das proficiências.

Como um teste adaptativo ajusta dinamicamente os itens que melhor avaliam uma dada proficiência, pode-se definir essencialmente três mecanismos para uma estimação eficiente dessa medida:

1. Adoção de um método adequado para estimação das proficiências;
2. Definição do critério de seleção dos itens e
3. Controle na exposição de itens.

Em relação ao primeiro fator, destaca-se que a Teoria de Resposta ao Item fornece grandes contribuições nessa área. Uma vez que diferentes examinandos podem receber diferentes conjuntos de itens em CAT (calibrados em uma mesma escala), por meio da TRI pode-se obter estimativas que são comparáveis entre si.

Birnbaum (1968) foi o primeiro a mostrar que estimativas de Máxima Verossimilhança para as proficiências dos examinandos podem ser obtidas de um conjunto arbitrário de itens na TRI. Ele apresentou as equações de Verossimilhança para os modelos ML2P e ML3P com probabilidades de acerto ao item em função da medida latente de proficiência e os parâmetros dos itens conhecidos. Ele também introduziu os conceitos de função de informação do item e da função de informação do teste comumente utilizados na seleção de itens em CAT.

Mais tarde, Samejima (1969) estendeu esses resultados para uma classe de modelos mais geral da TRI e propôs o estimador Bayesiano Modal (maximizando a densidade *a posteriori* em relação à proficiência, dado as respostas dos examinandos) que é um estimador alternativo ao de Máxima Verossimilhança. Ela não considerou a possibilidade do uso da média da distribuição *a posteriori* como estimativa da proficiência, pois nesta época ainda não estava claro como os valores esperados poderiam ser obtidos.

Bock e Aitkin (1981), por sua vez, apresentaram fórmulas baseadas em pontos de quadratura que tornam mais eficiente o cálculo numérico da média e da variância da distribuição da posteriori. Como na estimação dos parâmetros dos itens, eles propuseram o uso prático dessa fórmulas para a estimação das proficiências, introduzindo os termos MAP e EAP, já mencionados nesta dissertação. A Seção 3.2 apresentará as fórmulas matemáticas para a estimação proficiência por esses métodos.

Outro fator que precisa ser avaliado no desenvolvimento de testes adaptativos são os critérios de seleção dos itens. Geralmente, esses critérios baseiam-se em funções que dependem tanto dos parâmetros dos itens como das estimativas iterativas das proficiências. Por meio de alguma característica dessa função, faz-se possível a seleção iterativa dos itens no CAT. Na Seção 3.3 serão apresentadas três técnicas para uma seleção eficiente dos itens que comporão o CAT.

É comum no CAT assumir que os valores dos parâmetros dos itens estimados na construção do banco de itens foram calibrados com grande precisão a tal ponto que se pode tratar essas estimativas como valores verdadeiros para esses parâmetros (van der Linden e Pashley, 2003). Esses autores ainda destacam que ignorar erros nos valores das estimativas dos parâmetros dos itens é uma estratégia sem sérias conseqüências quando o tamanho da amostra na fase de calibração do banco é grande. Dessa maneira, os métodos de seleção dos itens e estimação da proficiência deste capítulo considerarão os parâmetros dos itens como conhecidos.

O último fator que pode influenciar a estimativa da proficiência são os métodos para controle da exposição de itens no CAT. Muitos programas operacionais de testes adaptativos encontram necessariamente uma base para seleção de itens não somente nos procedimentos estatísticos mas também impondo restrições no procedimento de seleção de itens para controlar certos atributos como balanceamento do conteúdo ou freqüência de exposição do item. Dada a importância desse estudo, a Seção 3.4 discutirá alguns pontos principais que precisam ser implementados no CAT para controle da exposição dos itens.

3.2 Métodos de estimação

Em um teste adaptativo informatizado, as proficiências dos examinandos são estimadas iterativamente após a resposta a cada item do teste. Para se obter uma medida para a proficiência dos indivíduos, existem na Teoria de Resposta ao Item diversos estimadores

para θ, sendo que os mais difundidos são o de Máxima Verossimilhança e os Bayesianos. Antes de definir esses estimadores, é necessário retomar a definição de Verossimilhança.

Considere um banco de itens com I itens. Seja denotado i o i-ésimo item do banco que será administrado no CAT. Por conveniência, um teste adaptativo de tamanho fixo n será assumido, embora os resultados apresentados também possam ser estendidos a um teste com critério de parada definido por um nível fixo de precisão na estimação da habilidade. Suponha que $k-1$ itens foram selecionados.

Como U_{ji} é variável dicotômica que assume o valor 1, quando o examinando j responde corretamente o item i, ou 0, caso contrário, considera-se que a distribuição de U_{ji} é a de Bernoulli de parâmetro $P_{ji}(\theta_j)$ dada pelo modelo ML3P. Assim, a função de Verossimilhança que relaciona a proficiência θ e as respostas do examinando j aos primeiros $k-1$ itens é dada por:

$$L(\theta_j; u_1, \ldots, u_{k-1}) = \prod_{i=1}^{k-1} P_{ji}(\theta_j)^{u_i} \left[1 - P_{ji}(\theta_j)\right]^{1-u_i}. \tag{3.1}$$

Vale ressaltar que a suposição de que as proficiências dos examinandos são estocasticamente independentes entre si é fundamental nesse processo de estimação. Dessa maneira, como examinandos diferentes não possuem informação de outros examinandos, pode-se estimar cada proficiência separadamente.

Os estimadores de Máxima Verossimilhança (MV) para a proficiência foram os primeiros a serem empregados no CAT. Basicamente, o procedimento para se obter estimativas por esse método consiste na maximização da função de Verossimilhança em 3.1 em relação a todos os possíveis valores de θ_j. Ou seja:

$$\hat{\theta}_{j_{u_1,\ldots,u_{k-1}}}^{MV} \equiv \arg\max_{\theta_j}\{L(\theta_j; u_1, \ldots, u_{k-1}) : \theta \in (-\infty, \infty)\}.$$

Como mencionado no Capítulo 2, o estimador de Máxima Verossimilhança apresenta algumas limitações como: para o modelo ML3P, um único máximo da função de Verossimilhança nem sempre existe; este também não está definido para alguns padrões de respostas (quando o examinando erra todos os itens ou acerta todos). Dessa maneira, os métodos Bayesianos apresentam-se como uma boa alternativa.

Na abordagem Bayesiana, uma distribuição *a priori* para o valores desconhecidos da proficiência, $g(\theta_j)$, é assumida. Combinando a função de Verossimilhança e a distribuição *a priori*, tem-se a distribuição *a posteriori* para θ_j:

$$g(\theta_j|u_1, \ldots, u_{k-1}) = \frac{L(\theta_j; u_1, \ldots, u_{k-1})g(\theta_j)}{\int L(\theta_j; u_1, \ldots, u_{k-1})g(\theta_j)d\theta}. \tag{3.2}$$

van der Linden e Pashley (2003) destacam que no CAT a densidade $g(\theta_j)$ é, tipicamente, considerada como Uniforme em um intervalo fechado ou, se os examinandos forem permutáveis, uma estimativa empírica da distribuição das proficiências na população dos examinandos pode ser considerada. Geralmente, modela-se a distribuição da população dos examinandos pela Normal. No entanto, esses autores também notam que para o modelo de resposta ML3P, uma *priori* Normal não fornecerá uma distribuição *a posteriori* Normal quando o teste for pequeno.

Os estimadores pontuais para θ_j pelo método Bayesiano baseia-se na distribuição $g(\theta_j|u_1,\ldots,u_{k-1})$ dada em 3.2. Freqüentemente, adota-se a moda (MAP) ou a média (EAP) como estimadores das proficiências.

No método MAP, deseja-se maximizar a distribuição a posteriori de θ_j:

$$\hat{\theta}_{j_{u_1,\ldots,u_{k-1}}}^{MAP} \equiv \arg\max_{\theta_j}\{g(\theta_j|u_1,\ldots,u_{k-1}) : \theta \in (-\infty,\infty)\}.$$

Diferentemente do estimador MV, o procedimento de estimação MAP sempre converge independentemente do padrão de resposta dos indivíduos. Contudo, esse método exige cálculos mais complexos do que o MV, já que a equação de estimação para θ é não-linear (Azevedo, 2003).

Uma alternativa aos métodos citados está na estimação pela Média a Posteriori (EAP). Esse estimador pode ser representado por:

$$\hat{\theta}_{j_{u_1,\ldots,u_{k-1}}}^{EAP} \equiv \int \theta_j g(\theta_j|u_1,\ldots,u_{k-1})d\theta_j. \tag{3.3}$$

Sob a abordagem dos pontos de quadratura apresentada na Seção 2.3.1, pode-se definir o estimador EAP de outra maneira:

$$\hat{\theta}_{j_{u_1,\ldots,u_{k-1}}}^{EAP} = \frac{\int\limits_{\mathbb{R}} \theta_j L(\theta_j;u_1,\ldots,u_{k-1})g(\theta_j)d\theta_j}{\int\limits_{\mathbb{R}} L(\theta_j;u_1,\ldots,u_{k-1})g(\theta_j)d\theta_j} \approx \frac{\sum\limits_{t=1}^{q} X_t L(X_t;u_1,\ldots,u_{k-1})A_t\Delta_t^{-1}}{\sum\limits_{t=1}^{q} L(X_t;u_1,\ldots,u_{k-1})A_t\Delta_t^{-1}}, \tag{3.4}$$

em que: X_t representa os pontos de quadraturas, $t = 1,\ldots,q$; A_t, o peso associado a X_t; Δ_t, o comprimento do intervalo.

No contexto da estimação pelo método EAP, os pesos são as probabilidades correspondentes aos pontos de uma distribuição *apriori* discretizada. Em certos casos, por exemplo, quando uma distribuição *apriori* Normal é assumida, os pontos e os pesos escolhidos podem ser os que possuem melhor precisão para a aproximação numérica da integral. A metodologia de Gauss-Hermite (descrita no Capítulo 2) fornece pontos de

quadratura e pesos para uma aproximação exata da integral de qualquer função, ponderada pela função erro Gaussiana, que pode ser expressa como uma função polinomial de q graus. Infelizmente, essa classe não inclui a função de Verossimilhança em testes adaptativos, tornando o uso dos pontos de quadratura Gaussiano não aplicáveis neste contexto (Bock e Mislevy, 1999). O que geralmente se faz em CAT consiste na determinação de pontos igualmente espaçados dentro de um intervalo (-3 a $+3$ desvios-padrão, por exemplo) que serão os pontos de quadraturas e os pesos serão iguais a probabilidade da *priori* discretizada nestes pontos.

Dados os pontos de quadratura, os cálculos para se obter estimativas pelo EAP não exigem nenhum processo iterativo como no caso do MAP ou MV, o que o torna bastante vantajoso do ponto de vista computacional. Da mesma maneira, pode-se obter a Variância a Posteriori (VAP) associada à estimativa EAP:

$$\begin{aligned} Var[\theta_j|u_1,\ldots,u_{k-1}] &= \frac{\int\limits_{\mathbb{R}}[\theta_j - \hat{\theta}_{j_{u_1,\ldots,u_{k-1}}}^{EAP}]^2 L(\theta_j|u_1,\ldots,u_{k-1})g(\theta_j)d\theta_j}{\int\limits_{\mathbb{R}} L(\theta_j|u_1,\ldots,u_{k-1})g(\theta_j)d\theta_j} \\ &\approx \frac{\sum\limits_{t=1}^{q}[X_t - \hat{\theta}_{j_{u_1,\ldots,u_{k-1}}}^{EAP}]^2 L(X_t;u_1,\ldots,u_{k-1})A_t\Delta_t^{-1}}{\sum\limits_{t=1}^{q} L(X_t;u_1,\ldots,u_{k-1})A_t\Delta_t^{-1}}. \end{aligned} \tag{3.5}$$

Em aplicações a populações reais, Bock e Mislevy (1999) mencionam que 80 pontos de quadratura entre ± 4 desvios-padrão devem ser usados para assegurar um nível de precisão menor que 0,20. O nível de precisão em EAP é dado pela raiz da VAP, conhecido como desvio-padrão *a posteriori* (*Posteriori Standard Deviation*, PSD). Em outras situações, por volta de 10 pontos de quadratura são necessários.

Uma propriedade interessante no CAT é que se um teste adaptativo apresentar itens aos examinandos até que um certo nível de precisão (PSD) for obtido, a confiabilidade das estimativas das proficiências será a mesma para todos os examinandos. Assumindo que a distribuição das proficiências dos examinandos assume uma distribuição $N(0,1)$, esse coeficiente de confiabilidade para a estimativa EAP é dado pela correlação $\rho = 1 - Var[\theta_j|u_1,\ldots,u_{k-1}]$ (Bock e Mislevy, 1999).

Dessa maneira, se as estimativas das proficiências forem utilizadas para seleção e classificação dos examinandos em diversos níveis da escala de proficiência, uma medida de precisão constante para todos os indivíduos assegura que os erros cometidos na classificação será uniforme para todas as decisões. Esse tipo de eqüidade nos procedimentos de seleção e classificação dos indivíduos não pode ser alcançado nos testes convencionais de

tamanho fixo.

Para uma distribuição *a priori* própria, a estimativa EAP sempre existirá, mesmo para aqueles padrões em que o MV não existe. Essa propriedade também é compartilhada pelo estimador MAP, contudo, o estimador EAP apresenta-se mais estável para todos os tamanhos de testes adaptativos, incluindo o primeiro item administrado (Bock e Mislevy, 1999). Dada essas vantagens, o método de estimação da proficiência utilizado ao longo dessa dissertação será o EAP.

3.3 Métodos de seleção adaptativa

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 proficiência) em busca de uma melhor escolha dos itens para a estimação das proficiências. Nesta parte da dissertação, três métodos serão avaliados.

3.3.1 Critério de Máxima Informação

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. Apesar de já se ter apresentado definições sobre a medida de Informação (Capítulo 2), nesta seção dar-se-á maiores detalhes a respeito desta função em CAT.

Conforme citado por Migon e Gamerman (1999), a noção de Informação está presente em todos os estudos desenvolvidos em Estatística. Como a incerteza é um dos principais ingredientes da modelagem estatística, quanto mais se tem informação a respeito da variável de interesse maior será a redução dessa incerteza. Para mensurar o grau de informação de uma variável, a Informação de Fisher é um dos conceitos mais conhecidos na literatura estatística. Define-se a Informação de Fisher da seguinte maneira:

Definição 3.3.1 (Informação de Fisher) *Seja* $\boldsymbol{U}$ *um vetor de variáveis aleatórias com*

função de probabilidade $P(\boldsymbol{u}|\theta)$*. A Informação de Fisher esperada de* θ *através de* $\boldsymbol{U}$ *é dada por:*

$$I_F(\theta) = E_{\boldsymbol{U}|\theta}\left[-\frac{\partial^2 \log P(u|\theta)}{\partial \theta^2}\right].$$

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

Na TRI, a Informação de Fisher permite analisar quanto um item contém de informação para a medida de habilidade e pode ser calculada para cada item individualmente a partir dos seus parâmetros. A Figura 3.2 mostra a curva de informação para cinco itens.

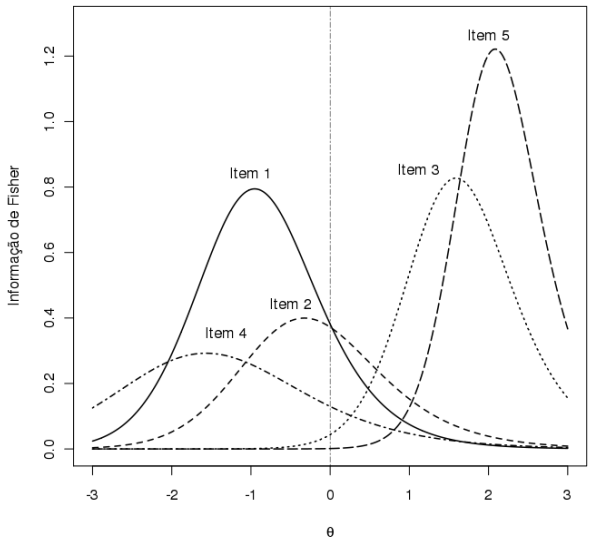


Figura 3.2: Função de Informação de cinco itens.

As curvas ilustradas em 3.2 sintetizam a idéia do MI. Seja $\hat{\theta}_{k-1} = 0,0$ o estimador de θ após $k-1$ respostas. O k-item selecionado por esse método será aquele que possuir a máxima Informação de Fisher avaliada em $\hat{\theta}_{k-1}$. Como já apresentado na Seção 2.4.1, para o Modelo Logístico de três Parâmetros (ML3P), a Informação de Fisher do k-ésimo item é dada por:

$$I_{F,k}(\theta) = \frac{\left[\frac{\partial P_k(\theta)}{\partial \theta}\right]^2}{P_k(\theta)[1-P_k(\theta)]} = D^2 a_k^2 \left[\frac{1-P_k(\theta)}{P_k(\theta)}\right]\left[\frac{P_k(\theta)-c_k}{1-c_k}\right]^2. \tag{3.6}$$

A Informação de Fisher é naturalmente relacionada à estimação da proficiência pela Máxima Verossimilhança (MV). Maximizar essa medida significa minimizar a variância assintótica de $\hat{\theta}_{k-1}$, o que leva a uma estimação de $\hat{\theta}_{k-1}$ mais eficiente. Em outras palavras,

a Informação de Fisher é inversamente proporcional ao erro-padrão do estimador MV. A seguir, será apresentada uma figura com as curvas de informação e do erro-padrão da estimativa de proficiência para um certo item.

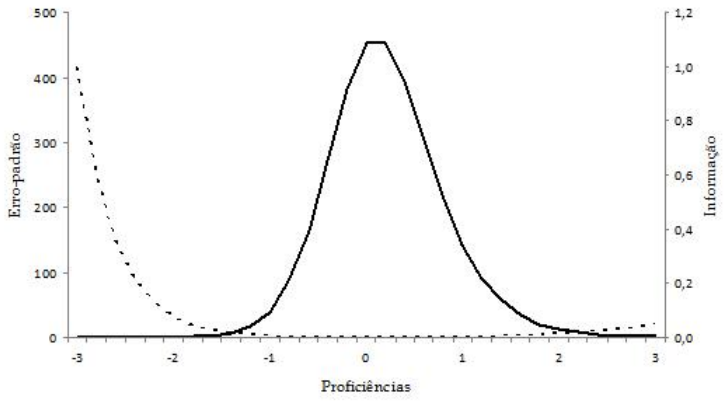


Figura 3.3: Curvas de informação e do erro-padrão.

Como ilustrado na Figura 3.3, a curva de informação do item tem formato de uma curva "Normal". Isto indica que o item é mais informativo em torno do ponto $\theta = 0,0$. Como já esperado, a medida de erro-padrão ao se estimar as proficiências em torno desse ponto é baixa. Fora desse intervalo, porém, o item não é capaz de estimar eficientemente as habilidades dos examinandos, isso se deve ao fato de que a medida de erro-padrão apresenta-se maior que a medida de informação.

Sob a TRI, maximizar a Informação de Fisher significa intuitivamente selecionar um item de dificuldade que corresponda exatamente ao nível de proficiência do examinando. Além do mais, itens com maior discriminação, ou equivalentemente, alto valor do parâmetro a, serão preferencialmente selecionado pelo algoritmo. Para melhor avaliar o impacto dos parâmetros dos itens no cálculo da informação, uma análise de sensibilidade será feita.

Análise de sensibilidade

Nesta seção serão apresentados alguns exemplos de curvas características e de curvas de Informação de itens com diferentes combinações dos parâmetros a, b e c. Pretende-se avaliar a influência dos parâmetros dos itens no cálculo da função de Informação.

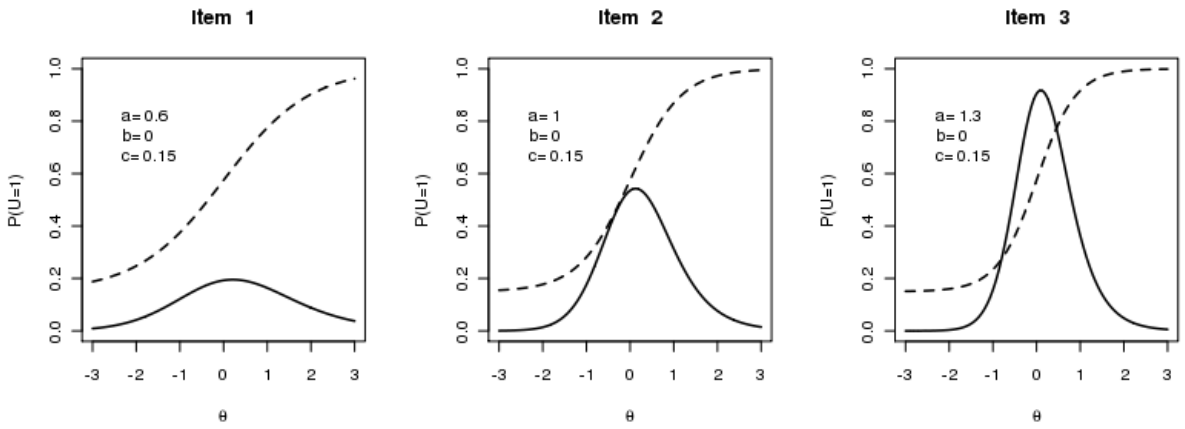


Figura 3.4: Influência do parâmetro a na Informação do item.

Parâmetro a

Considerando fixos os parâmetros b e c, a Figura 3.4 mostra que quanto maior o valor do parâmetro a, maior é a informação do item. Isso já era previsto, pois a informação do item é diretamente proporcional ao parâmetro a, ver a fórmula 3.6. Percebe-se também que quanto maior o valor do parâmetro de discriminação, mais acentuada será a curva característica. A conseqüência disso é que a diferença entre as probabilidades de resposta correta de dois examinandos com proficiência $-1,0$ e $0,0$, por exemplo, é maior no item 3 ($P(U = 1|\theta_0) - P(U = 1|\theta_{-1}) = 0,75$) do que em relação a item 1 ($P(U = 1|\theta_0) - P(U = 1|\theta_{-1}) = 0,07$) ou item 2 ($P(U = 1|\theta_0) - P(U = 1|\theta_{-1}) = 0,36$). Conclui-se, portanto, que o item 3 é mais apropriado (informativo) para discriminar estes dois indivíduos do que os outros itens.

Parâmetro b

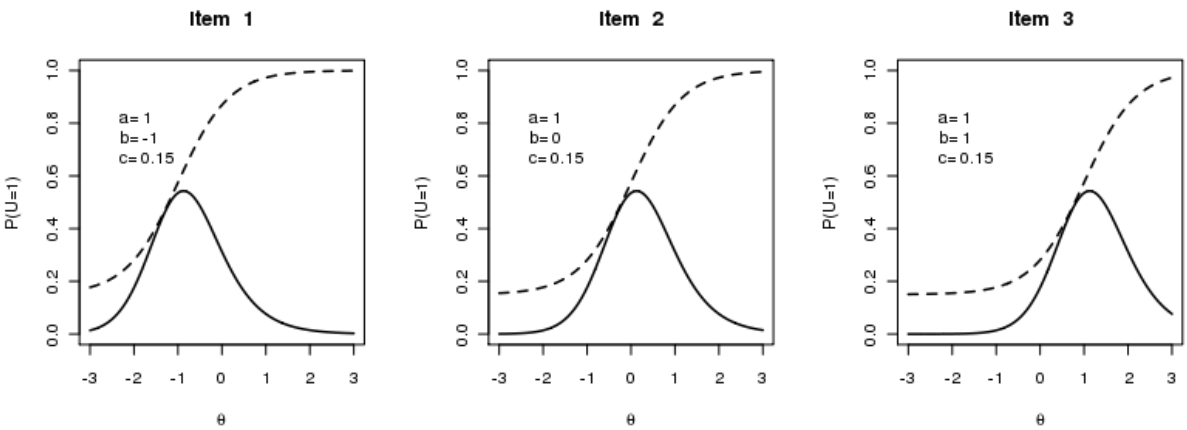


Figura 3.5: Influência do parâmetro b na Informação do item.

Por meio da Figura 3.5, nota-se que, fixando os parâmetros a e c, tanto a curva característica como a curva de informação deslocam-se ao longo dos níveis de proficiências. Em outras palavras, o parâmetro b está intrinsecamente ligado à posição da curva na escala de proficiência. Como já dito, maximizar a Informação de Fisher significa intuitivamente selecionar um item de dificuldade que corresponda exatamente ao nível de proficiência do examinando.

Parâmetro c

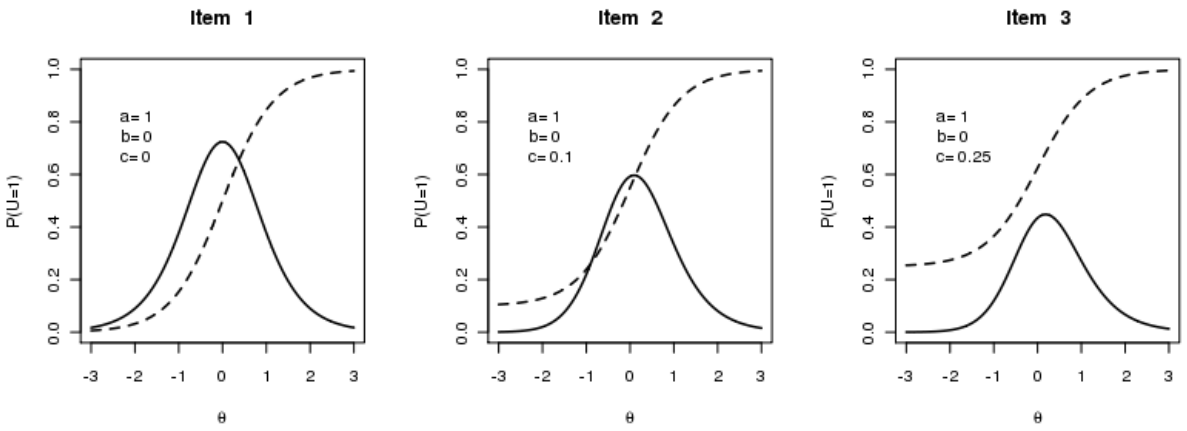


Figura 3.6: Influência do parâmetro c na Informação do item.

Diferentemente do parâmetro a, quanto menor o parâmetro c, maior será a altura da curva de Informação. Isso pode ser observado na Figura 3.6.

3.3.2 Critério de Máxima Informação Global

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, θ.

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 extremamente 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.

Chang e Ying (1996) denominaram a medida de Informação de Fisher como uma *Função de Informação Local*. Seja θ_0 o valor verdadeiro de θ. Se a informação ao redor de uma região pequena de θ_0 é vista como informação local, a informação fora desta região pode ser vista como informação global.

Na teoria estatística de testes de hipóteses, existem dois tipos de hipóteses alternativas para a hipótese Nula - local e fixa. Por exemplo, se a hipótese Nula é $H_0 : \theta = \theta_0$, então uma hipótese alternativa fixa pode ser $H_1 : \theta = \theta_1$, e uma hipótese alternativa local, relativa a uma amostra de tamanho m, pode ser: $H_1 : \theta = \theta_0 + (\theta_1/\sqrt{m})$. A hipótese alternativa local aproxima-se da hipótese Nula quando m cresce, diferentemente da hipótese alternativa fixa. É razoável esperar que informação local está relacionada ao poder de detectar uma hipótese alternativa local e informação global, hipótese alternativa fixa. Com relação ao CAT, informação local serve como referência para seleção de itens quando existe suficiente conhecimento sobre a localização de θ_0 e informação global deve ser preferido quando não há informações suficientes sobre esse parâmetro.

Dado um vetor de respostas $U_1, \ldots, U_{k-1}$ para os $k-1$ itens do teste, a quantidade que resume toda informação para a proficiência do examinando é a função de Verossimilhança $L(\theta_j; u_1, \ldots, u_{k-1})$ definida em 3.1. Para diferenciar qualquer valor fixo θ_1 de θ_0, a medida de Kullback-Leibler avalia a diferença entre os valores de L para θ_1 e θ_0. Tal diferença pode ser capturada pela razão de dois valores, resultando no mais conhecido Teste da Razão de Verossimilhança (Neyman e Pearson, 1936). Pela teoria de Neyman-Pearson, o método da Razão de Verossimilhança é ótimo para se testar $\theta = \theta_0$ contra $\theta = \theta_1$. Como definido por Chang e Ying (1996), essa é a melhor forma de avaliar θ_1 de θ_0 quando assume-se o modelo da TRI para as $u_1, \ldots, u_{k-1}$ observações.

Uma vez que os erros associados ao Teste da Razão de Verossimilhança decrescem para 0 exponencialmente rápido, faz-se conveniente tomar o logaritmo da razão da Verossimilhança. Além disso, como a função de Verossimilhança é um produto, o logaritmo dessa função é uma soma, muito mais fácil de se trabalhar. Uma das conseqüências disso consiste na aditividade da medida da Informação, que não seria possível sem a transformação logarítmica. O valor esperado da Razão do log-Verossimilhança é conhecido como a Informação de Kullback-Leibler (Kullback, 1959). Essa medida permite quantificar quão poderoso (eficiente) o teste estatístico é, além de avaliar a discrepância entre duas distribuições de probabilidade especificadas por θ_0 e θ_1.

Definição 3.3.2 (Informação de Kullback-Leibler) *Seja θ_0 o valor verdadeiro do*

parâmetro. Para qualquer valor de θ, a Informação KL para o i-ésimo item (com resposta u_i) é definido por:

$$K_i(\theta||\theta_0) \equiv E_{\theta_0} \log \left[\frac{L_i(\theta_0; u_i)}{L_i(\theta; u_i)} \right].$$

A notação de duas barras verticais é padrão para a Informação KL (Kullback, 1959). Essa simbologia evita a confusão com uma única barra que é tipicamente utilizada para indicar condicionamento.

Para o modelo da TRI, como $U_i \sim Bernoulli(P_i(\theta_0))$, a medida de Informação KL pode ser expressa por:

$$\begin{aligned} K_i(\theta||\theta_0) &= E_{\theta_0} \log \left[\frac{P_i(\theta_0)^{u_i} \left[1 - P_i(\theta_0)\right]^{1-u_i}}{P_i(\theta)^{u_i} \left[1 - P_i(\theta)\right]^{1-u_i}} \right] \\ &= P_i(\theta_0) \log \left[\frac{P_i(\theta_0)}{P_i(\theta)} \right] + \left[1 - P_i(\theta_0)\right] \log \left[\frac{1 - P_i(\theta_0)}{1 - P_i(\theta)} \right], \end{aligned} \tag{3.7}$$

em que: $K_i(\theta||\theta_0) \geq 0$ e K_i não é simétrica, ou seja: $K_i(\theta||\theta_0) \neq K_i(\theta_0||\theta)$.

Análogo à Informação de Fisher, pode-se obter uma medida de informação do teste por meio da função de KL. Para um teste com $k-1$ itens, a Informação do teste Kullback-Leibler consiste na soma de todas as informações individuais dos itens e denota-se por:

$$K(\theta||\theta_0) = \sum_{i=1}^{k-1} K_i(\theta||\theta_0) = E_{\theta_0} \left[\log L(\theta_0; u_i, , u_{k-1}) - \log L(\theta; u_i, , u_{k-1}) \right].$$

Uma importante propriedade de K é que esta medida é uma função de dois níveis, θ e θ_0. Chang e Ying (1996) destacam que K representa o poder discriminatório do item nos dois níveis. Isto não requer que θ seja próximo de θ_0. Dessa forma, K resume a informação contida no item com respeito a uma amplo espectro de θ. Em contraste, a Informação de Fisher é uma função de θ_0 e somente representa o poder discriminatório ao redor de θ_0 (Hambleton, Swaminathan e Rogers, 1991).

A proposta do CAT é estimar precisamente a proficiência do examinando θ_0 pela eficiente seleção dos itens. Para esta finalidade, é desejável encontrar uma quantidade que diferencie todos valores θ de θ_0 (onde $\theta \neq \theta_0$). Segundo Chang e Ying (1996), a Razão do log das Verossimilhanças é, de certa forma, a melhor quantidade construída dos dados para fazer essa distinção. K é o valor médio da Razão log-Verossimilhança. Para um item i e θ variando sobre o espaço paramétrico, Θ, a medida K gera um perfil global sobre o poder discriminatório do item. Isto não requer que θ esteja próximo de θ_0. Dessa maneira, Chang e Ying (1996) definiram K como uma medida de Informação Global.

Para cada θ_0, K é uma função de θ e I_F (Informação de Fisher) é um valor fixo. Essa é uma das principais diferenças entre K e I_F citadas por Chang e Ying (1996). Se θ_0 varia ao longo da escala, K se torna uma superfície de informação global num espaço tridimensional: (γ, ν, κ), onde: γ corresponde a θ_0, ν a $\hat{\theta}$ e κ a K. A Figura 3.7 ilustra uma superfície de Informação KL interceptada por um plano vertical ($\gamma = 0$), para um item do modelo ML3P. A curva resultante no plano é a medida de Informação para $\theta_0 = 0$. Um significado geométrico da função de Informação KL para um valor fixo de θ_0 é uma curva, que representa a interseção do plano vertical $\gamma = \theta_0$ e a superfície da informação. Da Figura 3.7, observa-se que a função de Informação KL muda sua forma conforme θ_0 muda seus valores. Não importa como é a mudança, K é sempre 0 quando $\hat{\theta} = \theta_0$.

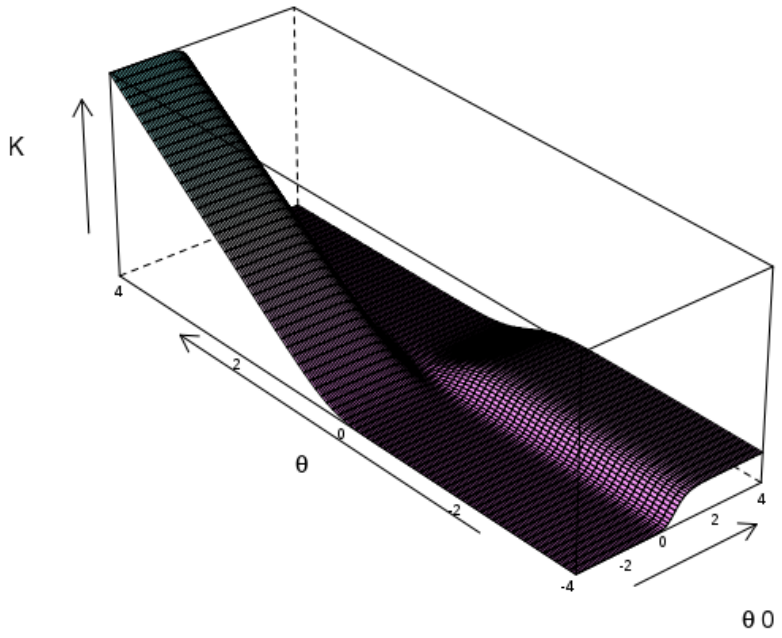


Figura 3.7: Superfície de Informação KL para um item com $a = 2,5; b = 0,0$ e $c = 0,12$; interceptando um plano vertical $\gamma = 0$.

Note que a curvatura de $\theta = \theta_0$ na Figura 3.7 equivale a I_F para θ_0. Isto não é difícil de ser avaliado matematicamente uma vez que $\frac{\partial^2 K(\theta||\theta_0)}{d\theta^2}|_{\theta=\theta_0} = I_F(\theta_0)$ (Chang e Ying, 1996).

A relação entre a segunda derivada e a Informação de Fisher indica que se o perfil de K é conhecido, então I_F também será conhecida. Contudo, o contrário não é verdadeiro. Dessa maneira, Chang e Ying (1996) concluíram que a Informação KL do teste (ou do item) é mais informativa que a informação mais convencional (Fisher).

Para melhor avaliar o impacto dos parâmetros dos itens do modelo ML3P no cálculo

da Informação KL, uma análise de sensibilidade será feita.

Análise de sensibilidade

A Figura 3.8 apresenta a função de Informação de Kullback-Leibler para cinco itens com $\theta_0 = 1$. Para cada função, o valor da curvatura em 1 é equivalente ao valor da Informação de Fisher em 1. Análogo à análise de sensibilidade dos parâmetros em I_F, a medida de KL também é influenciada pelos parâmetros dos itens.

A Informação de Fisher para o item 5 é de $0,52$ enquanto para o item 4 tem-se que $I_F = 0,27$. Dessa forma, o item 5 fornece maior Informação de Fisher que o item 4, contudo, este não é o caso da Informação de KL, que mostra uma relação mais complexa entre esses itens. Em relação ao item 5, o item 4 apresenta maior Informação KL quando $\theta < 1,0$ e valores próximos de informação quando $\theta \geq 1,0$.

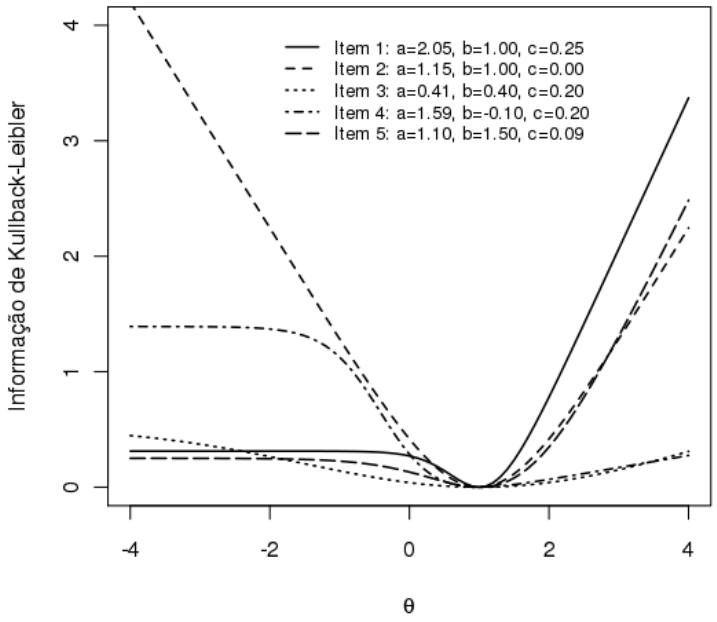


Figura 3.8: Função de Informação KL para cinco itens em $\theta_0 = 1,0$.

A Figura 3.9 ilustra tanto a Informação de Fisher como a Informação KL para dois itens quando $\theta_0 = 0$. Apesar de I_F do item 1 ser maior do que do item 2 ao redor de θ_0, o item 2 aparenta ser a melhor escolha para K, pois esse apresenta-se mais "robusto" e possui maior poder quando considera-se todo o espaço paramétrico.

O maior impedimento para se selecionar itens pela medida de Informação de KL está no fato de que para um dado θ_0, K é uma função no espaço paramétrico enquanto a Informação de Fisher fornece apenas um número. Embora K seja mais complicada de se

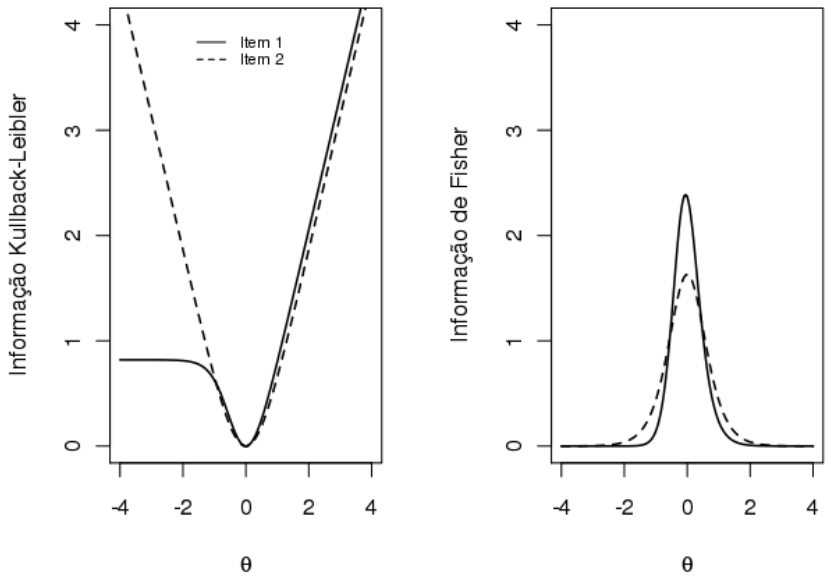


Figura 3.9: Funções de informação para dois itens. Item 1: $a = 2,0$; $b = -0,1$ e $c = 0,1$ e Item 2: $a = 1,5$;$b = 0,0$ e $c = 0,0$.

obter analiticamente do que I_F, Chang e Ying (1996) estabeleceram um procedimento lógico para o uso da medida de KL para seleção dos itens.

Segundo Chang e Ying (1996), uma forma mais simples para se construir um único valor para K consiste em se tomar o valor médio sobre um apropriado intervalo da proficiência estimada, $\hat{\theta}$. Um índice médio de Informação KL pode ser definido como:

$$K_i(\hat{\theta}) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} K_i(\theta||\hat{\theta})d\theta, \tag{3.8}$$

onde δ determina o tamanho do intervalo sobre o qual a média será computada.

O índice apresentado em 3.8 corresponde a área sobre a função KL de $\hat{\theta}-\delta$ a $\hat{\theta}+\delta$. O efeito da curvatura em $\hat{\theta}$ é claro. Para um pequeno valor de δ, este índice é essencialmente determinada pela curvatura de $K_i(\theta||\hat{\theta})$ em $\hat{\theta}$. Em outras palavras, a área máxima será equivalente à máxima curvatura e, conseqüentemente, no valor máximo de I_F. Para grandes valores de δ, a área é muito influenciada pelos extremos de $K_i(\theta||\hat{\theta})$. Dessa maneira, a seleção de itens baseadas na área máxima definida em 3.8 reflete a idéia da abordagem da informação global.

Note que a integração em 3.8 foi definida com relação à medida de Lebesgue no intervalo $(\hat{\theta}-\delta, \hat{\theta}+\delta)$, ou seja, a função de densidade K_i é uniforme no intervalo $(\hat{\theta}-\delta, \hat{\theta}+\delta)$.

A medida de Lebesgue foi selecionada por conveniência, embora outras medidas pu-

dessem ser consideradas (Chang e Ying, 1996). De modo geral, seja μ qualquer medida de probabilidade no espaço paramétrico. O índice KL pode ser definido como:

$$K_i^{\mu}(\hat{\theta}) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} K_i(\theta||\hat{\theta}) d\mu(\theta). \tag{3.9}$$

O índice 3.9 inclui a equação 3.8 com um caso especial, com μ sendo definida como a medida de Lebesgue dentro do intervalo $(\hat{\theta}-\delta, \hat{\theta}+\delta)$ e 0, a medida fora desse intervalo.

Implementação do índice médio da Informação KL requer a especificação de δ. Chang e Ying (1996) destacam que mais eficiente será o critério de seleção pela medida KL no contexto do CAT se δ decresce para 0 quando o tamanho do teste se aproxima de ∞. Na seleção de δ é esperado que o resultado do intervalo $(\hat{\theta}-\delta,\hat{\theta}+\delta)$ contenha θ_0. Pela teoria assintótica, o estimador de Máxima Verossimilhança de $\hat{\theta}$ é assintoticamente Normal com média θ_0 e variância $1/I_F(\theta_0)$. Dessa maneira, um intervalo de confiança para θ_0 deve ser do tipo $\left[\hat{\theta} - \frac{p}{\sqrt{I_F(\hat{\theta})}}, \hat{\theta} + \frac{p}{\sqrt{I_F(\hat{\theta})}}\right]$, onde a constante p é selecionada de acordo com uma específica probabilidade de cobertura. Para um teste de tamanho n, a medida de Informação de Fisher será de ordem n, logo, Chang e Ying (1996) concluíram que um valor razoável para δ seria igual a $\frac{p}{\sqrt{n}}$.

3.3.3 Critério da Máxima Informação Esperada

O critério da Máxima Informação Esperada (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 seqüencial. 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 com 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, 1999). 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:

$$P_s(u_s|u_1,\ldots,u_{k-1}) = \int P_s(u_s|\theta)g(\theta|u_1,\ldots,u_{k-1})d\theta, \tag{3.10}$$

em que: $P_s(u_s|\theta)$ é a probabilidade preditiva da resposta u_s ao item s dado θ e $g(\theta|u_1,\ldots,u_{k-1})$ é a densidade *a posteriori* após $k-1$ itens definida em 3.2 na página 49.

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

A motivação para a adoção do critério MIE vem de van der 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 conseqüência, a Informação de Fisher, definida como o valor esperado da variável aleatória U não é uma medida válida. Uma escolha Bayesiana típica neste caso é o uso da medida de Informação Observada:

$$J_{u_1,\ldots,u_{k-1}}(\theta) = -\frac{\partial^2 \log L(\theta; u_1,\ldots,u_{k-1})}{\partial\theta^2}, \tag{3.11}$$

que reflete a curvatura da função de Verossimilhança observada para o θ relativo à métrica escolhida para θ.

Porque a estimativa pontual $\hat{\theta}$ é atualizada iterativamente, a medida de Informação Observada não deve ser somente atualizada para a resposta ao k-ésimo item selecionado, e sim para todas as respostas anteriores da mesma maneira.

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 denotado 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 $S_{k-1} = \{i_1,\ldots,i_{k-1}\}$, enquanto os itens restantes formam o conjunto $R_k = \{1,\ldots,I\} \setminus$

S_{k-1}. A seleção do k-ésimo obedecerá à seguinte regra:

$$i_k = \arg\max_s \{P_s(0|u_1,\ldots,u_{k-1})J_{u_1,\ldots,u_{k-1},U_s=0}(\hat{\theta}_{u_1,\ldots,u_{k-1},U_s=0}) + P_s(1|u_1,\ldots,u_{k-1})J_{u_1,\ldots,u_{k-1},U_s=1}(\hat{\theta}_{u_1,\ldots,u_{k-1},U_s=1}) : s \in R_k\}.$$

3.4 Métodos de controle da exposição de itens

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 a controlar certos atributos como balanceamento do conteúdo ou freqüência 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-caneta 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.

Pretende-se, com esta seção, apresentar duas principais estruturas que devem ser observadas ao se aplicar um teste adaptativo: controle da freqüência da exposição dos itens e balanceamento do conteúdo. A restrição em relação à freqüência 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 decorá-los, adicionando assim um erro na estimativa da proficiência e, conseqüentemente, prejudicando a validade do teste.

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 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-caneta. Em Matemática, por exemplo, espera-se que a prova adaptativa aborde vários conteúdos, tais como: Geometria, Álgebra e Trigonometria.

3.4.1 Controle da freqüência de exposição de itens

A segurança de um teste adaptativo pode ser comprometida se a freqüência de exposição dos itens não for bem controlada. Uma vez que manter a segurança dos testes é importante, especialmente nas avaliações de larga escala, um teste adaptativo informatizado não pode ser implementado eficientemente na prática sem que a exposição dos itens seja bem administrada.

Se o CAT for delineado de forma que os examinandos iniciam o teste com a mesma estimativa provisória da proficiência, sob o critério de seleção dos itens pela Máxima Informação, o item mais informativo será o mesmo para todos os examinandos, o segundo item será um entre as duas escolhas (será um item se a resposta for correta ou outro após uma resposta incorreta) e assim por diante. Como conseqüência, a seqüência da administração do item será previsível e os itens iniciais serão usados com mais freqüência - tornando-se superexpostos.

Itens com alta taxa de exposição podem produzir efeitos indesejáveis. Os itens mais freqüentes podem rapidamente se tornar populares aos examinandos e perder suas propriedades psicométricas além de, conseqüentemente, diminuir a validade do teste (Revuelta e Ponsoda, 1998).

Georgiadou, Triantafillou e Economides (2007) citam diversas estratégias para controle da exposição de itens desde pesquisas iniciadas em 1983 até 2005. Nesta dissertação serão destacados os seguintes procedimentos: procedimentos probabilísticos e os métodos de estratificação do banco de itens.

Procedimento Probabilístico

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 (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 freqüência 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 sintetizado em 8 passos.

Algoritmo de Sympson-Hetter

Os passos 1 a 3 são realizados somente uma vez para cada CAT. Enquanto os passos 4 a 8 são repetidos até que um critério seja alcançado.

Passo 1 Especificar o valor máximo esperado para a taxa de exposição do item (r) para o teste.

Passo 2 Construir uma tabela de Informação. Conhecida como *infotable*, esta tabela consiste em uma lista de itens por proficiência (como na Tabela 2.4 da página 40). Em cada lista, todos os itens do banco são organizados em ordem decrescente do valor de suas funções de Informação (geralmente, Informação de Fisher) calculada em cada nível de proficiência.

Passo 3 Gerar os primeiros conjuntos de parâmetros para cada item. Se um banco contém I itens, gerar um vetor de tamanho I com todos os elementos iguais a 1. Esse vetor será denominado de $P(A|S)$ e representa o parâmetro de exposição dos itens do banco. Em outras palavras, o parâmetro de exposição pelo método SH representa a probabilidade do item ser administrado (A) dado que esse foi selecionado (S).

Passo 4 Simular um teste adaptativo informatizado para uma amostra aleatória de examinandos. Esse teste deve ser administrado para um grande grupo de examinandos simulados cujas proficiências "verdadeiras" são aleatoriamente amostradas da distribuição de proficiências representativa da real população de examinandos. Para cada item do CAT, identificar o mais informativo item da *infotable* mais próximo do valor estimado ($\hat{\theta}$). Gerar um número x pseudo-aleatório da distribuição Uniforme $(0, 1)$. Administrar o item i se x for menor ou igual ao correspondente $P_i(A|S)$. Se o item i foi ou não administrado, excluir esse item dos futuros itens selecionados para o mesmo examinando. Note que para a primeira simulação, todos os $P_i(A|S)$ são iguais a 1 e todos os itens serão administrados, se selecionados.

Passo 5 Anote o número de vezes que cada item do banco é selecionado (NS) e o número de vezes que é administrado (NA) no total de amostras simuladas. Quando a amostra completa foi testada, calcule $P(S)$, a probabilidade com que um item é selecionado, e $P(A)$, a probabilidade com que um item é administrado para cada item:

$$P(S) = \frac{NS}{NE} \quad \text{e} \quad P(A) = \frac{NA}{NE},$$

onde NE é o total de examinandos.

Passo 6 Utilizando o valor de r do Passo 1 e $P(S)$, calcule o novo valor de $P_i(A|S)$ como segue:

$$P_i(A|S) = \begin{cases} \frac{r}{P(S)}, & \text{se} \quad \text{P(S)>r,} \\ 1, & \text{se} \quad \text{P(S)}\leq r. \end{cases} \tag{3.12}$$

Passo 7 Para um teste adaptativo de tamanho n, assegure-se de que exista pelo menos n itens no banco que possuam novo $P_i(A|S)$ igual a 1. Itens com $P_i(A|S) = 1$ são sempre administrados quando selecionados, uma vez que o número aleatório gerado sempre será menor ou igual a 1. Se houver menos que n itens com $P_i(A|S) = 1$, faça que os n maiores valores de $P_i(A|S)$ sejam iguais a 1. Isto garante que todos os examinandos irão completar o teste de tamanho n antes do banco se desgastar.

Passo 8 Dado novos valores de $P(A|S)$, retorne ao Passo 4. Utilizando a mesma amostra de examinandos, repita os Passos 4, 5, 6 e 7 até que o máximo valor de $P(A)$ que será obtido no Passo 5 aproxime-se do limite levemente acima de r e então oscile nas sucessivas simulações.

Esse procedimento baseia-se no seguinte argumento: Se um item i for selecionado pelo algoritmo do CAT com probabilidade menor que o valor r máximo, a inequação 3.12 implica que este item possui taxa de exposição menor que o r máximo e, portanto, não necessita de controle ($P_i(A|S) = 1$). Por outro lado, se o item i for selecionado com probabilidade maior que r máximo, o parâmetro de controle deve ser ajustado tal que $P(A) = r$. Esse ajuste será igual a $P_i(A|S) = \frac{r}{P(S)}$.

van der Linden e Glas (2003) ressaltam que o valor de r nunca deve ser menor que $\frac{n}{I}$, onde n é o tamanho do teste adaptativo e I, o tamanho do banco de itens. Como na prática, um banco de itens é tipicamente 7 a 10 vezes o tamanho de um teste adaptativo, os valores mínimos de r correspondentes seriam de $0,10$ a $0,14$. Geralmente, os valores de r mais utilizados estão entre $0,20$ e $0,30$.

Os valores $P_i(A|S)$ obtidos ao final das simulações serão os parâmetros de controle da exposição dos itens no teste real. No CAT real, o uso de $P_i(A|S)$ durante o CAT processa-se como no Passo 4: (1) Seleciona-se o mais informativo item para a corrente estimativa da proficiência; (2) Gera-se um número pseudo-aleatório x de uma distribuição Uniforme $(0, 1)$; (3) Se x for menor ou igual ao item $P_i(A|S)$, o item será administrado, caso contrário, este não será administrado e será identificado outro item mais informativo. Posteriormente, (1), (2) e (3) serão repetidos até que se alcance algum critério de parada.

Uma das conseqüências da imposição de restrição no controle dos itens com tendência à superexposição é a possível redução na informação do teste para a estimação da proficiência. A presença da restrição proíbe que o algoritmo administre os itens com o valor máximo da Informação quando se utiliza o critério de seleção MI. Como menciona van der Linden e Glas (2003), a quantidade de informação perdida não é grande se os itens do banco são de alta qualidade e se o algoritmo conseguir administrar itens com informação levemente menor que o item mais informativo para a estimativa da proficiência.

Segundo Chang e Ying (1999), o procedimento probabilístico pode efetivamente controlar a taxa máxima de controle da exposição de um item. Contudo, existem duas limitações: primeiro, os itens que não foram selecionados não podem ser administrados; então, itens que possuem baixa probabilidade de serem selecionados continuam apresentando baixas taxas de exposição. Segundo, os parâmetros de controle da exposição necessitam de atualização através de um número grande de complicadas simulações a cada alteração do banco de itens ou se a distribuição das proficiência da população de interesse for modificada. Dessa forma, alguns métodos de estratificação do banco de itens foram propostos na literatura para complementar essa metodologia e tentar contornar tal situação.

Método de estratificação do banco de itens

Um dos primeiros métodos de estraficação do banco de itens foi proposto por Chang e Ying (1999). Esse método tem-se mostrado bastante eficiente na administração do banco de itens. Contudo, nenhuma simulação é feita para se obter os parâmetros de controle da exposição dos itens.

Nesse procedimento, o banco de itens é divido em diferentes estratos baseado nos valores dos parâmetros dos itens e o teste adaptativo é dividido em estágios. No método de estratificação pelo parâmetro a (Chang e Ying, 1999), por exemplo, divide-se os itens do banco em diversos estratos em ordem ascendente do parâmetro de discriminação. Cada teste consiste em um número idêntico de estágios e estratos. O primeiro estágio consiste em administrar itens com menores parâmetros a selecionados do primeiro estrato. Os estágios subseqüentes selecionarão itens mais discriminativos que pertencem aos diferentes estratos.

Uma das melhores justificativas para o uso de tal delineamento é que, nos estágios iniciais do teste, o ganho na informação usada pelos itens mais discriminativos não é

adequado porque a estimação da proficiência ainda é relativamente imprecisa. Conseqüentemente, itens com altos valores do parâmetro de discriminação devem ser usados nas fases finais do teste (Chang e Ying, 1999).

Sob essa perspectiva, Chang, Qian e Ying (2001) mencionam que, ao se estratificar o banco de itens pelo parâmetro a, alguns bancos de itens podem não possuir itens suficientes com baixos valores do parâmetro de dificuldade no último estrato do teste. Dessa forma, esses autores desenvolveram outro método de estratificação do banco pelo parâmetro a utilizando um bloco estabelecido pelo parâmetro b. Nesse método, o banco de itens é dividido em pequenos níveis baseados nos parâmetros de dificuldade dos itens. Em relação a cada nível, itens são classificados na ordem ascendente dos valores de a. Em seguida, itens com menores valores de discriminação de cada nível são agrupados no primeiro estrato, itens com os segundos menores valores de a são agrupados no segundo estrado e assim por diante. Eventualmente o último estrato conterá os itens mais discriminativos de cada nível do parâmetro da dificuldade.

3.4.2 Balanceamento do conteúdo

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. Por exemplo, um estudo piloto em CAT foi realizado para análise das habilidades dos estudantes do Ensino Fundamental (por exemplo, 4ª série) em Matemática. Dessa maneira, foram considerados quatro descritores para avaliar essa área do conhecimento. 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.

3.5 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. O objetivo de cada teste é mensurar cada indivíduo o melhor possível. No contexto do CAT, dois objetivos são operacionalizados para dois diferentes critérios de parada do teste.

Para fins de classificação, a proficiência de um examinando é comparado com algum valor de corte. Weiss e Kingsbury (1984) indicam que, para implementação no CAT, tanto a estimativa da proficiência como o erro-padrão da medida associado devem ser usados. No caso da estimação das proficiências 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) se a estimativa da proficiência 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 proficiência.

Quando um teste adaptativo informatizado não é usado para classificação, uma diferente regra de parada se aplica. Nesse caso, é desejável mensurar cada examinando para um nível de precisão fixo, ou seja, um nível pré-determinado do erro-padrão é fixado (Weiss e Kingsbury, 1984). Isso resultará em um conjunto de medidas "equiprecisas", em que todos os examinandos terão escores com equivalentes precisões. Para implementar a medida equiprecisa, o CAT permite que os usuários especifiquem o nível da medida de erro-padrão que seja desejável para cada examinando. Assumindo que o banco de itens tem um número suficiente de itens distribuídos em toda escala do traço latente e que o tamanho do teste seja o suficiente para cada examinando, o teste é finalizado quando o nível do erro-padrão da medida for satisfeito.

Alguns algoritmos em CAT são finalizados pelo administrador quando atingir um número fixo de itens ou por imposição de um tempo limite. Ambos os casos são usados por conveniência do administrador do teste e não é considerada uma boa prática nesse tipo de teste. No caso em que o teste é utilizado para classificação, a qualidade do teste pode prejudicar a estimativa de alguns examinandos. No outro caso, um término prematuro do CAT não irá resultar em uma medida equiprecisa, já que a medida de erro-padrão não decresce para todos os examinandos na mesma taxa. Para obter o máximo de benefícios do CAT, nem o tempo limite nem o tamanho do teste deveriam ser impostos como critérios de parada.

Capítulo 4

Aplicação à prova de Proficiência

Este capítulo tem por objetivo comparar os três métodos de seleção adaptativa de itens descritos no Capítulo 3: Máxima Informação (MI), Kullback-Leibler (KL) e Máxima Informação Esperada (MIE); para estimação das proficiências em CAT utilizando o banco de Proficiência em Inglês Instrumental I. Na Seção 4.1 será apresentada a metodologia empregada para avaliação dos algoritmos, enquanto na Seção 4.2, os resultados referentes a este estudo.

4.1 Metodologia

Com a finalidade de se avaliar os três procedimentos mais conhecidos na literatura para seleção de itens em CAT, cinco estudos de simulações foram conduzidos. Em todos eles, utilizaram-se os 246 itens da prova de Proficiência em Inglês Instrumental I calibrados pelo ML3P na escala $(0, 1)$.

Vale salientar que, para se obter as proficiências de Inglês Instrumental I na escala interpretável de média 100 e desvio 25, basta simplesmente realizar a transformação $\theta^* = 25 \times \theta + 100$ descrita no Capítulo 2, página 40.

Após a seleção de cada item, as respostas dos examinandos foram simuladas de uma distribuição de Bernoulli ($P(\hat{\theta})$), onde $P(\hat{\theta})$ representa a probabilidade de acerto ao item pelo ML3P definida em 1.1, considerando fixos os parâmetros dos itens do banco e a estimativa da proficiência corrente.

O método iterativo para estimação das proficiência foi o da *Média a Posteriori* (EAP). Após a observação de cada resposta (1-resposta correta e 0-resposta incorreta), o estima-

dor EAP para θ foi calculado numericamente com base na Equação 3.4. A distribuição *a priori* assumida foi a Uniforme $(-6{,}00; 6{,}00)$ e, dessa maneira, supôs-se que a distribuição dos examinandos estariam totalmente concentrada no intervalo $-6{,}00$ a $6{,}00$.

Ainda para o cálculo das estimativas pelo método EAP, 40 pontos igualmente espaçados no intervalo ± 6 desvios-padrão foram empregados e os pesos associados a esses pontos foram iguais à densidade da *priori*.

Os procedimentos de seleção dos itens podem ser sintetizados da seguinte maneira:

Início: Especificou-se um valor inicial, $\hat{\theta}_0$, para a proficiência do examinando.

Iteração: Estimou-se a proficiência do examinando $\hat{\theta}$ após a resposta ao item. A escolha do próximo item administrado no CAT levou em conta a medida de informação dos itens no ponto $\hat{\theta}$. Dessa forma, seja denotado i o i-ésimo item do banco, $i = 1, ..., I$, e k a ordem em que o i-ésimo item aparece no teste adaptativo. Suponha que $k-1$ itens foram administrados no CAT. Os índices dos itens administrados formam o conjunto $S_{k-1} = \{i_1, \ldots, i_{k-1}\}$, enquanto os itens restantes formam o conjunto $R_k = \{1, \ldots, I\} \setminus S_{k-1}$. A seleção do k-ésimo obedeceu às seguintes regras:

Máxima Informação (MI): Dado $\hat{\theta}$, o k-ésimo item foi selecionado de maneira que $I_{F,k}(\hat{\theta}_{k-1})$ era o maior valor entre todos os itens do banco, ou seja:

$$i_k \equiv \arg\max_s \left\{ I_{F,U_s}\left(\hat{\theta}_{u_1,\ldots,u_{k-1}}\right) : s \in R_k \right\}.$$

Kullback-Leibler (KL): Dado $\hat{\theta}$, o k-ésimo item foi selecionado de maneira que $K_k(\hat{\theta}_{k-1})$ era o maior valor entre todos os itens do banco, ou seja:

$$i_k \equiv \arg\max_s \left\{ \int_{\hat{\theta}_{k-1}-\delta_k}^{\hat{\theta}_{k-1}+\delta_k} K_s\left(\theta || \hat{\theta}_{k-1}\right) : s \in R_k \right\}.$$

Para o cálculo de K utilizou-se $\delta_k = \frac{3}{\sqrt{n}}$, onde p=3 configura uma probabilidade de cobertura de aproximadamente 99%.

Máxima Informação Esperada (MIE): Dado $\hat{\theta}$, o k-ésimo item foi selecionado de maneira que $J_k(\hat{\theta}_{k-1})$ ponderada pela probabilidade preditiva $P_k(u_k|u_1, \ldots, u_{k-1})$ era o maior valor entre todos os itens do banco, ou seja:

$$\begin{aligned} i_k \equiv \arg\max_s \{ & P_s(0|u_1, \ldots, u_{k-1}) J_{u_1,\ldots,u_{k-1},U_s=0}(\hat{\theta}_{u_1,\ldots,u_{k-1},U_s=0}) \\ & + P_s(1|u_1, \ldots, u_{k-1}) J_{u_1,\ldots,u_{k-1},U_s=1}(\hat{\theta}_{u_1,\ldots,u_{k-1},U_s=1}) : s \in R_k \}. \end{aligned}$$

A distribuição preditiva *a posteriori* para a resposta ao k-ésimo item foi aproximada utilizando os mesmos pontos de quadratura e pesos da estimação EAP.

Parada: O algoritmo finalizou-se quando um número total de itens pré-estabelecido foi administrado ou quando um nível de precisão para θ foi atingido.

As medidas de Informação de Fisher, Kullback-Leibler e Observada (denotadas respectivamente por: I_F, K e J) para o modelo de três parâmetros da TRI estão descritas detalhadamente no Apêndice desta dissertação.

As simulações dos testes adaptativos foram feitas no *software* R (Team, 2008). Em cada estudo, diferentes inicializações e critérios de parada dos algoritmos foram utilizados.

É importante lembrar que alguns itens devem ser administrados seqüencialmente no CAT por fazerem referência a um mesmo elemento (texto, gravura), muito comum em provas de Línguas. Na literatura, quando um banco de itens foi desenvolvido dessa maneira, esse conjunto de questões conectadas é conhecido como *Testlets* (Wainer, Bradlow e Du, 2003). Assim, faz-se necessária a aplicação de restrições ao algoritmo de seleção dos itens de forma que, para cada *testlet*, os examinandos respondam às questões pertencentes a esse grupo e que este não seja mais repetido ao longo do CAT. Contudo, para a avaliação dos algoritmos de seleção adaptativa deste capítulo, esse componente não foi levado em consideração.

4.2 Resultados

4.2.1 Estudo de simulação 1

Este estudo se propôs a avaliar a quantidade de itens necessária no CAT para se obter estimativas das proficiências com medida de erro-padrão menor ou igual a 0,40 e 0,20. Para tanto, 500 testes adaptativos foram simulados para cada um dos métodos de seleção adaptativa de itens (MI, KL e MIE). Nesse estudo, os indivíduos foram simulados como proficiências variando no intervalo $-3{,}00$ e $+3{,}00$. Para essa simulação, determinou-se que os examinandos responderiam às questões do banco de Proficiência em Inglês Instrumental I até que o nível de precisão da estimativa estabelecido fosse alcançado.

A Tabela 4.1 apresenta a distribuição do número de itens necessário no CAT para se alcançar uma medida de erro-padrão de 0,40 e 0,20 para cada método de seleção de itens. Destaca-se que a medida de erro-padrão associada a estimativa θ é a raiz da variância *a posteriori* (PSD) descrita na página 51.

Para um critério de parada até que a medida de erro-padrão seja menor ou igual a

Tabela 4.1: Distribuição dos 500 casos simulados para se avaliar o número de itens requerido no teste adaptativo ao se fixar duas medidas de erro-padrão.

Número de itens	**Erro-padrão=0,4**			**Erro-padrão=0,2**		
requerido no CAT	**MI**	**KL**	**MIE**	**MI**	**KL**	**MIE**
Menor igual a 10	41	38	81	-	-	-
Entre 11 e 20	300	310	270	-	-	-
Entre 21 e 30	95	99	95	-	-	-
Entre 31 e 40	32	23	20	-	-	-
Entre 41 e 50	10	10	11	77	79	68
Entre 51 e 60	4	8	8	60	57	68
Entre 61 e 70	2	2	4	42	50	36
Maior ou igual a 71	16	10	11	321	314	328

0,40, observou-se que 341 dos 500 testes adaptativos simulados sob o critério de seleção de itens MI teriam até 20 itens. Já para KL, essa quantidade aumenta para 348 itens e no critério MIE obteve-se um percentual de 70,2% testes com tamanho menor ou igual a 20. Por outro lado, ao se finalizar o teste quando o erro-padrão da estimativa da proficiência for menor ou igual a 0,20, todos os casos simulados apresentaram uma quantidade total de itens superior a 40.

A Tabela 4.2 apresenta o número médio de itens, bem como o desvio-padrão, necessário no CAT para se obter estimativas das proficiências com erro menor ou igual a 0,40 e 0,20 para os três métodos.

Tabela 4.2: Número médio de itens nas 500 simulações.

	Erro-padrão=0,4			**Erro-padrão=0,2**		
Medida	**MI**	**KL**	**MIE**	**MI**	**KL**	**MIE**
Média	23	22	21	136	137	138
DP	30,4	22,1	27,1	84,2	85,4	84,5

Observou-se que o número médio de itens nos 500 testes simulados até se obter um erro-padrão menor ou igual a 0,40 pelo método Máxima Informação foi de 23, contra 22 do método Kullback-Leibler e 21, pelo método da Máxima Informação Esperada. Quando a medida de erro-padrão foi de 0,20, observou-se que o número médio de itens nos 500 testes simulados para se alcançar esse erro foi seis vezes maior (136 itens pelo método MI, 137 pelo KL e 138 pelo MIE).

4.2.2 Estudo de simulação 2

Neste estudo, 500 respostas de examinandos com proficiências entre −4,00 e +4,00 foram simuladas. O valor inicial de θ para cada examinando foi igual a $\hat{\theta}_0 = 0{,}00$. O primeiro item do CAT foi selecionado aleatoriamente entre todos os parâmetros de dificuldade abaixo do valor inicial $\hat{\theta}_0$ (ou seja: $b < 0{,}00$) e o teste finalizava quando a quantidade total de itens era de 25.

O Gráfico 4.2.2 apresenta a distribuição das estimativas em relação ao seu valor θ verdadeiro para cada critério de seleção de itens.

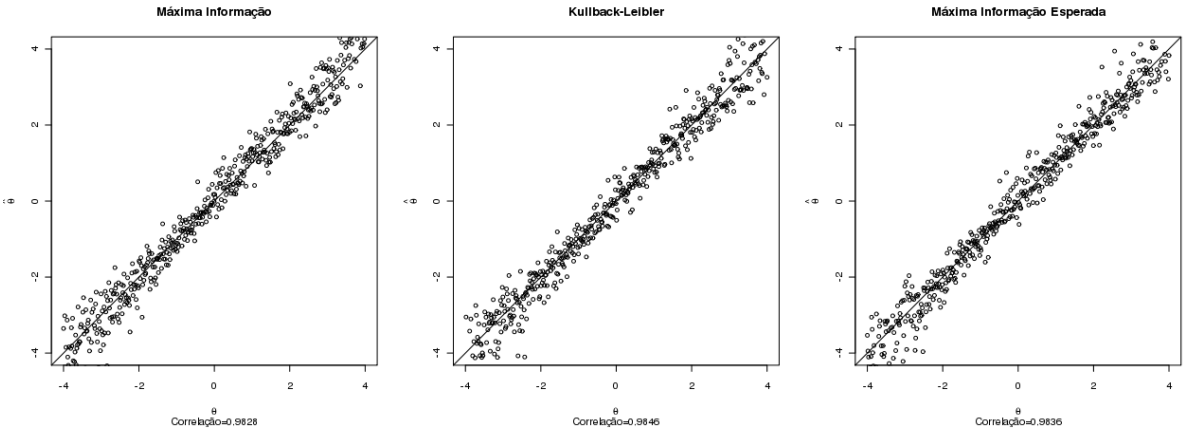


Figura 4.1: Valor verdadeiro × valor estimado para cada um dos métodos.

Observa-se que todos os três métodos conseguiram estimar bem as proficiências para um teste de tamanho 25. A menor correlação entre as estimativas e seu valor real foi do método de Máxima Informação (0,9828).

Como esperado, proficiências situadas nos extremos da distribuição ($|\theta| \geq -3{,}50$) foram as que obtiveram pior desempenho pelos três métodos de seleção de itens, uma vez que existem poucos itens do banco com parâmetro de dificuldade nessa faixa de valores.

Vale destacar que para se gerar as 500 simulações de CAT pelo método de seleção MI, o algoritmo implementado no *software* R levou cerca de 50 minutos. O método de seleção adaptativa KL levou cerca uma hora e 25 minutos enquanto o método MIE, uma hora e 34 minutos.

As tabelas subseqüentes resumem a lógica do procedimento adaptativo. Serão apresentados os itens na ordem em que foram administrados no CAT, bem como os parâmetros dos itens pela TRI (Par.a, Par.b e Par.c), a resposta do examinando, as estimativas iterativas da proficiências e erros-padrão associados.

Três exemplos de níveis de proficiências ($\theta = -1{,}50$, $\theta = 0{,}00$ e $\theta = 1{,}50$) foram simulados utilizando o critério de Máxima Informação de Fisher para seleção iterativa dos itens. Para cada simulação, o valor inicial da proficiência foi de $\hat{\theta}_0 = 0{,}00$. O item que inicializou todos os algoritmos foi o ING10836, cujo parâmetro b é de $-2{,}51$.

Tabela 4.3: Teste adaptativo simulado para $\theta = -1{,}50$ para o método MI.

Ordem	**Item**	**Par.a**	**Par.b**	**Par.c**	**Resposta**	$\hat{\theta}$	**Erro-padrão**
1	ING10836	0,80	-2,51	0,10	0	-4,00	1,36
2	ING10833	0,88	-3,56	0,11	1	-3,18	1,36
3	ING10834	0,97	-2,40	0,11	1	-2,30	1,29
4	ING20626	1,05	-1,93	0,04	0	-2,94	1,04
5	ING10701	0,92	-2,32	0,12	1	-2,44	0,89
6	ING10622	0,93	-1,69	0,04	1	-1,92	0,73
7	ING10614	1,12	-1,58	0,09	1	-1,57	0,67
8	ING20627	1,68	-1,03	0,03	0	-1,79	0,56
9	ING20625	1,20	-1,17	0,04	0	-1,92	0,54
10	ING20611	0,96	-1,58	0,04	0	-2,07	0,54
11	ING10831	0,91	-1,73	0,10	1	-1,89	0,47
12	ING10624	1,00	-1,50	0,10	1	-1,74	0,43
13	ING20705	1,15	-1,38	0,17	1	-1,61	0,41
14	ING10613	1,17	-1,14	0,08	0	-1,70	0,40
15	ING10547	0,89	-1,61	0,05	1	-1,60	0,37
16	ING20622	1,29	-0,91	0,04	0	-1,65	0,36
17	ING20807	1,08	-1,24	0,11	0	-1,72	0,36
18	ING20605	0,87	-1,42	0,04	1	-1,63	0,33
19	ING20701	1,07	-1,12	0,11	0	-1,68	0,33
20	ING10839	0,90	-1,68	0,11	1	-1,62	0,31
21	ING10724	0,91	-1,51	0,11	1	-1,56	0,30
22	ING10937	0,97	-1,31	0,12	1	-1,50	0,29
23	ING10634	1,28	-0,91	0,11	0	-1,54	0,29
24	ING10628	0,91	-1,08	0,03	0	-1,58	0,29
25	ING10629	0,90	-1,10	0,04	1	-1,51	0,27

Por meio da Tabela 4.3 verificou-se que ao responder incorretamente a questão ING10836, a estimativa da proficiência que era $\hat{\theta}_0 = 0{,}00$ foi para $\hat{\theta}_1 = -4{,}00$, com erro-padrão igual a 1,36. Vale lembrar que a medida de erro-padrão associada a estimativa θ é a raiz da variância *a posteriori* (PSD).

Após o cálculo da medida de Informação de Fisher para $\hat{\theta}_1 = -4{,}00$, o algoritmo MI selecionou o segundo item do teste, ING10833. Para esse item, o examinando simulado respondeu corretamente, o que acarretou um aumento na estimativa de sua proficiência ($\hat{\theta}_1 = -3{,}18$). O procedimento iterativo foi realizado até o 25º item. O valor da estimativa

ao final desse processo foi de $\hat{\theta}_{25} = -1{,}51$, próximo ao seu valor verdadeiro $\theta = -1{,}50$.

É interessante notar que, dos 25 itens aplicados, o indivíduo com esse nível de proficiência acertou 56% dos itens e obteve uma estimativa de proficiência com precisão igual ou superior ao teste tradicional de 50 itens.

Tabela 4.4: Teste adaptativo simulado para $\theta = 0{,}00$ para o método MI.

Ordem	**Item**	**Par.a**	**Par.b**	**Par.c**	**Resposta**	$\hat{\theta}$	**Erro-padrão**
1	ING10836	0,80	-2,51	0,10	1	1,42	2,81
2	ING20725	1,15	1,31	0,05	0	-0,99	1,80
3	ING20627	1,68	-1,03	0,03	1	0,21	1,12
4	ING20618	1,46	0,17	0,04	0	-0,53	0,88
5	ING10633	1,28	-0,52	0,09	1	-0,16	0,67
6	ING10632	1,33	-0,19	0,10	0	-0,50	0,62
7	ING20622	1,29	-0,91	0,04	1	-0,32	0,48
8	ING10532	1,29	-0,40	0,12	0	-0,54	0,44
9	ING10634	1,28	-0,91	0,11	1	-0,44	0,40
10	ING10719	1,13	-0,31	0,08	1	-0,31	0,38
11	ING10625	1,37	0,19	0,10	0	-0,39	0,36
12	ING10601	1,06	-0,65	0,09	1	-0,32	0,35
13	ING10639	1,09	-0,59	0,15	1	-0,26	0,34
14	ING20519	0,99	-0,29	0,09	1	-0,19	0,33
15	ING10941	1,03	-0,32	0,13	1	-0,13	0,32
16	ING20501	0,95	-0,27	0,11	1	-0,07	0,32
17	ING20610	1,00	0,45	0,03	0	-0,12	0,31
18	ING10936	1,05	-0,74	0,10	1	-0,08	0,30
19	ING20614	1,04	-0,84	0,04	1	-0,05	0,29
20	ING10816	0,90	-0,57	0,07	1	-0,02	0,29
21	ING20510	0,96	0,10	0,16	1	0,04	0,29
22	ING10641	0,88	0,07	0,10	0	-0,02	0,28
23	ING10636	0,85	-0,06	0,09	1	0,02	0,28
24	ING20545	0,86	-0,26	0,11	1	0,06	0,28
25	ING10930	0,88	0,26	0,11	0	0,01	0,27

A Tabela 4.4 apresenta um teste adaptativo quando $\theta = 0{,}00$. Como esperado, o examinando nesse nível de proficiência respondeu corretamente ao item ING10836, cujo parâmetro de dificuldade é de $-2{,}51$. Com o acerto, a medida de proficiência do examinando foi para $\hat{\theta}_1 = 1{,}42$ com erro-padrão igual a 2,81. O segundo item (com dificuldade superior ao primeiro, $b = 1{,}31$) foi apresentado ao indivíduo. Como o examinando errou, a estimativa da proficiência sofreu um decréscimo. Após os 25 itens do teste adaptativo, observou-se uma estimativa de proficiência de $\hat{\theta}_{25} = 0{,}01$ com erro igual a 0,27.

Para o teste simulado, o examinando com $\theta = 0{,}00$ acertou 68% da prova adaptativa

de Proficiência em Inglês Instrumental I.

Tabela 4.5: Teste adaptativo simulado para $\theta = 1{,}50$ para o método MI.

Ordem	**Item**	**Par.a**	**Par.b**	**Par.c**	**Resposta**	$\hat{\theta}$	**Erro-padrão**
1	ING10836	0,80	-2,51	0,10	1	1,42	2,81
2	ING20725	1,15	1,31	0,05	0	-0,99	1,80
3	ING20627	1,68	-1,03	0,03	1	0,21	1,12
4	ING20618	1,46	0,17	0,04	1	0,86	0,89
5	ING20813	1,35	0,69	0,20	1	1,18	0,83
6	ING10527	1,15	1,21	0,12	1	1,53	0,79
7	ING20642	0,94	2,00	0,03	0	1,26	0,66
8	ING10631	0,95	0,79	0,07	1	1,44	0,61
9	ING20512	0,96	1,54	0,17	0	1,21	0,54
10	ING20610	1,00	0,45	0,03	1	1,31	0,50
11	ING20511	0,83	1,11	0,08	0	1,13	0,47
12	ING10611	0,87	0,77	0,10	1	1,23	0,45
13	ING20544	0,86	1,08	0,13	0	1,08	0,43
14	ING10625	1,37	0,19	0,10	1	1,13	0,41
15	ING10640	0,83	0,74	0,05	1	1,21	0,40
16	ING10940	0,76	0,85	0,08	1	1,27	0,39
17	ING10506	0,82	1,54	0,15	1	1,36	0,39
18	ING20635	0,81	0,44	0,04	1	1,40	0,39
19	ING10806	0,76	0,88	0,13	1	1,45	0,38
20	ING20831	0,85	2,09	0,12	0	1,39	0,37
21	ING10635	0,76	0,58	0,08	1	1,43	0,36
22	ING20804	0,79	0,55	0,11	1	1,46	0,36
23	ING10545	0,82	0,67	0,21	1	1,50	0,35
24	ING20532	1,03	2,57	0,04	0	1,47	0,34
25	ING20525	0,69	0,82	0,08	1	1,50	0,34

Para um examinando simulado com proficiência igual a $\theta = 1{,}50$, obteve-se uma estimativa final de $\hat{\theta}_{25} = 1{,}50$, com erro de 0,34. Esse indivíduo acertou 72% dos 25 itens do CAT. Os resultados referentes ao processo iterativo para essa simulação estão na Tabela 4.5.

No Anexo desta dissertação, tem-se a reprodução das Tabelas 4.3-4.5 para os métodos de seleção de itens Kullback-Leibler e Máxima Informação Esperada.

4.2.3 Estudo de simulação 3

O estudo desta seção pretende retratar a diferença entre os métodos de seleção adaptativa para diferentes valores iniciais da proficiência. Dessa maneira, três valores foram fixados para inicialização do algoritmo. O primeiro com $\hat{\theta}_0 = -1{,}50$, o segundo, $\hat{\theta}_0 = 0{,}00$ e o terceiro, $\hat{\theta}_0 = 1{,}50$. Como ilustrado na Figura 4.2, esse estudo simulado teve por objetivo avaliar se, inicializando o teste com diferentes estimativas das proficiências, os métodos de seleção adaptativa de itens convergiriam da mesma forma ou se algum apresentaria melhor desempenho do que os restantes.

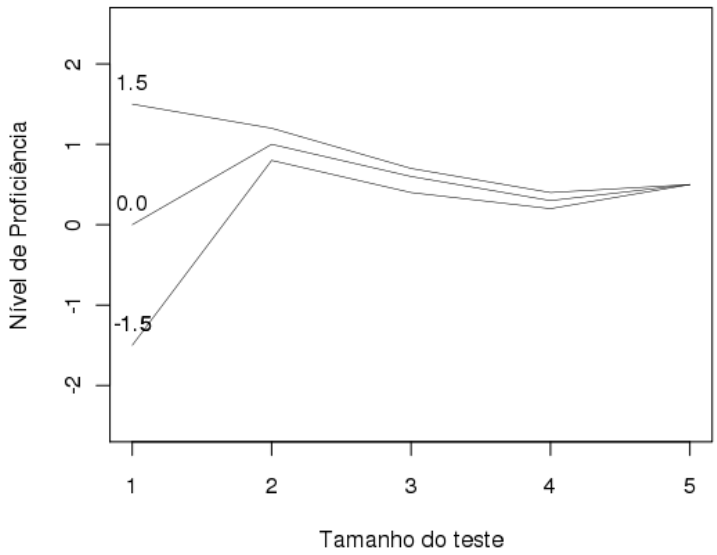


Figura 4.2: Exemplificação da simulação 3.

Para tanto, três diferentes valores de θ foram gerados para a simulação: $\theta = -1{,}50$, $\theta = 0{,}00$ e $\theta = 1{,}50$. Cada proficiência simulada θ foi replicada 100 vezes para se avaliar a variabilidade das estimativas, haja vista que indivíduos com mesma proficiência podem ter diferentes padrões de resposta e, conseqüentemente, obter diferentes estimativas para a proficiência. O tamanho máximo do teste foi de 25 para todos os examinandos. Os resultados desse estudo podem ser visualizados nas Figuras 4.3 a 4.5.

A Figura 4.3 ilustra como as estimativas da proficiência de indivíduos com $\theta = -1{,}50$ estão distribuídas para diferentes valores iniciais $\hat{\theta}_0$. Para $\hat{\theta}_0 = -1{,}50$, observou-se que, apesar de se ter fixado um valor inicial para a proficiência dos examinandos exatamente igual ao valor real θ, encontrou-se estimativas variando de aproximadamente $-1,77$ a $-1,35$. De uma maneira geral, pode-se dizer que os três métodos apresentaram desempenhos próximos para $\hat{\theta}_0 = -1{,}50$ com 50% das medidas de erros-padrão menores ou

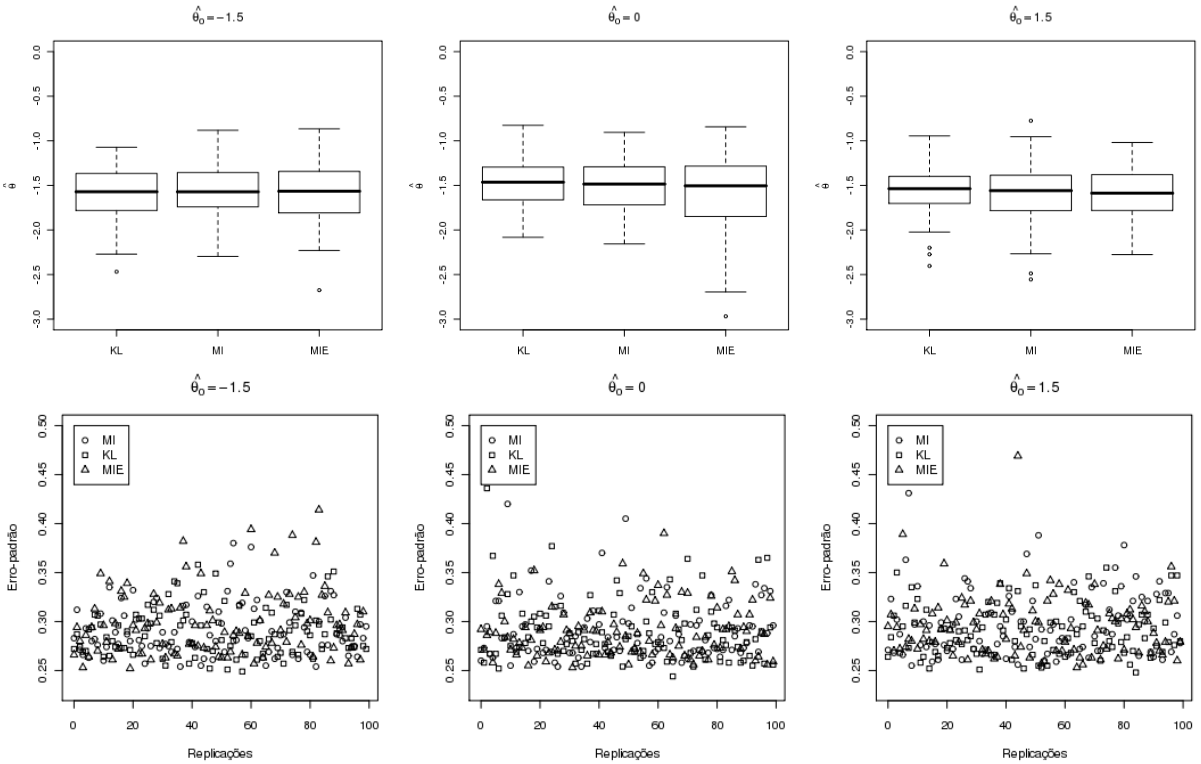


Figura 4.3: Distribuição das estimativas quando $\theta = -1{,}50$ para diferentes $\hat{\theta}_0$.

iguais a 0,28.

Quando o valor inicial da proficiência foi de $\hat{\theta}_0 = 0{,}00$, observou-se que o método de Máxima Informação Esperada foi o que apresentou maior amplitude em relação aos valores das estimativas (menor valor de $\hat{\theta}$ foi de $-2{,}97$ e o maior foi de $-1{,}28$). Em geral, os valores dos erros-padrão na estimação de $\theta = -1{,}50$ para os três métodos variaram de 0,24 a 0,30.

Em relação a $\hat{\theta}_0 = 1{,}50$, observou-se que os três critérios de seleção de itens apresentaram estimativas próximas do seu valor verdadeiro mesmo inicializando o algoritmo com uma proficiência bem acima do real. As medidas de erros-padrão nesta simulação oscilaram, em média, entre 0,25 e 0,32.

A seguir, serão apresentados os gráficos referentes à estimação de $\theta = 0{,}00$ e $\theta = 1{,}50$ utilizando os mesmos valores $\hat{\theta}_0$ para inicialização do algoritmo.

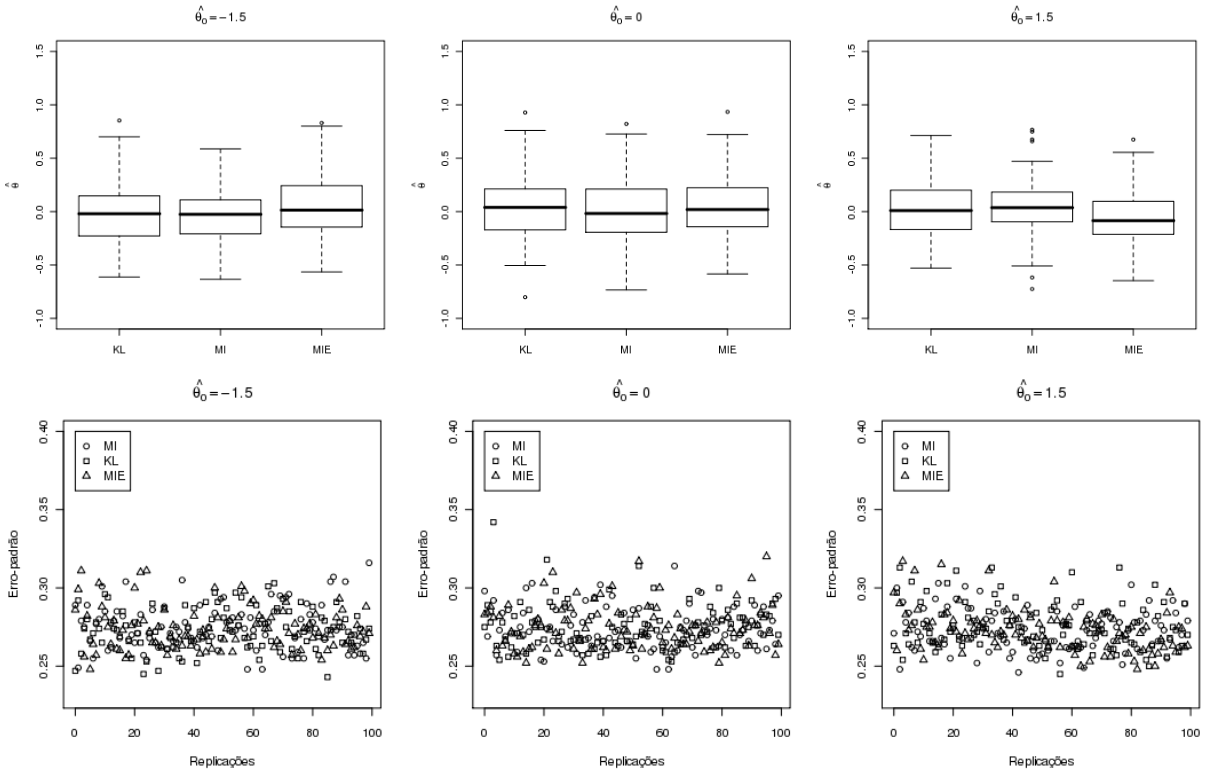


Figura 4.4: Distribuição das estimativas quando $\theta = 0{,}00$ para diferentes $\hat{\theta}_0$.

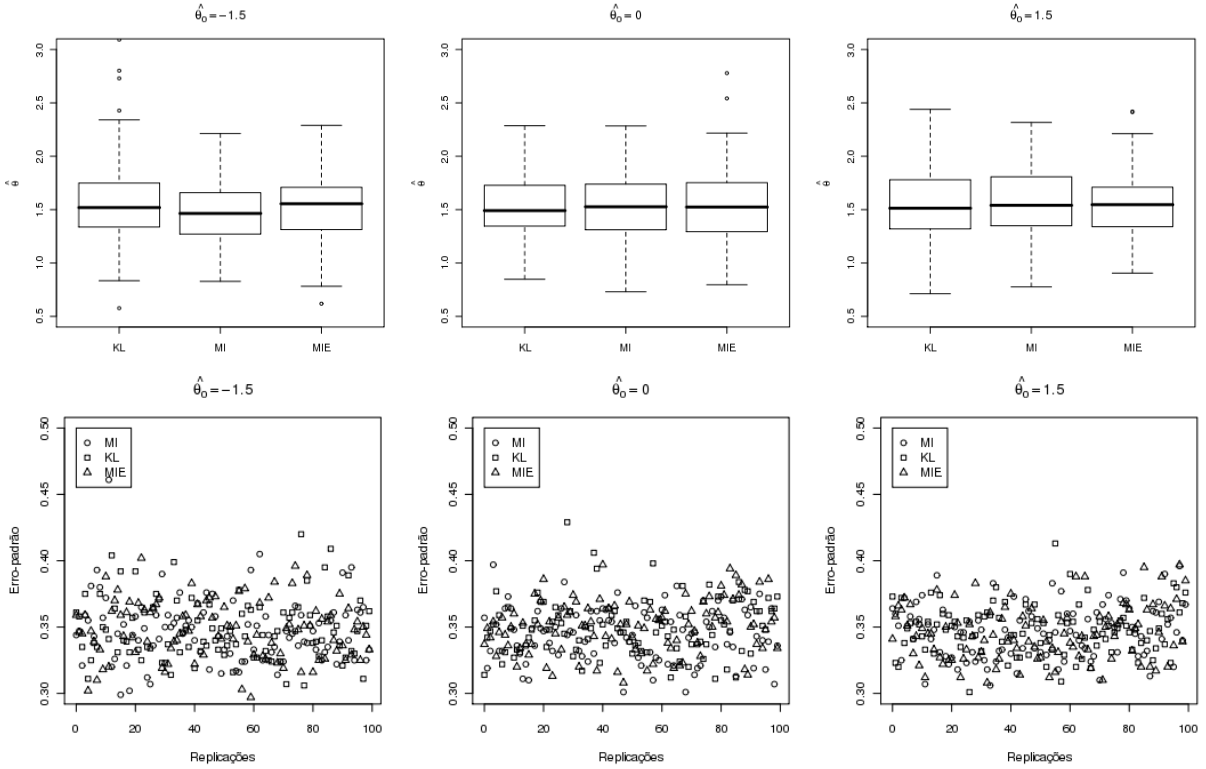


Figura 4.5: Distribuição das estimativas quando $\theta = 1{,}50$ para diferentes $\hat{\theta}_0$.

Por meio das Figuras 4.4 e 4.5, observa-se que os métodos de seleção adaptiva conseguem estimar eficientemente diferentes níveis de proficiência em Inglês Instrumental I,

independentemente do valor inicial atribuído ao algoritmo para a estimação das proficiências desses examinandos. Para todas as simulações desse estudo, verificou-se que o desempenho entre os métodos de seleção adaptativa foi bem próximo quando o tamanho de teste foi fixado em 25.

4.2.4 Estudo de simulação 4

Neste estudo, pretende-se avaliar a qualidade da medida de proficiência para diferentes parâmetros de dificuldade para o primeiro item no CAT. Três diferentes valores de θ foram usados na simulação: $\theta = -1{,}50$, $\theta = 0{,}00$ e $\theta = 1{,}50$. Cada proficiência simulada θ foi replicada 100 vezes. O tamanho máximo do teste foi de 25 para todos os examinandos.

O valor inicial da proficiência dos examinandos foi de $\hat{\theta}_0 = 0{,}00$. Dessa forma, três análises foram feitas: a primeira, fixando o primeiro item com parâmetro de dificuldade bem abaixo dos verdadeiros valores de θ. A segunda, fixando o primeiro item com valor do parâmetro b próximo aos valores reais e o terceiro caso, fixando o primeiro item com parâmetro b acima das efetivas proficiências. Os itens selecionados em cada análise estão apresentados na Tabela 4.6.

Tabela 4.6: Parâmetros dos itens iniciais na simulação 4.

Análise	Item	Par.a	Par.b	Par.c
Par.$b<\theta$	ING10836	0,80	-2,51	0,10
Par.$b\approx\theta$	ING10706	0,60	-1,50	0,12
	ING20602	0,61	0,00	0,04
	ING10516	0,59	1,43	0,19
Par.$b>\theta$	ING20532	1,03	2,57	0,04

Para $\theta = -1{,}50$, observou-se que tanto as estimativas como os erros associados às proficiências de cada método de seleção de itens (MI, KL e MIE) tiveram desempenhos próximos quando um item de dificuldade variada é apresentado como a primeira questão do teste. Para qualquer valor do parâmetro de dificuldade do primeiro item, observou-se que a medida de erro associado às proficiências foi, em média, 0,29 após 25 itens. Quanto às medidas de proficiências, observou-se também que, para qualquer um dos itens iniciais da análise, os métodos de seleção de itens apresentaram valores próximos ao real. As distribuições das estimativas das proficiências e dos erros-padrão para $\theta = -1{,}50$ estão apresentadas na Figura 4.6.

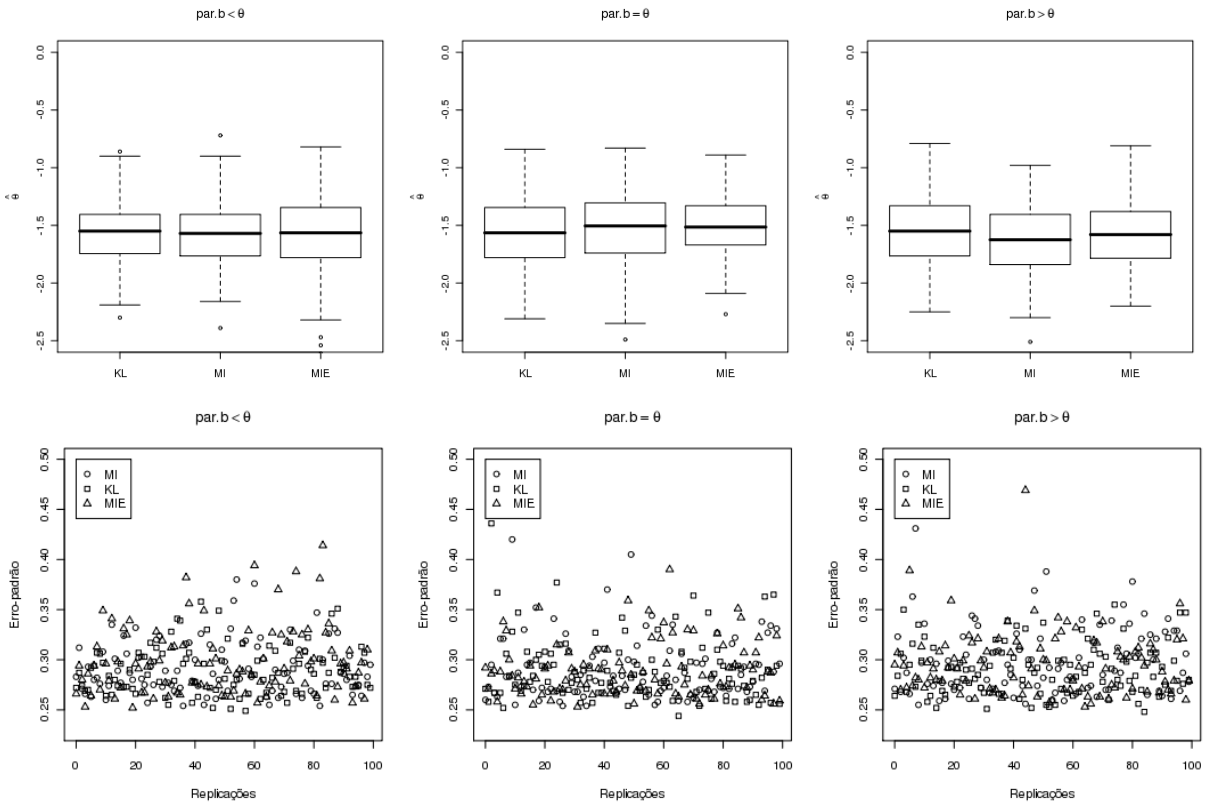


Figura 4.6: Distribuição das estimativas quando $\theta = -1{,}50$ para diferentes itens iniciais.

Em relação a $\theta = 0{,}00$, também se observou que tanto as estimativas como os erros associados às proficiências de cada método de seleção de itens (MI, KL e MIE) tiveram desempenhos próximos quando um item de dificuldade variada é apresentado como a primeira questão do teste. As distribuições das estimativas das proficiências e dos erros-padrão nesta simulação estão apresentadas na Figura 4.7.

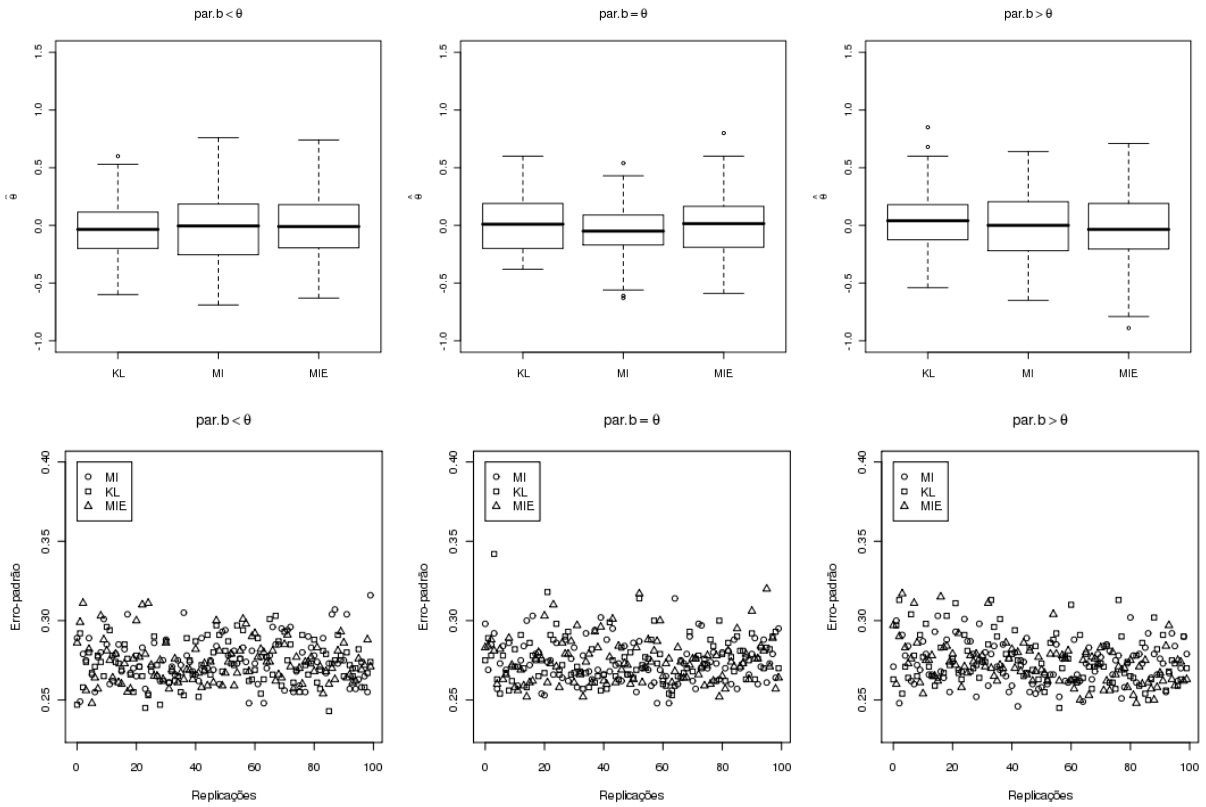


Figura 4.7: Distribuição das estimativas quando $\theta = 0{,}00$ para diferentes itens iniciais.

A Figura 4.8 apresenta as distribuições das estimativas e dos erros-padrão para os diferentes métodos de seleção de itens quando $\theta = 1{,}0$. Como nos casos já expostos, pode-se concluir que, independentemente do parâmetro de dificuldade do item inicial, os três métodos de seleção adaptativa conseguiram estimar as proficiência em Inglês Instrumental I com praticamente a mesma precisão para um teste com 25 itens.

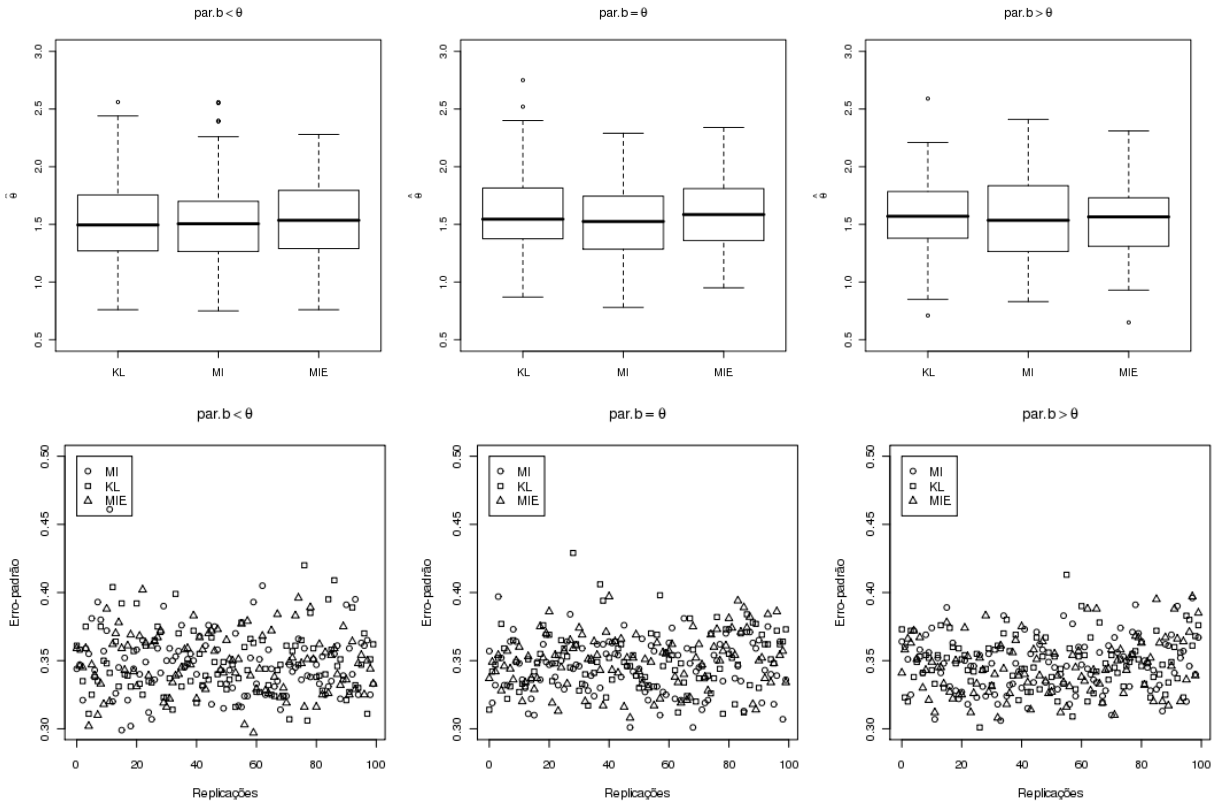


Figura 4.8: Distribuição da estimativas quando $\theta = 1{,}50$ para diferentes itens iniciais.

4.2.5 Estudo de simulação 5

O procedimento de simulação utilizado neste estudo pretende avaliar as medidas de *Viés* e *Erro Quadrático Médio* (EQM) associados a 10 diferentes valores de proficiências: $\theta = -3{,}50$, $\theta = -2{,}50$, $\theta = -1{,}50$, $\theta = -0{,}50$, $\theta = 0{,}00$, $\theta = 0{,}50$, $\theta = 1{,}50$, $\theta = 2{,}50$, $\theta = 3{,}50$ e $\theta = 4{,}00$. Foram utilizadas 100 replicações para cada θ. As proficiências foram estimadas até que a quantidade total de itens no teste fosse igual a 30.

O Viés e o Erro Quadrático Médio das estimativas foram calculados quando o tamanho do teste (n) era igual 1, 2 até 30 itens da seguinte maneira:

$$\text{Viés}_n = \frac{\sum_{j=1}^{100} \hat{\theta}_{j,n}}{100} - \theta, \quad \text{n} = \{1, \ldots, 30\}.$$

$$\text{EQM}_n = \frac{\sum_{j=1}^{100} (\hat{\theta}_{j,n} - \theta)^2}{100}, \quad \text{n} = \{1, \ldots, 30\}.$$

Por meio da Figura 4.9 observou-se que à medida que o tamanho do teste aumenta, o viés diminui para todos os três métodos de seleção de itens.

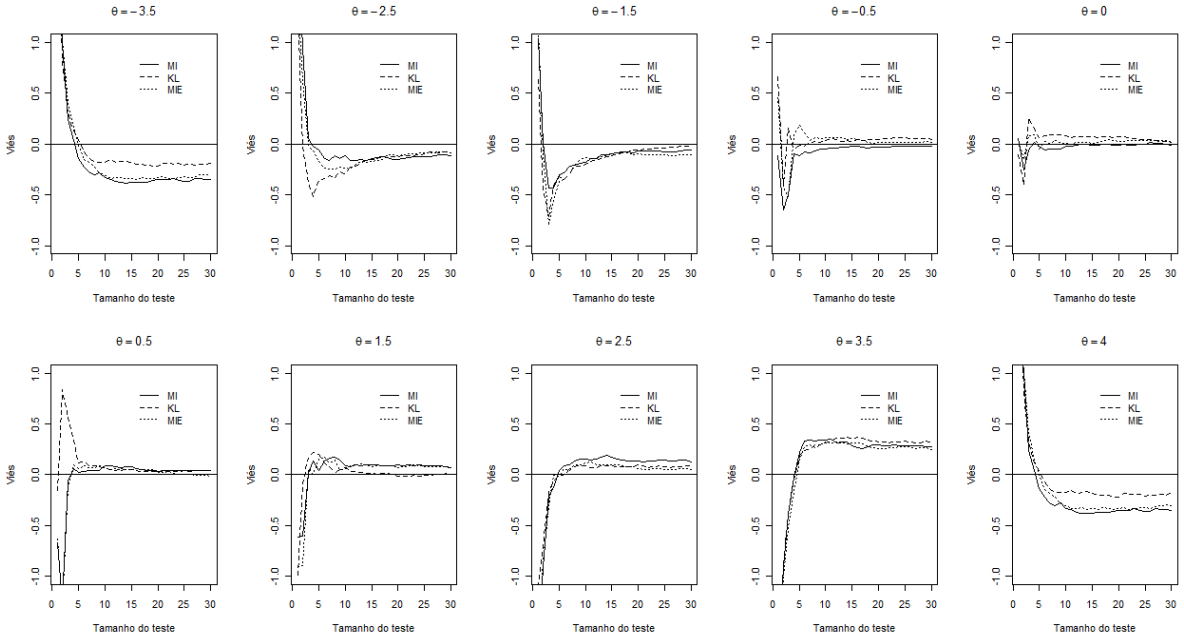


Figura 4.9: Viés das estimativas.

Como também era esperado, o EQM das estimativas diminui com o aumento da quantidade de itens no teste adaptativo.

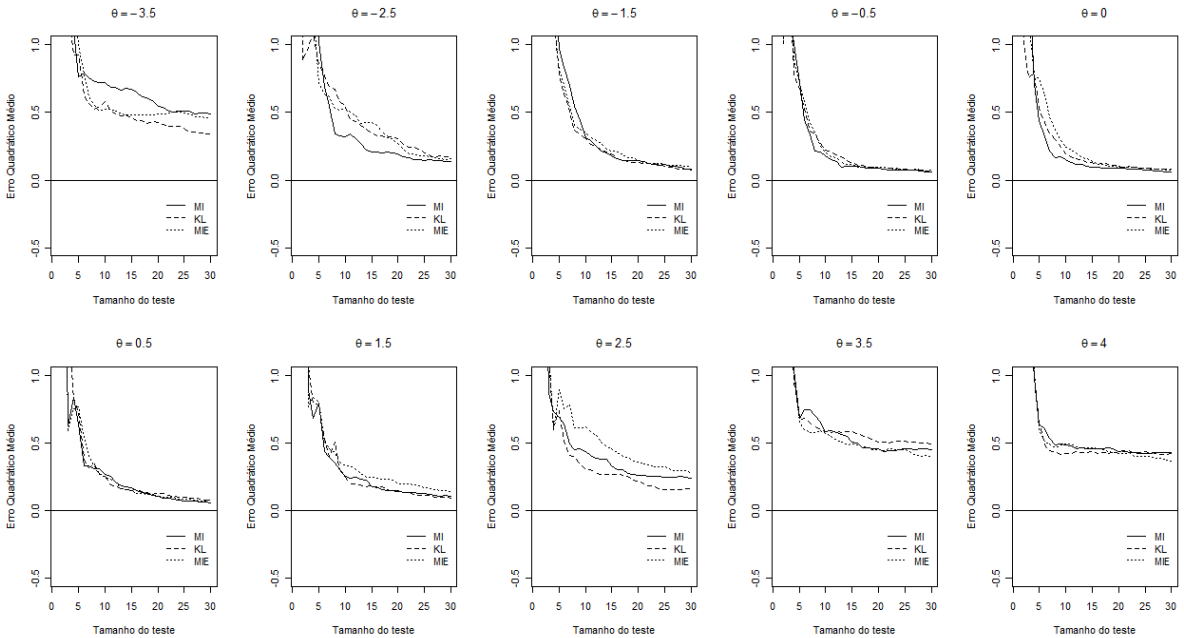


Figura 4.10: Erro Quadrático Médio das estimativas.

Capítulo5

Conclusões e trabalhos futuros

A possibilidade de se obter uma avaliação individualizada e que corresponda ao nível de capacidade dos examinandos fez crescer o número de pesquisas em Testes Adaptativos Informatizados (CAT). Diferentemente dos testes papel-e-caneta, os testes adaptativos administram itens adequados a cada examinando. Essa seleção baseia-se tanto na característica dos itens (parâmetro de dificuldade do item, discriminação) quanto na estimativa da proficiência do candidato. Para tanto, dois componentes são necessários: a construção de um banco de itens e o desenvolvimento de um algoritmo para seleção adaptativa dos itens. Esta dissertação se propôs a discutir os métodos estatísticos que envolvem essas duas componentes.

Na construção do banco foi apresentada a metodologia de estimação dos parâmetros dos itens quando se tem grupos múltiplos. Esse procedimento foi utilizado na calibração dos itens da prova de Proficiência em Inglês Instrumental I da Universidade de Brasília. Uma análise dessas questões por meio da correlação bisserial e pelas estimativas dos parâmetros da TRI foi necessária para se eliminar aqueles de baixa qualidade psicométrica. Posteriormente, os parâmetros dos itens foram transformados em uma nova métrica com média 100 e desvio 25. Nessa nova escala, os itens foram categorizados em níveis, pela técnica de Ancoragem, para que uma avaliação pedagógica pudesse ser feita com a finalidade de se dar um sentido qualitativo à medida de proficiência.

Após a análise do banco de itens, foram apresentados três métodos de seleção adaptativa em CAT. O primeiro deles referiu-se ao critério de Máxima Informação, um dos métodos mais populares para seleção de itens em CAT. O segundo método baseou-se no método da Informação Global definido por Chang e Ying (1996) utilizando a medida de

Kullback-Leibler. Já o terceiro fundamentou-se na análise preditiva definida pelo critério de Máxima Informação Esperada proposto por van der Linden (1998). Para avaliar os procedimentos, cinco tipos de simulações foram feitas para o banco de Proficiência em Inglês Instrumental I. As proficiências iterativas dos examinandos foram estimadas pelo método da *Média a Posteriori* (EAP).

O estudo simulado 1 teve por foco apresentar a quantidade de itens necessários para se obter diferentes níveis de erro-padrão associado às proficiências dos examinandos em CAT. Observou-se que, dentre os três métodos de seleção adaptativa analisados, o de Máxima Informação Esperada foi o que apresentou, em média, menor número de itens (22) para um erro-padrão igual a 0,40 (usualmente encontrado nos testes papel-e-caneta). Já para um erro de 0,20, o método de Máxima Informaçaõ foi o que apresentou menor quantidade de itens do banco de itens de Inglês Instrumental I.

O estudo simulado 2 objetivou avaliar o grau de correlação existente nas estimativas geradas dos testes adaptativos (utilizando os três métodos de seleção de itens) com seu valor verdadeiro. Nesse estudo, observou-se uma alta correlação entre essas variáveis e que ambos os métodos apresentaram tal tendência.

As simulações do estudo 3 foram delineadas para se comparar o desempenho dos critérios de seleção dos itens para diferentes valores iniciais das proficiências. Os resultados demonstraram que, independentemente do valor inicial da proficiência, a precisão das estimativas foi praticamente a mesma para os três métodos.

No estudo simulado 4, avaliou-se diferentes itens iniciais com parâmetros de dificuldade variados para inicialização dos algoritmos. Observou-se, entretanto, que, independentemente do parâmetro de dificuldade do primeiro item do teste, mediu-se com praticamente a mesma precisão as proficiências dos examinandos simulados para os três métodos de seleção adaptativa.

O estudo 5 apresentou os resultados de uma simulação para 10 valores diferentes de proficiência e avaliou-se a medida de Viés e Erro Quadrático Médio das estimativas ao longo de um teste de 30 itens. Como esperado, essas medidas aproximam-se de zero à medida que o tamanho do teste aumenta. Vale ainda destacar que os algoritmos não conseguiram estimar eficientemente as proficiências dos indivíduos situados fora do intervalo $(-2{,}50\ ;\ 2{,}50)$. Isto é justificável, uma vez que o banco de questões de Proficiência em Inglês Instrumental I possui poucos itens com parâmetro de dificuldade fora dessa faixa de valores.

Por meio das simulações, pode-se concluir que os três métodos adaptativos avaliados apresentaram desempenho semelhante, não tendo nenhum com performance superior aos demais ao se utilizar o banco de Inglês Instrumental I. Como era desejado, esses métodos conseguiram reduzir substancialmente a quantidade de itens para um teste de Inglês sem comprometer a precisão das estimativas das proficiências. Contudo, é importante salientar que o CAT é extremamente sensível às características (parâmetros) dos itens do banco. Por isso, tomou-se o devido cuidado para descrever toda a metodologia para estimação dos parâmetros dos itens (Capítulo 2), uma vez que uma boa testagem adaptativa só ocorrerá se os itens do banco tiverem boa qualidade.

Acredita-se que este estudo apresentou boas contribuições na área dos testes adaptativos informatizados, já que foram apresentados a metodologia de construção do banco de itens, exposição teórica do modelo da Teoria de Resposta ao Item e métodos que regem o comportamento dos testes adaptativos. Além disso, avaliou-se a adequabilidade prática do banco de itens de Proficiência em Inglês Instrumental I da UnB para implementação dessa tecnologia. Sem dúvidas, as análises psicométricas e pedagógicas realizadas com esses dados impulsionarão novas discussões e esclarecimentos sobre a metodologia aplicada à certificação desses estudantes.

Assim, ao final da pesquisa acredita-se ter alcançado o objetivo: o domínio da teoria estatística no desenvolvimento do CAT. Espera-se que este trabalho ajude a impulsionar o desenvolvimento dessa técnica no Brasil, pois certamente contribuirá para o progresso das avaliações educacionais.

Em futuros trabalhos, pretende-se aperfeiçoar os algoritmos de seleção dos itens; incorporar restrições ao procedimento de seleção de itens, tais como: balanceamento do conteúdo, procedimentos de controle da exposição dos itens (algoritmos como o de Sympson-Hetter apresentado no Capítulo 3), *testlets*; avaliar as estimativas das proficiências utilizando outros métodos de estimação como MCMC (*Markov Chain Monte Carlo*); estudar e propor critérios de decisão para a parada dos algoritmos e, por fim, incorporar a estimação dos parâmetros de novos itens ao próprio algoritmo adaptativo.

Apêndice

Informação Observada e Esperada

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

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

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

Logaritmo da Verossimilhança: $\ln L(\theta; u_i) = u_i \ln P_i(\theta) + (1 - u_i) \ln [1 - P_i(\theta)]$

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

$$J_{u_i}(\theta) = -\frac{\partial^2}{\partial\theta^2} \ln L(\theta; u_i) = -\frac{u_i P_i''(\theta)}{P_i(\theta)} + \frac{u_i [P_i'(\theta)]^2}{P_i^2(\theta)} - \frac{P_i''(\theta)[u_i - 1]}{1 - P_i(\theta)} - \frac{[P_i'(\theta)]^2[u_i - 1]}{[1 - P_i(\theta)]^2}.$$

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

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

Como $U_i \sim Bernoulli(P_i)$, a Informação de Fisher do j-ésimo item será igual a:

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

Sob o Modelo Logístico de três Parâmetros, tem-se que:

$$P_i'(\theta) = \frac{Da_i(1-c_i)e^{-Dai(\theta-bi)}}{(1+e^{-Dai(\theta-bi)})^2} \quad \text{e} \quad P_i''(\theta) = \frac{D^2a_i^2(1-c_i)e^{Dai(\theta-bi)}\left(1-e^{Dai(\theta-bi)}\right)}{(1+e^{Dai(\theta-bi)})^3}.$$

Logo, a Informação Observada será igual a:

$$J_{u_i}(\theta) = \left[\frac{a_i^2(1-c_i)^2\exp\left[-2Dai\left(\theta-bi\right)\right]}{(1+\exp\left[-ai\left(\theta-bi\right)\right])^4}\right]\left[\frac{1+\exp\left[-ai\left(\theta-bi\right)\right]}{c_i(1+\exp\left[-ai\left(\theta-bi\right)\right])+(1-c_i)}\right]$$

Enquanto a Informação de Fisher do item para esse modelo será denotada por:

$$\begin{aligned}
I_{U_i}(\theta) &= \left[\frac{a_i^2(1-c_i)^2\exp\left[-2Dai\left(\theta-bi\right)\right]}{(1+\exp\left[-ai\left(\theta-bi\right)\right])^4}\right]\left[\frac{1+\exp\left[-ai\left(\theta-bi\right)\right]}{c_i(1+\exp\left[-ai\left(\theta-bi\right)\right])+(1-c_i)}\right]\cdot \\
&\cdot\left[\frac{1+\exp\left[-ai\left(\theta-bi\right)\right]}{1+\exp\left[-ai\left(\theta-bi\right)\right]-c_i(1+\exp\left[-ai\left(\theta-bi\right)\right])-(1-c_i)}\right] \\
&= \left[\frac{a_i^2(1-c_i)^2\exp\left[-2ai\left(\theta-bi\right)\right]}{(1+\exp\left[-ai\left(\theta-bi\right)\right])^2}\right]\left[\frac{1}{c_i\exp\left[-ai\left(\theta-bi\right)\right])+1}\right]\left[\frac{1}{(1-c_i)\exp\left[-ai\left(\theta-bi\right)\right]}\right] \\
&= \left[\frac{a_i^2}{(1+\exp\left[-ai\left(\theta-bi\right)\right])^2}\right]\left[\frac{(1-c_i)\exp\left[-ai\left(\theta-bi\right)\right]}{1+c_i\exp\left[-Dai\left(\theta-bi\right)\right]}\right]\left[\frac{(1-c_i)\exp\left[-Dai\left(\theta-bi\right)\right]}{(1-c_i)\exp\left[-Dai\left(\theta-bi\right)\right]}\right] \\
&= \frac{a_i^2(1-c_i)}{\{1+\exp\left[-ai\left(\theta-bi\right)\right]\}^2\{c_i+\exp\left[ai\left(\theta-bi\right)\right]\}}.
\end{aligned}$$

Medida de Kullback-Leibler

Como definido no Capítulo 3, a medida de Informação Kullback-Leibler pode ser expressa por:

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

Para o Modelo Logístico de três Parâmetros, tem-se:

$$\begin{aligned}
K_i(\theta||\theta_0) = &\left[\frac{1+c_ie^{-Da_i(\theta_0-b_i)}}{1+e^{-Da_i(\theta_0-b_i)}}\right]\cdot\log\left[\frac{(1+c_ie^{-Da_i(\theta_0-b_i)})(1+e^{-Da_i(\theta-b_i)})}{(1+e^{-Da_i(\theta_0-b_i)})(1+c_ie^{-Da_i(\theta-b_i)})}\right] \\
&+\left[\frac{(1-c_i)e^{-Da_i(\theta_0-b_i)}}{1+e^{-Da_i(\theta_0-b_i)}}\right]\cdot\log\left[\frac{(1-c_i)e^{-Da_i(\theta-b_i)}(1+e^{-Da_i(\theta-b_i)})}{(1-c_i)e^{-Da_i(\theta-b_i)}(1+e^{-Da_i(\theta_0-b_i)})}\right]
\end{aligned}$$

Como $P(\theta) = \frac{1+c_ie^{-Da_i(\theta-b_i)}}{1+e^{-Da_i(\theta-b_i)}}$ e $1-P(\theta) = \frac{(1-c_i)e^{-Da_i(\theta-b_i)}}{1+e^{-Da_i(\theta-b_i)}}$, pode-se simplificar a medida de KL por:

$$\begin{aligned}
K_i(\theta||\theta_0) = &\left[\frac{c_ie^{Da_i(\theta_0-b_i)}}{1+e^{Da_i(\theta_0-b_i)}}\right]\cdot\log\left[\frac{c_ie^{Da_i(\theta_0-b_i)}}{(1+e^{Da_i(\theta_0-b_i)})P(\theta)}\right] \\
&+\left[\frac{1-c_i}{1+e^{Da_i(\theta_0-b_i)}}\right]\cdot\log\left[\frac{(1-c_i)}{(1+e^{Da_i(\theta_0-b_i)}(1-P(\theta))}\right].
\end{aligned}$$

Anexo

As Figuras de 5.1 a 5.3 referem-se ao material utilizado na oficina de interpretação da escala de Proficiência em Inglês Instrumental I discutido no Capítulo 2, enquanto as Tabelas de 5.1 a 5.6 referem-se aos exemplos de três testes adaptativos para as proficiências $\theta = -1{,}50$, $\theta = 0{,}00$ e $\theta = 1{,}50$ sob os métodos de seleção de Kullback-Leibler e Máxima Informação Esperada comentadas no Capítulo 4.

Figura 5.1: Exemplo de um item apresentado na oficina da interpretação da escala.

cespeUnB
Centro de Seleção e de Promoção de Eventos

Universidade de Brasília

INGLÊS INSTRUMENTAL I– ITEM: ING10502* – NÍVEL: 125

While larger banks can afford to maintain their own data-processing operations, many smaller regional and community banks are finding that the cost associated with upgrading data-processing equipment and with the development and maintenance of new products and technical staff are prohibitive.

a) cost associated with

b) costs associated with

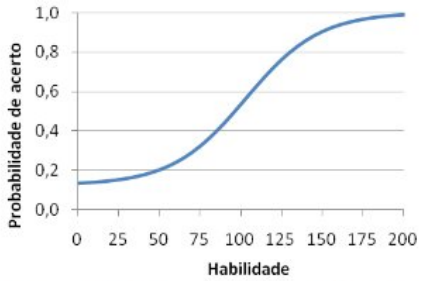
c) costs arising from

d) cost of

e) costs of

Estatísticas do item ING10502*

GAB	Percentual por alternativa					Teoria de Resposta ao Item		
	%A	*%B*	*%C*	*%D*	*%E*	*par.a*	*par.b*	*par.c*
B	6,07%	64,02%	8,88%	7,71%	12,15%	0,03	102,80	0,12



Curva Característica do item ING10502*

* Item meramente ilustrativo. Não há semelhanças com o item ING10502 do banco.

Figura 5.2: Distribuição dos itens nos níveis da escala de proficiência.

Figura 5.3: Distribuição dos itens nos níveis da escala de proficiência.

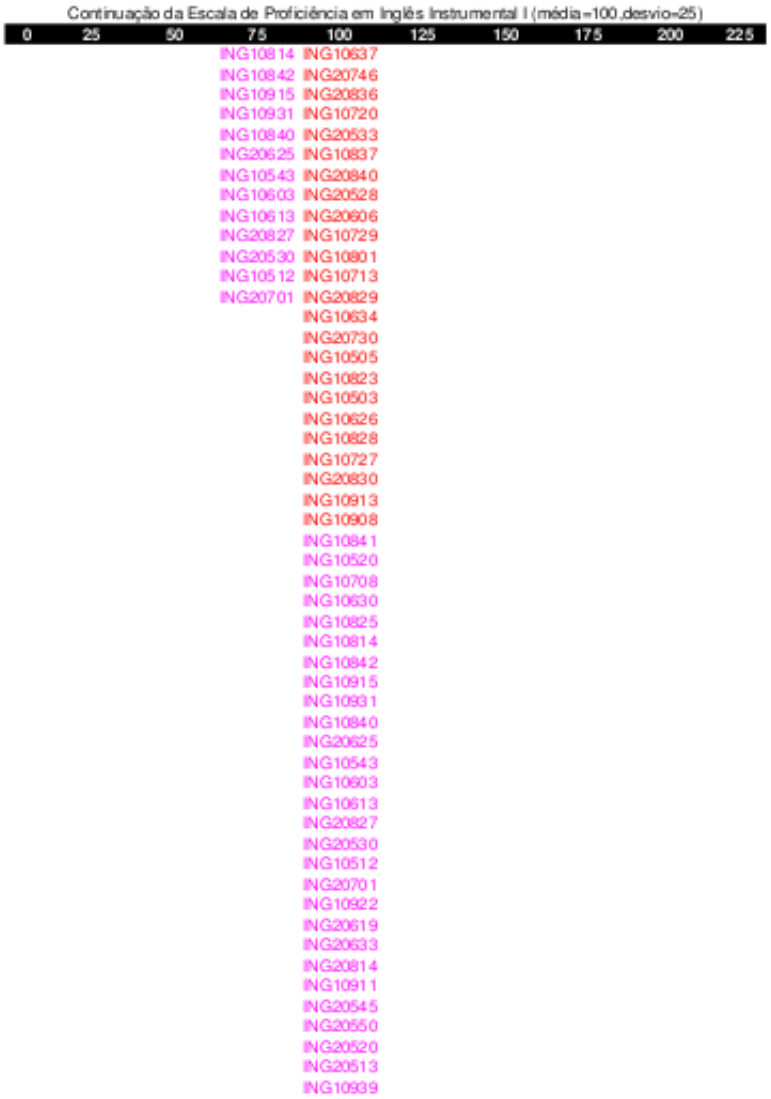


Tabela 5.1: Teste adaptativo simulado para $\theta = -1{,}50$ para o método KL.

Ordem	**Item**	**Par.*a***	**Par.*b***	**Par.*c***	**Resposta**	$\hat{\theta}$	**Erro-padrão**
1	ING10836	0,80	-2,51	0,10	1	1,42	2,81
2	ING20618	1,46	0,17	0,04	0	-1,68	1,54
3	ING20627	1,68	-1,03	0,03	0	-2,51	1,31
4	ING20626	1,05	-1,93	0,04	1	-1,64	0,83
5	ING20625	1,20	-1,17	0,04	0	-1,93	0,82
6	ING10614	1,12	-1,58	0,09	0	-2,23	0,84
7	ING10834	0,97	-2,40	0,11	1	-1,95	0,63
8	ING20611	0,96	-1,58	0,04	1	-1,68	0,50
9	ING20622	1,29	-0,91	0,04	0	-1,77	0,48
10	ING10622	0,93	-1,69	0,04	1	-1,62	0,43
11	ING10613	1,17	-1,14	0,08	1	-1,45	0,40
12	ING10634	1,28	-0,91	0,11	0	-1,53	0,39
13	ING20705	1,15	-1,38	0,17	0	-1,64	0,38
14	ING10624	1,00	-1,50	0,10	1	-1,54	0,35
15	ING20807	1,08	-1,24	0,11	1	-1,45	0,34
16	ING20701	1,07	-1,12	0,11	1	-1,36	0,32
17	ING20614	1,04	-0,84	0,04	0	-1,41	0,32
18	ING10628	0,91	-1,08	0,03	0	-1,46	0,32
19	ING20605	0,87	-1,42	0,04	0	-1,53	0,32
20	ING10547	0,89	-1,61	0,05	1	-1,47	0,30
21	ING10629	0,90	-1,10	0,04	0	-1,52	0,30
22	ING10937	0,97	-1,31	0,12	1	-1,46	0,29
23	ING10831	0,91	-1,73	0,10	1	-1,42	0,28
24	ING10512	1,00	-1,12	0,12	0	-1,46	0,28
25	ING20634	0,90	-1,03	0,04	0	-1,50	0,27

Tabela 5.2: Teste adaptativo simulado para $\theta = 0{,}00$ para o método KL.

Ordem	**Item**	**Par.** a	**Par.** b	**Par.** c	**Resposta**	$\hat{\theta}$	**Erro-padrão**
1	ING10836	0,80	-2,51	0,10	1	1,42	2,81
2	ING20618	1,46	0,17	0,04	0	-1,68	1,54
3	ING20627	1,68	-1,03	0,03	1	-0,45	0,92
4	ING20622	1,29	-0,91	0,04	1	-0,13	0,70
5	ING10632	1,33	-0,19	0,10	1	0,16	0,65
6	ING10625	1,37	0,19	0,10	1	0,44	0,63
7	ING10532	1,29	-0,40	0,12	1	0,55	0,61
8	ING20813	1,35	0,69	0,20	0	0,29	0,47
9	ING10633	1,28	-0,52	0,09	0	-0,04	0,43
10	ING10719	1,13	-0,31	0,08	1	0,06	0,40
11	ING20610	1,00	0,45	0,03	0	-0,03	0,38
12	ING10634	1,28	-0,91	0,11	1	0,02	0,37
13	ING20519	0,99	-0,29	0,09	1	0,09	0,36
14	ING10941	1,03	-0,32	0,13	1	0,14	0,35
15	ING10601	1,06	-0,65	0,09	1	0,18	0,34
16	ING20501	0,95	-0,27	0,11	1	0,23	0,34
17	ING20510	0,96	0,10	0,16	0	0,14	0,32
18	ING10639	1,09	-0,59	0,15	0	0,00	0,31
19	ING10936	1,05	-0,74	0,10	1	0,03	0,30
20	ING20614	1,04	-0,84	0,04	1	0,06	0,30
21	ING10641	0,88	0,07	0,10	1	0,11	0,29
22	ING10636	0,85	-0,06	0,09	0	0,04	0,28
23	ING10816	0,90	-0,57	0,07	1	0,07	0,28
24	ING10930	0,88	0,26	0,11	0	0,02	0,27
25	ING20632	0,90	-0,70	0,04	1	0,05	0,27

Tabela 5.3: Teste adaptativo simulado para $\theta = 1{,}50$ para o método KL.

Ordem	**Item**	**Par.a**	**Par.b**	**Par.c**	**Resposta**	$\hat{\theta}$	**Erro-padrão**
1	ING10836	0,80	-2,51	0,10	1	1,42	2,81
2	ING20618	1,46	0,17	0,04	1	2,95	1,88
3	ING20736	1,08	2,97	0,05	1	4,20	1,26
4	ING10835	0,76	3,23	0,03	0	3,08	1,30
5	ING20532	1,03	2,57	0,04	0	2,02	1,19
6	ING20725	1,15	1,31	0,05	1	2,45	0,86
7	ING20642	0,94	2,00	0,03	0	1,90	0,80
8	ING10527	1,15	1,21	0,12	0	1,24	0,81
9	ING20813	1,35	0,69	0,20	1	1,44	0,66
10	ING10631	0,95	0,79	0,07	1	1,59	0,58
11	ING20512	0,96	1,54	0,17	0	1,36	0,54
12	ING20610	1,00	0,45	0,03	1	1,44	0,50
13	ING20511	0,83	1,11	0,08	0	1,25	0,47
14	ING10625	1,37	0,19	0,10	1	1,30	0,45
15	ING10611	0,87	0,77	0,10	1	1,38	0,44
16	ING20544	0,86	1,08	0,13	1	1,46	0,43
17	ING10640	0,83	0,74	0,05	0	1,29	0,41
18	ING10940	0,76	0,85	0,08	0	1,16	0,39
19	ING20635	0,81	0,44	0,04	1	1,22	0,38
20	ING10635	0,76	0,58	0,08	1	1,27	0,38
21	ING10506	0,82	1,54	0,15	1	1,34	0,38
22	ING10806	0,76	0,88	0,13	1	1,39	0,37
23	ING20804	0,79	0,55	0,11	1	1,43	0,37
24	ING20831	0,85	2,09	0,12	1	1,52	0,37
25	ING10545	0,82	0,67	0,21	1	1,55	0,36

Tabela 5.4: Teste adaptativo simulado para $\theta = -1{,}50$ para o método MIE.

Ordem	**Item**	**Par.** a	**Par.** b	**Par.** c	**Resposta**	$\hat{\theta}$	**Erro-padrão**
1	ING10836	0,80	-2,51	0,10	1	1,42	2,81
2	ING20725	1,15	1,31	0,05	0	-0,99	1,80
3	ING20627	1,68	-1,03	0,03	0	-2,47	1,33
4	ING10834	0,97	-2,40	0,11	1	-1,81	0,98
5	ING20626	1,05	-1,93	0,04	1	-1,38	0,71
6	ING20625	1,20	-1,17	0,04	0	-1,67	0,64
7	ING10614	1,12	-1,58	0,09	0	-1,94	0,63
8	ING10622	0,93	-1,69	0,04	1	-1,69	0,51
9	ING20611	0,96	-1,58	0,04	0	-1,86	0,50
10	ING10624	1,00	-1,50	0,10	1	-1,69	0,46
11	ING10547	0,89	-1,61	0,05	1	-1,56	0,42
12	ING20705	1,15	-1,38	0,17	1	-1,45	0,40
13	ING10613	1,17	-1,14	0,08	0	-1,55	0,38
14	ING20807	1,08	-1,24	0,11	0	-1,64	0,38
15	ING10831	0,91	-1,73	0,10	1	-1,56	0,36
16	ING20605	0,87	-1,42	0,04	0	-1,64	0,35
17	ING10839	0,90	-1,68	0,11	0	-1,73	0,35
18	ING10724	0,91	-1,51	0,11	1	-1,65	0,34
19	ING10701	0,92	-2,32	0,12	1	-1,61	0,32
20	ING10628	0,91	-1,08	0,03	0	-1,66	0,32
21	ING10937	0,97	-1,31	0,12	0	-1,72	0,32
22	ING10715	0,85	-1,59	0,11	1	-1,66	0,31
23	ING10629	0,90	-1,10	0,04	1	-1,58	0,29
24	ING20622	1,29	-0,91	0,04	0	-1,61	0,29
25	ING20701	1,07	-1,12	0,11	1	-1,54	0,28

Tabela 5.5: Teste adaptativo simulado para $\theta = 0{,}00$ para o método MIE.

Ordem	**Item**	**Par.** a	**Par.** b	**Par.** c	**Resposta**	$\hat{\theta}$	**Erro-padrão**
1	ING10836	0,80	-2,51	0,10	1	1,42	2,81
2	ING20725	1,15	1,31	0,05	1	3,37	1,78
3	ING20736	1,08	2,97	0,05	0	1,82	1,53
4	ING10527	1,15	1,21	0,12	0	0,37	1,63
5	ING20618	1,46	0,17	0,04	0	-1,16	1,60
6	ING20627	1,68	-1,03	0,03	1	-0,10	0,87
7	ING10632	1,33	-0,19	0,10	1	0,27	0,69
8	ING10625	1,37	0,19	0,10	1	0,55	0,59
9	ING20610	1,00	0,45	0,03	1	0,75	0,53
10	ING20813	1,35	0,69	0,20	1	0,91	0,51
11	ING10631	0,95	0,79	0,07	0	0,72	0,47
12	ING10611	0,87	0,77	0,10	1	0,84	0,45
13	ING10640	0,83	0,74	0,05	0	0,70	0,43
14	ING20510	0,96	0,10	0,16	1	0,77	0,42
15	ING20635	0,81	0,44	0,04	0	0,63	0,40
16	ING10930	0,88	0,26	0,11	0	0,49	0,39
17	ING10719	1,13	-0,31	0,08	0	0,27	0,40
18	ING10532	1,29	-0,40	0,12	1	0,33	0,37
19	ING10633	1,28	-0,52	0,09	0	0,09	0,37
20	ING10941	1,03	-0,32	0,13	1	0,15	0,35
21	ING20519	0,99	-0,29	0,09	0	0,02	0,35
22	ING10639	1,09	-0,59	0,15	1	0,07	0,34
23	ING10601	1,06	-0,65	0,09	1	0,11	0,33
24	ING20501	0,95	-0,27	0,11	0	0,00	0,32
25	ING20622	1,29	-0,91	0,04	1	0,03	0,31

Tabela 5.6: Teste adaptativo simulado para $\theta = 1{,}50$ para o método MIE.

Ordem	**Item**	**Par.** a	**Par.** b	**Par.** c	**Resposta**	$\hat{\theta}$	**Erro-padrão**
1	ING10836	0,80	-2,51	0,10	1	1,42	2,81
2	ING20725	1,15	1,31	0,05	1	3,37	1,78
3	ING20736	1,08	2,97	0,05	0	1,82	1,53
4	ING10527	1,15	1,21	0,12	1	2,36	1,08
5	ING20532	1,03	2,57	0,04	0	1,86	0,97
6	ING20642	0,94	2,00	0,03	1	2,28	0,69
7	ING20831	0,85	2,09	0,12	0	1,97	0,66
8	ING20512	0,96	1,54	0,17	1	2,12	0,59
9	ING10506	0,82	1,54	0,15	0	1,84	0,58
10	ING20544	0,86	1,08	0,13	0	1,53	0,66
11	ING20813	1,35	0,69	0,20	1	1,63	0,53
12	ING10631	0,95	0,79	0,07	0	1,32	0,61
13	ING20610	1,00	0,45	0,03	1	1,43	0,47
14	ING20511	0,83	1,11	0,08	1	1,53	0,44
15	ING10611	0,87	0,77	0,10	1	1,59	0,42
16	ING10640	0,83	0,74	0,05	1	1,64	0,41
17	ING10940	0,76	0,85	0,08	1	1,69	0,40
18	ING10806	0,76	0,88	0,13	0	1,55	0,39
19	ING20635	0,81	0,44	0,04	1	1,59	0,38
20	ING10635	0,76	0,58	0,08	1	1,62	0,38
21	ING20525	0,69	0,82	0,08	0	1,51	0,37
22	ING10545	0,82	0,67	0,21	1	1,54	0,36
23	ING20804	0,79	0,55	0,11	1	1,58	0,36
24	ING20812	0,70	0,83	0,10	0	1,47	0,35
25	ING20817	0,68	0,95	0,11	1	1,51	0,34

Referências Bibliográficas

Andrade, D. F. (2001) *Comparando desempenhos de grupos de alunos por intermédio da Teoria de Resposta ao Item. Estudos em Avaliação Educacional*, **23**.

Andrade, D. F., Tavares, H. R. e Valle, R. C. (2000) *Teoria de Resposta ao Item: conceitos e aplicações.* Minas Gerais: SINAPE.

Azevedo, C. L. N. (2003) *Métodos de estimação na Teoria de Resposta ao Item.* Mestrado em estatística, Universidade de São Paulo, São Paulo.

Baker, F. B. (2001) *The basics of Item Response Theory.* EUA: ERIC Clearinghouse on Assessment and Evaluation, 2a edn.

Baker, F. B. e Kim, S. H. (2004) *Item Response Theory: Parameter Estimation Techniques.* New York: Marcel Dekker, 2a edn.

Birnbaum, A. (1968) *Some latent trait models and their use in inferring an examinee's ability.* Em *Statistical theories of mental test scores* (eds. F. M. Lord e M. R. Novick). Reading, MA: Addison-Wesley.

Bock, R. D. e Aitkin, D. (1981) *Marginal Maximum Likelihood estimation of Item parameters: An application of* a EM *algorithm. Psychometrika*, **46**, 433–459.

Bock, R. D. e Lieberman, M. (1970) *Fitting a response model for n dichotomously scored items. Psychometrika*, **35**, 179–197.

Bock, R. D. e Mislevy, R. J. (1999) *Adaptive* EAP *Estimation of Ability in a Microcomputer Environment. Applied Psychological Measurement*, **6**, 431–444.

Bock, R. D. e Zimowski, M. F. (1997) *Multiple Group Item Response Theory*. Em *Handbook of Modern Item Response Theory* (eds. W. J. van der Linden e R. K. Hambleton). New York: Spring-Verlag.

Chang, H. H., Qian, J. e Ying, Z. (2001) *A-Stratified Multistage Computerized Adaptive Testing with b-Blocking*. *Applied Psychological Measurement*, **25**, 333–341.

Chang, H. H. e Ying, Z. (1996) *A Global Information Approach to Computerized Adaptive Testing*. *Applied Psychological Measurement*, **20**, 213–229.

— (1999) *a-stratified Multistage Computerized Adaptive Testing*. *Applied Psychological Measurement*, **23**, 211–222.

Dempster, A. P., Laird, N. M. e Rubin, D. B. (1977) *Maximum likelihood from incomplete data via the* EM *algorithm (with discussion)*. *Journal of the Royal Statistical Society*, **39**, 1–38.

Flaugher, R. (2000) *Item Pool*. Em *Computerized Adaptive Testing: A Primer* (ed. H. Wainer). New Jersey: awrence Erlbaum Associates.

Georgiadou, E., Triantafillou, E. e Economides, A. (2007) *A review of item exposure control strategies for Computerized Adaptive Testing developed from 1983 to 2005*. *Journal of Technology, Learning, and Assessment*, **5**.

Gonçalves, F. B. (2006) *Análise bayesiana da Teoria de Resposta ao Item: uma abordagem generalizada*. Mestrado em estatística, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Gonçalves, J. P. (2004) *A integração de testes adaptativos informatizados e ambientes computacionais de tarefas para o aprendizado do Inglês Instrumental*. Mestrado em ciências de computação e matemática computacional, Universidade de São Paulo, São Carlos.

Hambleton, R. K., Swaminathan, H. e Rogers, H. J. (1991) *Fundamental of item response theory*. Califórnia: SAGE Publications.

Hetter, R. D. e Sympson, B. (1997) *Item exposure control in* CAT-ASBAV. Em *Computerized Adaptive Testing: from inquiry to operation* (eds. W. A. Sands, B. K. Waters e J. R. McBride). Washington: American Psychological Association.

Hildebrand, F. B. (1956) *Introduction to Numerical Analysis*. New York: McGraw-Hill.

Kingsbury, G. G. e Zara, A. R. (1989) *Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education*, **4**, 359–375.

Kullback, S. (1959) *Information theory and statistics*. New York: Wiley.

van der Linden, W. J. (1998) *Bayesian Item Selection Criteria for Adaptive Testing. Psychometrika*, **63**, 201–216.

van der Linden, W. J. e Glas, C. A. W. (2003) *Computerized Adaptive Testing: Theory and Practice*. Netherlands: Kluwer Academic.

van der Linden, W. J. e Pashley, P. J. (2003) *Item selection and ability estimation in Adaptive Testing*. Em *Computerized Adaptive Testing: Theory and Practice* (eds. W. J. van der Linden e C. A. W. Glas). Netherlands: Kluwer Academic.

Lindley, D. V. e Smith, A. F. M. (1972) *Bayesian estimates for the linear model. Journal of the Royal Statistical Society*, **34**, 1–41.

Lord, M. F. (1952) *A Theory of Test Scores. Psychometric Monograph*, **7**.

— (1971) *Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement*, **31**, 3–31.

— (1980) *Applications of item response theory to practical testing problems*. Hillsdale,NJ: Erlbaum.

Migon, H. S. e Gamerman, D. (1999) *Statistical Inference: an Integrated Approach*. London: Arnold.

Mislevy, R. J. (1986) *Bayes modal estimation in Item Response models. Psychometrika*, **51**, 177–195.

Neyman, J. e Pearson, E. S. (1936) *Contributions to the theory of testing statistical hypotheses. Statistical Research Memorandum*, **1**, 1–37.

Nojosa, R. T. (2001) *Modelos multidimensionais para a Teoria de Resposta ao Item*. Mestrado em estatística, Universidade Federal de Pernambuco, Pernambuco.

Oliveira, L. H. M. (2002) *Testes adaptativos sensíveis ao conteúdo do banco de itens: uma aplicação em exames de proficiência em Inglês para programas de Pós-graduação*. Mestrado em ciências de computação e matemática computacional, Universidade de São Paulo, São Carlos.

Owen, R. J. (1975) *A Bayesian sequential procedure for quantal response in the context of adaptive mental testing*. *Journal of the American Statistical Association*, **70**, 351–356.

Pasquali, L. (2003) *Psicometria: teoria dos testes na psicologia e na educação*. Petrópolis: Editora Vozes.

Revuelta, J. e Ponsoda, V. (1998) *A comparison of item exposure control methods in Computerized Adaptive Testing*. *Journal of Educational Measurement*, **35**, 311–327.

Samejima, F. (1969) *Estimation of latent ability using a pattern of graded scores*. *Psychometrika Monograph Supplement*.

Sands, W. A. e Waters, B. K. (1997) *Introduction to* ASVAB *and* CAT. Em *Computerized Adaptive Testing: from inquiry to operation* (eds. W. A. Sands, B. K. Waters e J. R. McBride). Washington: American Psychological Association.

Segall, D. O., Moreno, K. E. e Hetter, R. D. (1997) *Item Pool Development and Evaluation*. Em *Computerized Adaptive Testing: from inquiry to operation* (eds. W. A. Sands, B. K. Waters e J. R. McBride). Washington: American Psychological Association.

Swaminathan, H. e Gifford, J. A. (1986) *Bayesian Estimation in the three-parameter logistic model*. *Psychometrika*, **51**, 589–601.

Team, R. D. C. (2008) *R: A language and environment for statistical computing*. Austria. URL`http://www.R-project.org`.

Valle, R. C. (2001) *A construção e a interpretação das escalas de conhecimento - considerações gerais e uma visão do que vem sendo feito no* SARESP. *Estudos em Avaliação Educacional*, **23**.

Wainer, H. (2000) *Computerized Adaptive Testing: A Primer*. New Jersey: Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E. T. e Du, Z. (2003) *Testlet Response Theory: An Analog for the 3PL Model Useful in Testlet-Based Adaptive Testing*. Em *Computerized Adaptive*

Testing: Theory and Practice (eds. W. J. van der Linden e C. A. W. Glas). Netherlands: Kluwer Academic.

Weiss, D. J. e Kingsbury, G. G. (1984) *Application of Computerized Adaptive Testing to Educational Problems. Journal of Educational Measurement*, **21**, 361–375.

Zimowski, M. F., Muraki, E., Mislevy, R. J. e Bock, R. D. (1996) *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items.* Chicago: Scientif Software, Inc.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo