Universidade de São Paulo – USP

Formalização dos Testes Adaptativos Informatizados: A Teoria de Resposta de Itens (TRI)

Leandro Henrique Mendonça de Oliveira
leandroh@icmc.sc.usp.br
Sandra Maria Aluísio
sandra@icmc.sc.usp.br

NILC TR–02-03

Série de Relatórios Técnicos do Núcleo Interistitucional de Lingüística Computacional – NILC
Caixa Postal 668 – CEP: 13560-970 – São Carlos - SP - Brasil

Sumário

Lista de Figuras **5**

Lista de Tabelas **7**

1 Introdução **9**

2 A Teoria de Resposta de Itens (TRI) **11**
2.1 Conceito e Características 11
2.2 Modelos de Resposta de Itens 13
2.2.1 Modelo Normal de Dois Parâmetros 13
2.2.2 Modelo Logístico de Dois Parâmetros 14
2.2.3 Modelo Logístico de Três Parâmetros 16
2.2.4 Modelo Logístico de Um Parâmetro (*Rasch Model*) 16
2.3 Função de Informação de um Item 17
2.4 Abordagens de Seleção de Itens Baseados na TRI 20
2.4.1 Método da Máxima Informação 20
2.4.2 Método Bayesiano 22
2.4.3 Técnicas de Busca de Itens 24
2.5 Estimação de Parâmetros 24
2.5.1 Método da Estimativa de Máxima Verossimilhança (EMV) 25
2.6 Procedimentos de Pontuação 30

3 Conclusão **33**

Bibliografia e Referências **35**

Lista de Figuras

2.1 Passos para a estimação dos parâmetros dos itens (Hambleton and Swaminathan, 1985) 12
2.2 Curva Característica de um Item, com valores de a positivo e negativo (Baker, 1992) 14
2.3 Curvas Características de um Item baseado nos modelos Normal e Logístico com b = 0.6 e a = 1.2 (Baker, 1992) 15
2.4 Curva Característica de um Item baseado no modelo logístico de três parâmetros, com b=0.6, a=1.2 e c=0.15 (Baker, 1992) 17
2.5 Curva Característica de um Item baseado no Rasch Model, com b = 1.822 (Baker, 1992) 18
2.6 Curvas de informação de 10 itens hipotéticos (Weiss, 1985) 19
2.7 Representação das curvas de informação de 10 itens e o nível de habilidade com valor 0 (Weiss, 1985) 21
2.8 Representação das curvas de informação de 9 itens e o nível de habilidade com valor -1 (Weiss, 1985) 22
2.9 Representação das curvas de informação de 8 itens e o nível de habilidade com valor -0,5 (Weiss, 1985) 23

Lista de Tabelas

2.1 Principais considerações sobre modelos de resposta de itens (Hambleton and Swaminathan, 1985) 11
2.2 Modelos de resposta de um item, para respostas objetivas e subjetivas (Hambleton and Swaminathan, 1985) 13

Capítulo 1

Introdução

De maneira geral uma avaliação informatizada é aquela que utiliza o computador como meio principal de apresentação das questões de um teste, bem como a coleta e o armazenamento de resposta, o cálculo e a interpretação dos resultados. Nesse contexto, existem diferentes graus de informatização de uma avaliação, podendo variar desde um nível baixo, em que a apresentação das questões se dá de forma tradicional (em folhas de papel) e são corrigidas por meio de leitura óptica (cartão resposta), até um nível mais alto, em que as questões podem assumir novos formatos de apresentação, além do processamento das respostas e cálculos dos resultados serem realizados eletronicamente.

As avaliações informatizadas podem ser implementadas por meio dos *métodos convencionais* ou *métodos alternativos*. Os métodos convencionais correspondem àqueles que aplicam os testes tradicionais de lápis e papel em um sistema computacional, apresentando-os de forma eletrônica e obtendo correção automática das respostas e resultados. Os métodos alternativos consistem na implementação de técnicas que buscam avaliar de maneira confiável o "real" conhecimento de um aluno, implementando técnicas que tentam quantificar a habilidade individual em uma avaliação em um determinado domínio do conhecimento.

Dentre os métodos alternativos destacam-se os **Testes Adaptativos Informatizados (TAI's)**[1], que empregam características específicas, como, por exemplo, o nível de dificuldade e o fator de discriminação de cada questão do teste, além da habilidade de cada aluno para identificar o nível de capacidade individual. Assim, a idéia fundamental dos testes adaptativos é ajustar as questões de um teste ao nível de habilidade de cada estudante que participa de uma avaliação, de maneira que é apresentada uma seqüência de questões que mais se adapta ao conhecimento de cada aluno. Portanto, a principal característica que o define é que um teste pode não ser o mesmo para todos os participantes da avaliação, possibilitando uma medida mais fiel a respeito da competência de cada um (Olea et al. 1999).

Os primeiros testes adaptativos foram desenvolvidos e administrados via lápis e papel. Dentre estes se destacam principalmente o *Teste de Binet* (primeiro teste adaptativo) e os *Testes Adaptativos de Dois Estágios*[2], *Piramidais*[3] *e Estratificados*[4] (Weiss, 1985; Oliveira, 2002). A principal diferença entres os testes adaptativos de lápis e papel e os informatizados está na capacidade de

[1]Computer Adaptive Testing - em inglês.
[2]Two-Stage Adaptive Testing - em ingles.
[3]Pyramidal Adaptive Testing - em inglês.
[4]Stratified Adaptive Testing - em inglês.

"adaptação" que cada um pode desempenhar, visto que neste último a eficiência do teste pode ser melhorada com o uso da **Teoria de Resposta de Itens (TRI[5])**.

A Teoria de Resposta de Itens é um conjunto de modelos estatísticos que descrevem tanto os itens de um teste, quanto as suas características (índices de dificuldade e/ou discriminação e/ou adivinhação), como os estudantes, quanto às habilidades ou competências (Baker, 1992; Hambleton and Swaminathan, 1985; Rudner, 1998). Tal teoria reúne os modelos de resposta de itens (questões) que, por sua vez, especificam a relação entre a pontuação observada e não observada de um estudante durante um teste. Dessa forma, pode-se dizer que o desenvolvimento da **TRI** foi um marco importante na aplicação dos testes adaptativos, já que tornou possível a administração de diferentes itens com diversas características a distintos indivíduos, respeitando e considerando os traços intelectuais de cada um.

Para otimizar e melhorar sua execução, os TAI's mais recentes utilizam um banco de itens que facilita a busca e recuperação dos mesmos durante uma avaliação. Assim, é necessário realizar uma calibração do banco de itens, ajustando-o a um modelo de resposta (pertencente à TRI), ou seja, é preciso escolher e avaliar quais os modelos de respostas mais indicados para que os objetivos de uma avaliação sejam atingidos. Para um estudo mais detalhado da calibração de um banco de itens veja Oliveira (2002).

A seleção de itens, a calibração do banco, a estimação da habilidade e os procedimentos de pontuação são atividades inerentes aos testes adaptativos, pois estes regem a sua execução e aplicação. A seleção de itens corresponde à tarefa de selecionar um determinado item a ser fornecido a um estudante (dado seu nível de habilidade corrente) durante um teste. A calibração dos itens consiste na tarefa de estimar os parâmetros dos itens (**a** e/ou **b** e/ou **c**) identificando suas características e possibilitando seu ajuste a um modelo de resposta da TRI. Os métodos de estimação de habilidade têm a função de estimar a habilidade de um estudante durante um teste após o mesmo ter respondido uma determinada questão, de forma que o próximo item seja selecionado e administrado. Finalmente, os procedimentos de pontuação têm a responsabilidade de converter a habilidade estimada durante todo o teste em pontos convencionais (usados nas avaliações tradicionais) de maneira a classificar o estudante após a avaliação.

Tanto os modelos de resposta da TRI, como as atividades de aplicação dos TAI's acima descritas, capacitam a implementação da uma avaliação adaptativa a qualquer domínio de conhecimento, desde que para isto, seja realizado um estudo cuidadoso sobre as questões dos testes, de maneira a selecionar os modelos e métodos mais adequados aos objetivos que se deseja alcançar com a avaliação, tendo como resultado um cenário ideal para aplicação dos testes adaptativos.

O propósito deste relatório técnico é apresentar e descrever os formalismos dos testes adaptativos, ressaltando principalmente os modelos de resposta de itens (da TRI), bem como os métodos de seleção de itens, estimação de parâmetros e os procedimentos de pontuação que permitem a implantação da avaliação adaptativa.

[5]Item Response Theory - ITR - em inglês

Capítulo 2

A Teoria de Resposta de Itens (TRI)

2.1 Conceito e Características

A Teoria de Resposta de Itens (TRI) é uma reunião de modelos estatísticos usados para fazer predições, estimativas ou inferências sobre as habilidades (ou competências) medidas em um teste. Através dos modelos estatísticos inerentes ao seu contexto, é possível predizer tais habilidades por meio da correspondência entre a pontuação obtida por um estudante em uma situação de teste e os itens a ele fornecido (Hambleton and Swaminathan, 1985; Rudner, 1998).

Um modelo de resposta de itens especifica a relação entre a pontuação (desempenho) *observada* e *não observada* de um estudante em um teste. Essa relação é descrita por uma função matemática baseada nos parâmetros dos itens do teste e nas respostas dos indivíduos envolvidos na avaliação. Assim, diferentes modelos estatísticos são formados pela especificação dos parâmetros dos itens e das habilidades dos estudantes (o que se deseja avaliar) que influenciam seus desempenhos.

A Tabela 2.1 especifica as principais considerações sobre os Modelos de Reposta de Itens, quando ocorre o adequado ajuste entre um modelo de reposta escolhido e os parâmetros dos itens.

Tabela 2.1: Principais considerações sobre modelos de resposta de itens (Hambleton and Swaminathan, 1985)

Considerações sobre os modelos de resposta
São modelos que supõem que o desempenho de um estudante em um teste pode ser predito em termos de uma ou mais habilidades pessoais;
O modelo de resposta de um item especifica a relação entre a pontuação observada de um estudante e as habilidades esperadas ou assumidas que fundamentam os testes;
O desempenho de um estudante em um teste deve ser estimado a partir da pontuação observada em um conjunto de itens de teste.

A Figura 2.1 mostra uma seqüência de passos para estimar os parâmetros dos itens de um banco e obter a escala de pontuação da habilidade.

Primeiro, para estimar os parâmetros dos itens o projetista do teste analisa um conjunto de respostas de itens já administrados a um grupo de estudantes, que o conduz a levantar várias características preliminares sobre seu comportamento, por exemplo, como quais itens foram respondidos freqüentemente de forma correta ou incorreta. Depois da análise desse conjunto de respostas,

é necessário selecionar o modelo de resposta que se ajusta da melhor forma possível aos itens observados. O próximo passo é estimar os parâmetros dos itens, combinando com o melhor modelo de resposta escolhido sob os dados analisados. Uma vez realizado a estimação dos parâmetros é definido um procedimento que efetua a conversão da escala de habilidade observada para uma escala conveniente de pontuação real (utilizada nos testes tradicionais de lápis e papel).

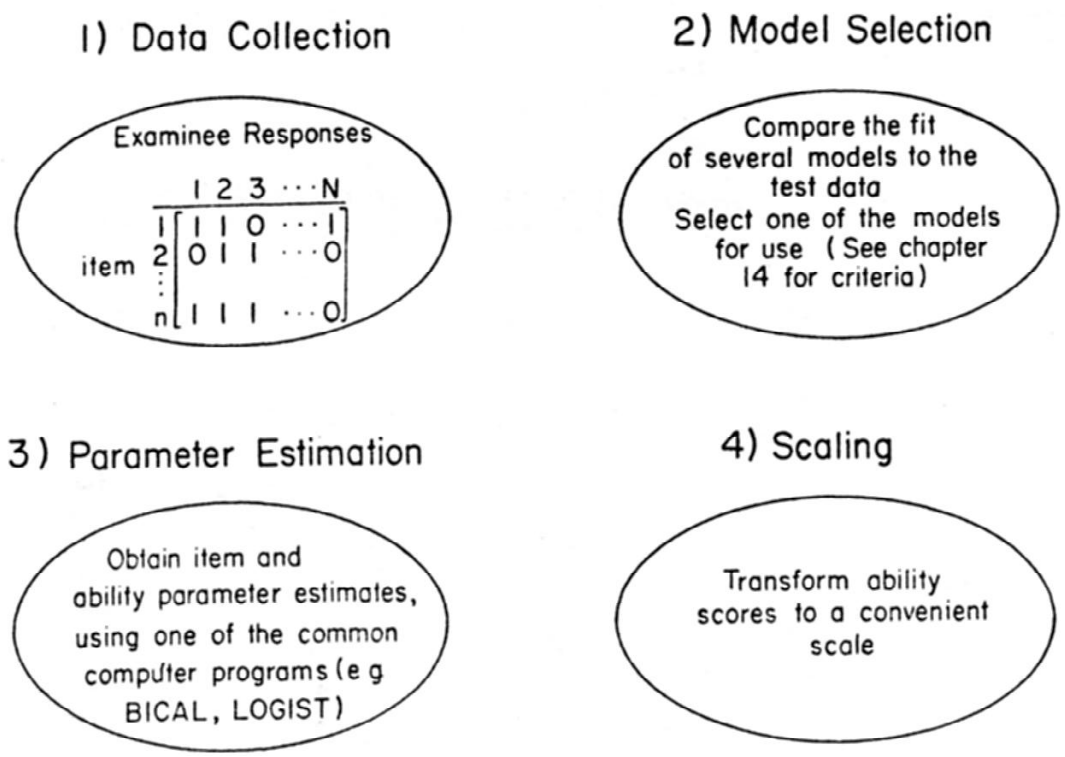

Figura 2.1: Passos para a estimação dos parâmetros dos itens (Hambleton and Swaminathan, 1985)

Talvez a vantagem mais importante dos modelos de resposta de itens é que, dado um conjunto de itens de teste que são adequadamente ajustados a um determinado modelo isto é, os parâmetros dos itens são conhecidos e bem especificados, é possível estimar a habilidade de um indivíduo na mesma escala na qual os dados foram ajustados. Dessa forma, sem levar em consideração o número de itens administrados, a estimação de habilidade para cada estudante será teoricamente verdadeira e mais justa. Portanto, os modelos de respostas fornecem uma maneira de comparar o desempenho de estudantes, mesmo que eles tenham recebido diferentes conjuntos de itens de teste em uma avaliação.

Quando um modelo de reposta é devidamente escolhido para o ajuste dos itens de teste de um determinado domínio de conhecimento, é obtida uma distribuição de habilidade de acordo com os parâmetros de cada item, fornecendo assim informações sobre o comportamento do mesmo dentro do processo de teste. Tal distribuição é chamada de Curva Característica de um Item[1] (CCI) e está presente em todos os modelos de resposta, tendo suas particularidades em cada um deles. Estes modelos serão estudados na próxima subseção.

[1] *Item Curve Caracteristic* - ICC em inglês

2.2 Modelos de Resposta de Itens

Os modelos de resposta de itens são utilizados para uma análise detalhada dos itens de um teste observando principalmente seu comportamento em uma avaliação. Geralmente, um modelo de resposta consiste na aplicação e implementação de várias equações matemáticas (estas com objetivos variados), por exemplo: nível de informação de um item, estimação de parâmetros e habilidade, e medida de erros.

Uma das maneiras de classificar os modelos de resposta de itens, baseia-se nos tipos de resposta que podem ser empregadas. Considerando a forma como o aluno responde os testes, objetiva (usando testes objetivos) ou subjetiva (respostas nominais e dissertativas) é possível aplicar diferentes modelos de resposta. Na Tabela 2.2 são mostrados os principais modelos de respostas.

Tabela 2.2: Modelos de resposta de um item, para respostas objetivas e subjetivas (Hambleton and Swaminathan, 1985)

Tipos de Respostas	***Modelo***
Respostas Objetivas	Modelo Linear (*Latent Linear*)
	Modelo da Escala Perfeita (*Perfect Scale*)
	Modelo da Distância Latente (*Latent Distance*)
	Modelo Normal de 1,2 e 3 parâmetros (One,Two-,Tree Parameter Normal)
	Modelo Logístico de 1,2 e 3 parâmetros (One,Two-,Tree Parameter Logistic)
	Modelo Logístico de 4 parâmetros(Four-Parameter Logistic)
Respostas Subjetivas	Resposta Nominal (*Nominal Response*)
	Resposta de Classificação (*Graded Response*)
	Modelo de Crédito Parcial (*Partial Credit Model*)

Todos os modelos descritos na Tabela 2.2 são unidimensionais, ou seja, assumem que os itens do teste foram ajustados para medir uma única habilidade. As principais características que diferenciam os modelos de resposta são suas fórmulas matemáticas, conseqüentemente suas curvas características, e a forma de pontuação empregada. Neste trabalho veremos apenas os principais modelos de respostas aplicados aos testes objetivos, especificamente os modelos de um, dois e três parâmetros, pois a aplicação desta teoria neste projeto contempla os testes objetivos para avaliação da proficiência em língua inglesa.

2.2.1 Modelo Normal de Dois Parâmetros

Segundo Hambleton and Swaminathan (1985), Lord (1952, 1953a) propôs o primeiro modelo de resposta de dois parâmetros, no qual a curva característica do item assume sua forma sob a distribuição normal dada por:

$$P_i(\theta) = \int_{-\infty}^{a_i \cdot (\theta - b_i)} \frac{1}{\sqrt{2\pi}} \cdot \exp^{-z^2/2} dz$$

onde $P_i(\theta)$ é a probabilidade que um determinado aluno selecionado aleatoriamente com habilidade θ, responda a um item i corretamente; as variáveis b_i e a_i são parâmetros que caracterizam o item i; e z é o desvio normal da distribuição com média em b_i e desvio padrão $1/a_i$. O

parâmetro b_i é comumente referenciado como índice de *dificuldade* e representa o nível de dificuldade de um determinado item, que inicialmente é atribuído pelo projetista do teste baseado em seu conhecimento sobre o item, podendo ser alterado a medida que um estudante o responde correta ou incorretamente. O parâmetro a_i é chamado de índice de *discriminação* e representa a capacidade que um item possui de discriminar indivíduos que apresentam nível de habilidade muito próximo.

Geralmente, os valores de b_i variam tipicamente, neste modelo, entre -2.0 e +2.0. Valores de b_i próximos ou iguais a -2.0 correspondem a item fáceis, e valores de b_i próximos a +2.0 correspondem a item difíceis. Por sua vez, o item de discriminação a_i, é definido teoricamente no intervalo de $-\infty$ a $+\infty$. Entretanto, itens que possuem o valor de a_i negativos são descartados para os testes de habilidade, pois não possuem boa capacidade de discriminação. Do mesmo modo, valores de a_i muito altos não são usuais, já que resultam em curvas características de item muito acentuadas. Contundo, valores baixos de a_i permitem uma distribuição gradual da curva, em função e ao longo da escala de habilidade do estudante. Assim, valores típicos do índice de discriminalidade situam-se entre 0 e +2.

A Figura 2.2 mostra a curva característica de um item para o modelo normal de dois parâmetros. Note que quando o valor de a é negativo ocorre uma inversão de sentido da curva, indicando que o item possui baixo poder discriminativo. Para os valores de b vemos que seu aumento é gradativo, de acordo como os valores de θ, que significa dizer que quanto mais alto o valor da habilidade maior será a probabilidade de um indivíduo acertar um determinado item.

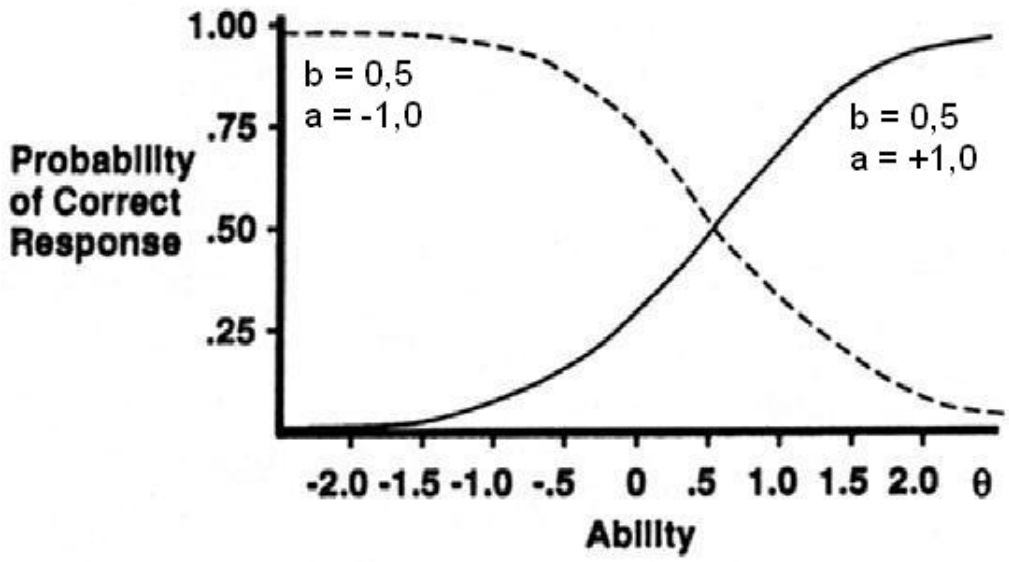

Figura 2.2: Curva Característica de um Item, com valores de a positivo e negativo (Baker, 1992)

2.2.2 Modelo Logístico de Dois Parâmetros

A curva característica de um item nos modelos logísticos de dois parâmetros é dada pela fórmula matemática:

$$P_i(\theta) = \frac{\exp^{D \cdot a_i \cdot (\theta - b_i)}}{1 + \exp^{D \cdot a_i \cdot (\theta - b_i)}}$$

com $i = 1, 2, 3, ..., n$.Existe uma maneira alternativa para escrever a função $P_i(\theta)$ acima. Multiplicando o numerador e o denominador da equação pelo fator: $\exp^{-D \cdot a_i \cdot (\theta - b_i)}$ a fórmula torna-se:

$$P_i(\theta) = (1 + \exp^{-D \cdot a_i \cdot (\theta - b_i)})^{-1}$$

e que pode ser escrita como uma melhor alternativa e de leitura mais fácil da seguinte forma:

$$P_i(\theta) = \frac{1}{1 + \exp^{-D \cdot a_i \cdot (\theta - b_i)}}$$

com $i = 1, 2, 3, ..., n$

Os parâmetros a, b e θ possuem essencialmente a mesma interpretação que no modelo normal, e a constante D é o fator de escala. A maior vantagem do modelo logístico de dois parâmetros em relação ao modelo normal, é a facilidade de lidar com a equação, já que esta é mais tratável matematicamente e não envolve a função de integração presente no modelo normal. Além disso, o modelo logístico é computacionalmente mais simples de implementar (Hambleton and Swaminathan, 1985; Revuelta and Ponsoda, 1997).

A Figura 2.3 mostra a forma geral das curvas características, tanto do modelo normal como logístico. Note que suas aparências são similares, diferenciando apenas no ponto de inclinação, pois compartilham os mesmos parâmetros. Os dois modelos tem o ponto de inflexão situado no parâmetro b, visto que este valor tem o mesmo significado nos dois modelos. Entretanto, os mesmos valores numéricos do parâmetro a produzem diferentes pontos de inclinação.

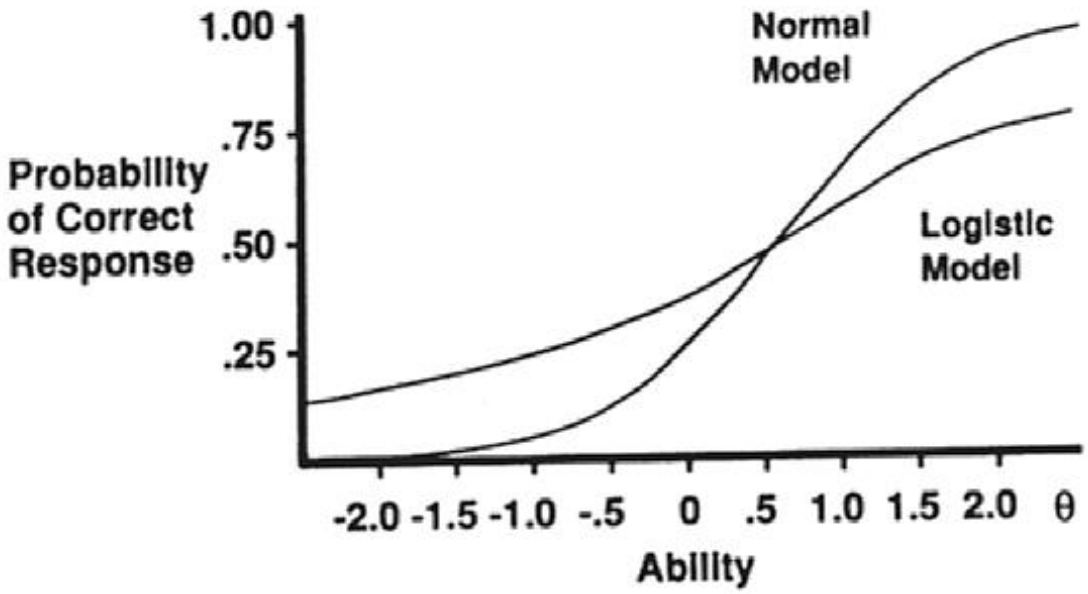

Figura 2.3: Curvas Características de um Item baseado nos modelos Normal e Logístico com b = 0.6 e a = 1.2 (Baker, 1992)

Os modelos de dois parâmetros, tanto o normal como o logístico assumem que os estudantes não podem acertar um determinado item dada a uma baixa habilidade, mas existem possíveis situações em que um indivíduo consegue responder a um item corretamente tendo um baixo nível de habilidade ou competência. O modelo logístico de três parâmetros adiciona em suas equações essa possibilidade, e será estudado na próxima subseção.

2.2.3 Modelo Logístico de Três Parâmetros

O modelo logístico de três parâmetros pode ser obtido a partir do modelo logístico de dois parâmetros pela adição de um terceiro parâmetro, chamado c ou índice de *adivinhação*. Tal índice representa a possibilidade de se ter uma resposta correta dado um baixo nível de habilidade, ou seja, é a mínima probabilidade de acerto.

A fórmula matemática do modelo logístico de três parâmetros é a seguinte:

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{[1 + \exp^{-D \cdot a_i \cdot (\theta - b_i)}]}$$

na qual $P_i(\theta)$ é a probabilidade que um indivíduo com nível de habilidade θ responda um item i corretamente, b_i é o índice de dificuldade, a_i é o índice de discriminação, c_i é o de adivinhação e D é o fator de escala.

O valor do parâmetro c pode ser atribuído a um item por meio do desenvolvimento das opções distratoras pertencentes a um teste objetivo. Os distratores são as opções atrativas, mas incorretas, de uma determinada questão. Os estudantes que possuem uma baixa habilidade são atraídos por essas "opções incorretas" (Hambleton and Swaminathan, 1985). Os valores do índice de adivinhação devem ser baixos, já que um valor alto indica uma grande possibilidade de acerto de uma questão dada uma baixa habilidade.

A Figura 2.4 fornece um exemplo de uma curva característica de um item baseado no modelo logístico de três parâmetros. Primeiramente, observamos na parte inferior do gráfico uma linha assintótica a um valor de $P_i(\theta) = 0.15$, que define o limite inferior da curva. Segundo, o limite superior da curva do gráfico é de 1.0. Finalmente, os valores de a e b definem as características da curva de um item (ao longo da escala de habilidade) pela equação logística de três parâmetros, a qual é compreendida pelo limite superior de 1.0 e o valor de c que define o limite inferior, nesse caso $c = 0.15$. Assim, os valores de $P_i(\theta)$ dependerão diretamente dos valores de a, b, c e θ, sendo que o valor de c será o limite inferior da curva (Baker, 1992).

Para obter o modelo logístico de dois parâmetros a partir do modelo de três parâmetros, basta assumir que o valor do índice de adivinhação c é sempre igual a zero ($c = 0$).

2.2.4 Modelo Logístico de Um Parâmetro (*Rasch Model*)

O modelo logístico de um parâmetro, também chamado de *Rasch Model* em homenagem a seu criador Georg Rasch, citado por Hambleton and Swaminathan (1985), pode ser visto como um modelo de resposta, no qual a curva característica de um item é baseada em uma função de apenas um parâmetro.

O Rasch Model é um caso especial do modelo logístico de três parâmetros de maneira que todos os itens assumem uma capacidade semelhante de discriminação (mesmo valor do parâmetro a) e mínima possibilidade de adivinhação (baixos ou nulos valores do parâmetro c). Dessa forma, o único parâmetro que o modelo utiliza é o índice de dificuldade, representado pelo valor de b. Assim, este modelo pode ser considerado com um caso especial dos modelos de dois e três parâmetros (Hambleton and Swaminathan, 1985; Baker, 1992), pois possui algumas propriedades específicas que o torna atrativo: primeira, o modelo envolve um único parâmetro facilitando o trabalho de administração dos itens; e segundo, agiliza a tarefa de estimação do parâmetro em relação a outros modelos. A equação do Rasch Model é a seguinte:

$$P_i(\theta) = \frac{\exp^{D \cdot a_i \cdot (\theta - b_i)}}{1 + \exp^{D \cdot a_i \cdot (\theta - b_i)}}$$

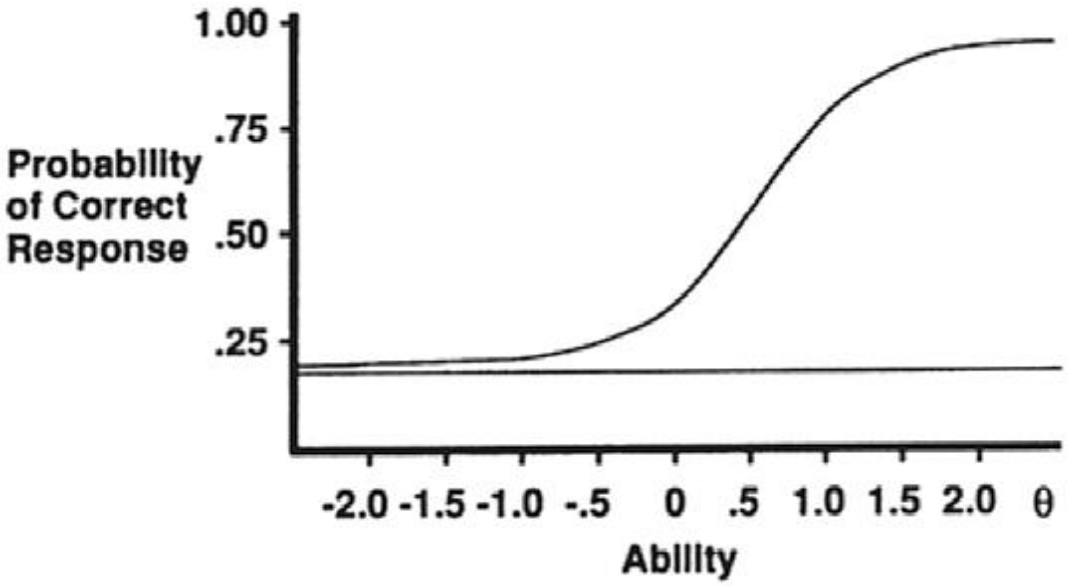

Figura 2.4: Curva Característica de um Item baseado no modelo logístico de três parâmetros, com b=0.6, a=1.2 e c=0.15 (Baker, 1992)

A Figura 2.5 mostra a curva característica de um item regida pela equação logística de um parâmetro. O gráfico da Figura 2.5 relata a probabilidade de uma resposta correta de um item dada uma escala de habilidade.

Conforme pode ser visto, a forma da curva não parece com qualquer uma das curvas vistas anteriormente (modelos de dois e três parâmetros), já que a diferença básica desta curva característica é que tanto a escala de habilidade θ quanto o valor de b (índice de dificuldade) variam no intervalo de 0 ao ∞ (infinito) com uma unidade arbitrária de medida, isso porque neste modelo, as chances de sucesso de um aluno com habilidade θ em um item com nível de dificuldade b é dada pela relação direta entre a habilidade do aluno e dificuldade do item.

Dadaos as principais formas e modelos de repostas dos itens, na próxima subseção veremos como os testes adaptativos podem ser implementados usando os métodos da Teoria de Resposta de Itens e seus diversos modelos de resposta.

2.3 Função de Informação de um Item

Os parâmetros que caracterizam um item, além de ditarem a forma das curvas características dos modelos de resposta, também são combinados de maneira a medir o nível de informação que um determinado item fornece ao longo da distribuição da escala de habilidade. Dessa forma, podem descrever quão precisamente um item ajusta-se a um dado nível de habilidade.

O nível de informação de um item representa a quantidade proporcionada, baseado em seus parâmetros, a um determinado ponto da escala de habilidade. A *curva de informação de um item* tem seu "pico" representado pelo índice de dificuldade (parâmetro b) e a altura desse "pico" representa o grau de precisão do item em um ponto da escala de habilidade, e é regida pelo índice de discriminação (parâmetro a) (Weiss, 1985); em outras palavras, o parâmetro de discriminação a revela a extensão da curva de informação ao longo da escala de habilidade. Assim, um item que

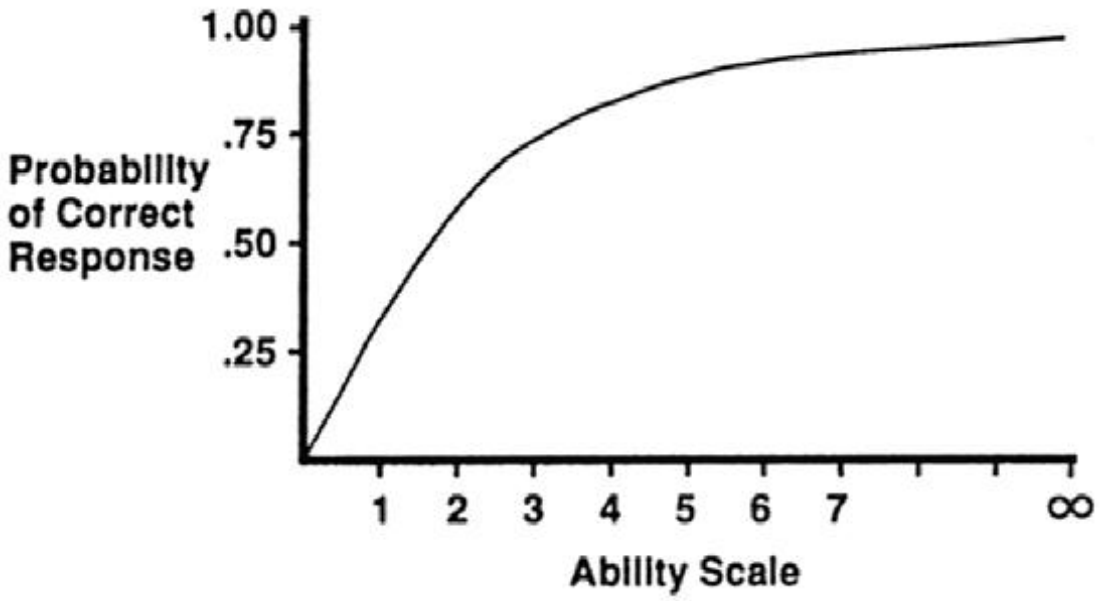

Figura 2.5: Curva Característica de um Item baseado no Rasch Model, com b = 1.822 (Baker, 1992)

possui baixo valor de a terá um baixo pico e sua distribuição será mais achatada, enquanto que um pico alto da curva de informação com achatamento mínimo indica um alto valor do índice de discriminalidade.

Dessa maneira, dado os parâmetros da TRI para um conjunto de itens, o valor do nível de informação pode ser calculado para cada possível valor da habilidade avaliada.

A equação abaixo representa o nível de informação de um item, dado um valor θ da escala de habilidade. A função de informação de um item pode variar de acordo com a quantidade de parâmetros envolvidos no modelo de resposta adotado, e assume a seguinte forma:

$$I_i(\theta) = \frac{(P\prime_i(\theta))^2}{P_i(\theta) \cdot Q_i(\theta)}$$

no qual

$$Q_i(\theta) = 1 - P_i(\theta)$$

e $P\prime_i(\theta)$ é a primeira derivada da função $P_i(\theta)$.

A Figura 2.6 mostra 10 curvas de informações de itens hipotéticos. O eixo horizontal representa a habilidade θ sendo medida e o eixo vertical representa o nível de informação de um item dada uma habilidade qualquer.

Valores altos no nível de informação indicam que um item fornece maior quantidade de informação naquele nível de habilidade. O item 9 por exemplo, tem o maior nível de informação, e é o item mais discriminativo entre os 10, seguidos pelos itens 4 e 1. Os itens menos discriminativos são os itens 3 e 7 que possuem a curvas mais achatadas entre os 10 itens. Como mostrado da Figura 2.6 a curva de pico mais alto (item 9) possui menor distância nas bases. Isso significa que o item tem capacidade de diferenciar níveis de habilidade muito próximos.

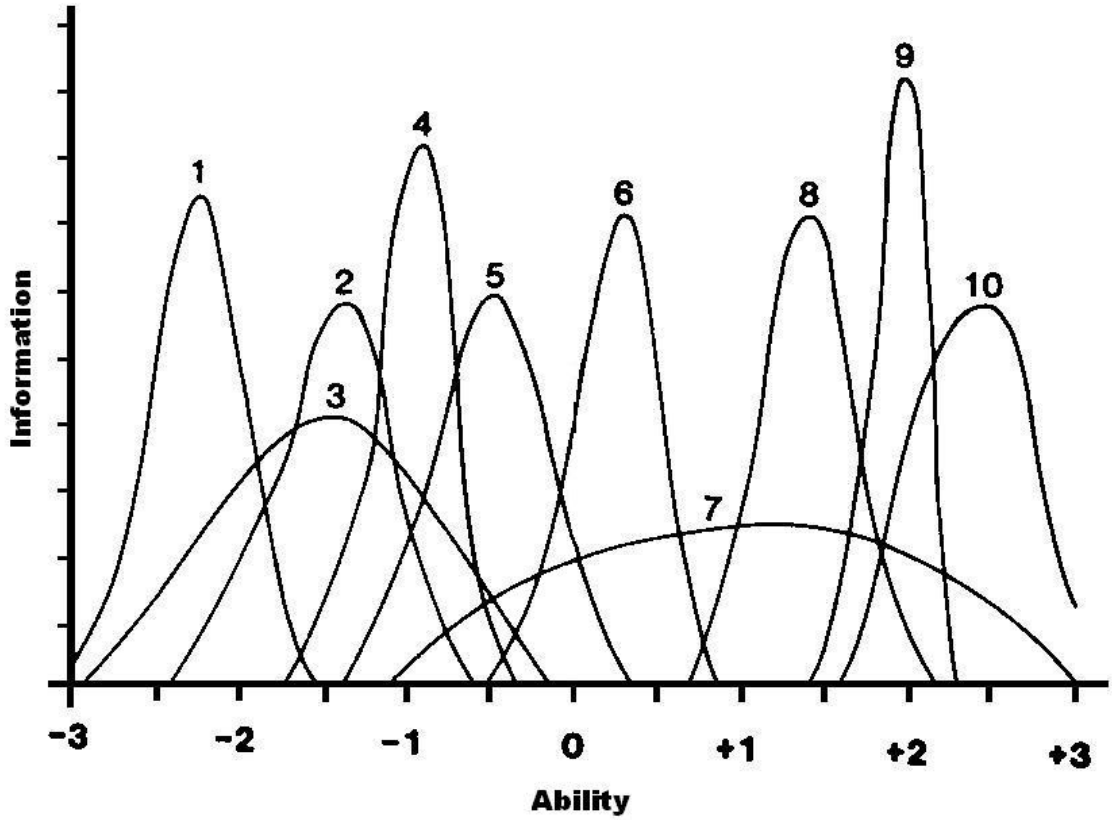

Figura 2.6: Curvas de informação de 10 itens hipotéticos (Weiss, 1985)

Note, ainda na Figura 2.6, que as curvas de informação dos itens diferem sua localização ao longo da escala de habilidade. Por exemplo, os itens 1, 6 e 8 possuem curvas de informação bastante parecidas, tanto na altura do pico quanto no poder de discriminação, mas são localizadas em diferentes regiões. Observe que o pico de informação da curva do item 1 corresponde ao valor de habilidade θ próximo de -2.5, enquanto que o pico da curva do item 6 é próximo a 0.5 e do item 8 a 1.5. Apesar de terem formas parecidas o item 1 diferencia indivíduos com baixa habilidade, e é mais fácil que os outros dois itens (6 e 8), já o item 8 é mais difícil que o item 6, porque em geral o item 8 diferencia indivíduos com maiores níveis de habilidade.

Como os índices de dificuldade e discriminalidade (parâmetros b e a) são independentes em cada item, pode ocorrer que dois ou mais itens tenham o mesmo valor de dificuldade mas discriminalidades distintas. Quando isso acontece, as curvas de informação dos itens terão pico no mesmo ponto da escala de habilidade, mas a altura e o achatamento da curva serão diferenciados. Tal fato pode ser visto entre os itens 2 e 3 na Figura 2.6, pois ambos possuem o mesmo índice de dificuldade -1.5, mas o item 2 é mais discriminativo do que o item 3 pois fornece maior informação.

2.4 Abordagens de Seleção de Itens Baseados na TRI

O advento da Teoria da Resposta de Itens proporcionou o desenvolvimento de várias abordagens de seleção de itens voltadas para os testes adaptativos. Esta subseção faz uma revisão dos procedimentos clássicos e alternativos de seleção de itens destacando suas vantagens e desvantagens.

Os procedimentos de seleção de itens podem ser categorizados em dois grupos: os *pré-estruturados* e os de *ramificação variável*. A maior diferença entre estes dois grupos está na capacidade de adaptação permitida. Os procedimentos pré-estruturados resultam em uma adaptação determinística, com ramificação mais contida durante a execução do teste e os procedimentos com ramificação variável permitem que o processo de adaptação e ajuste dos itens aos alunos seja mais específico e eficiente, quase ilimitado ao logo do teste (Kingsbury and Zara, 1989).

O grupo de procedimento *pré-estruturados* foi criado quando os testes adaptativos começaram a se desenvolver, e os mais conhecidos são: *Teste de Dois Estágios* (Two Stage Test), o *Teste Piramidal* (Pyramidal Test) e os *Testes Adaptativos Estratificados* (Stratified Adaptative Testing). Para um estudo mais preciso e completo desses procedimentos veja (Kingsbury and Zara, 1989) e (Oliveira, 2002).

As técnicas pré-estruturadas de seleção de itens não são muito adequadas para as avaliações adaptativas, pois elas limitam a capacidade de adaptação que pode ser realizada sobre um indivíduo, além de não usarem todas as informações disponíveis dos estudantes para a realização da seleção dos itens (Kingsbury and Zara, 1989; Olea et al., 1999).

O grupo de procedimento de *ramificação variável* é o de desenvolvimentos mais recente na seleção de itens aplicados aos testes adaptativos, destacando entre eles o **Método da Máxima Informação** e o **Método Bayesiano**.

2.4.1 Método da Máxima Informação

O Método da Máxima Informação consiste na escolha de itens que irão maximizar as informações existentes em cada item quando o estudante responde uma questão. Para selecionar o "melhor" item a informação fornecida por cada questão disponibilizada no banco de itens é calculada, dada a habilidade corrente do indivíduo e os parâmetros (a e/ou b e/ou c) dos itens. Aquele que fornecer mais informação em função da habilidade do indivíduo será selecionado e administrado, assim os cálculos para descobrir qual será o melhor item são sempre realizados depois que o estudante responde uma questão e sua nova habilidade é computada.

As vantagens fornecidas pelo Método da Máxima Informação sobre os métodos pré-estruturados são consideráveis. Primeiro, porque sempre o item mais informativo será administrado, e segundo, devido a esse fato, a eficiência do teste é aprimorada. Além disso, este método permite diferentes ramificações ao longo do teste, devido a busca constante do melhor item a cada habilidade estimada do indivíduo, fornecendo uma medida precisa (Kingsbury and Zara, 1989).

As desvantagens do método de seleção de Máxima Informação estão intimamente ligadas aos procedimentos de busca de itens, principalmente quando existe um grande banco de questões, pois a constante e repetida tarefa de seleção pode consumir tempo de administração desses itens, fazendo com que o indivíduo fique esperando, o que pode ser impraticável em algumas situações de teste.

Especificação do Modelo de Máxima Informação e sua aplicação nos TAI's

A maioria dos testes adaptativos informatizados desenvolvidos utiliza as informações de um item para executar a tarefa de adaptação ao nível do estudante. Assim, a estratégia da máxima informação aplicada aos testes adaptativos indica que o item a ser administrado a qualquer momento

do teste é o que fornece maior quantidade de informação, dado o nível corrente de habilidade do aluno (Weiss, 1985).

Para exemplificar esta estratégia de seleção observe as Figuras 2.7 2.8 e 2.9. Como exposto na Figura 2.6 o eixo horizontal representa o nível de habilidade e o eixo vertical representa o nível de informação. Na Figura 2.7, a linha vertical pontilhada ao nível 0 da escala de habilidade representa a estimação inicial da habilidade de um determinado estudante no início do teste. Neste exemplo, com a habilidade no valor 0, três itens fornecem níveis de informação diferentes: os itens 5, 6 e 7. As curvas de informação dos itens 5 e 7 se cruzam no mesmo ponto em intersecção com a linha pontilhada, de maneira que estes itens fornecem a mesma quantidade de informação a este nível. Entretanto o item 6 fornece mais informação em relação aos dois itens anteriores. Desse modo, como nenhum outro item fornece mais informações, o item 6 é administrado ao estudante, que por sua vez o responde e um novo nível de habilidade é calculado.

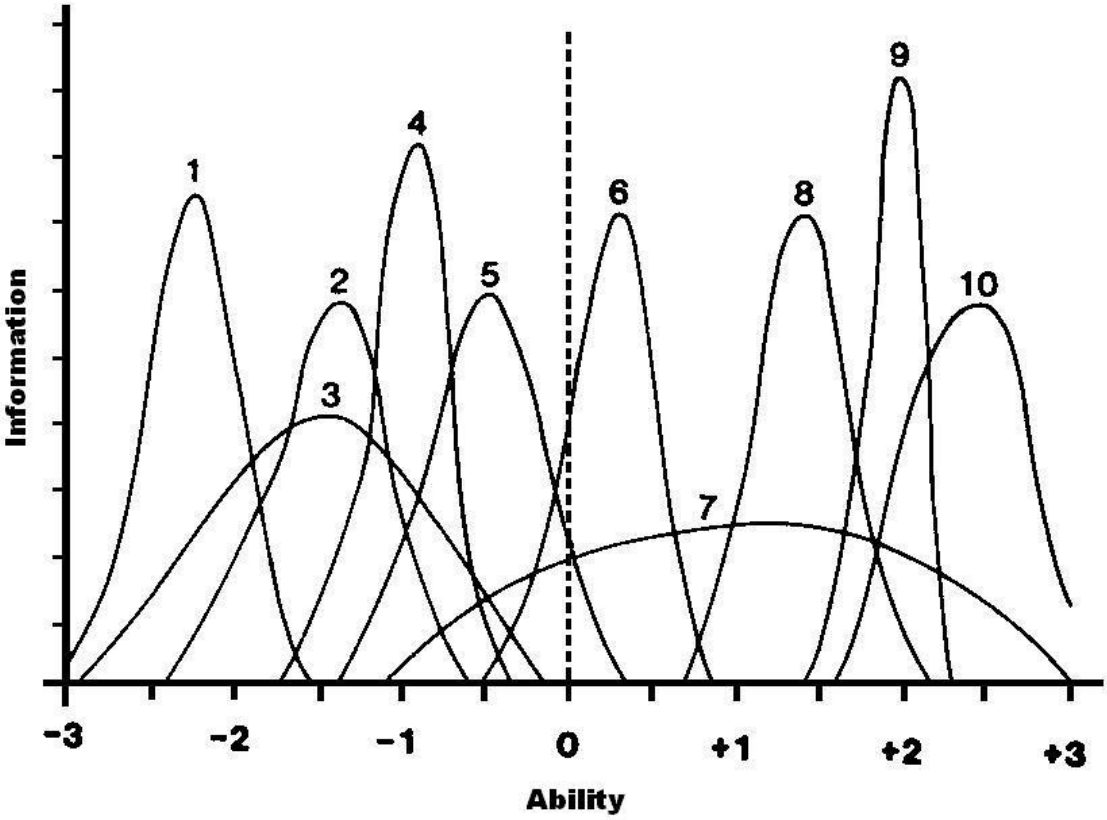

Figura 2.7: Representação das curvas de informação de 10 itens e o nível de habilidade com valor 0 (Weiss, 1985)

Quando um aluno responde corretamente um item sua estimação de habilidade é levemente elevada, enquanto que uma resposta incorreta conduz a um decréscimo da estimação de habilidade. Assumindo que o aluno respondeu incorretamente o item 6, o novo nível de habilidade estimado será menor, valor de $\theta = -1.0$ (como representado na Figura 2.8 na linha pontilhada). Para o nível de habilidade com valor -1.0, o item 7 fornece a menor quantidade de informação, sendo que os

itens 5 e 2 fornecem, de certa forma, a mesma quantia. O item 3, por sua vez, fornece ligeiramente mais informação que os outros anteriores e o item 4 fornece o maior grau de informação.

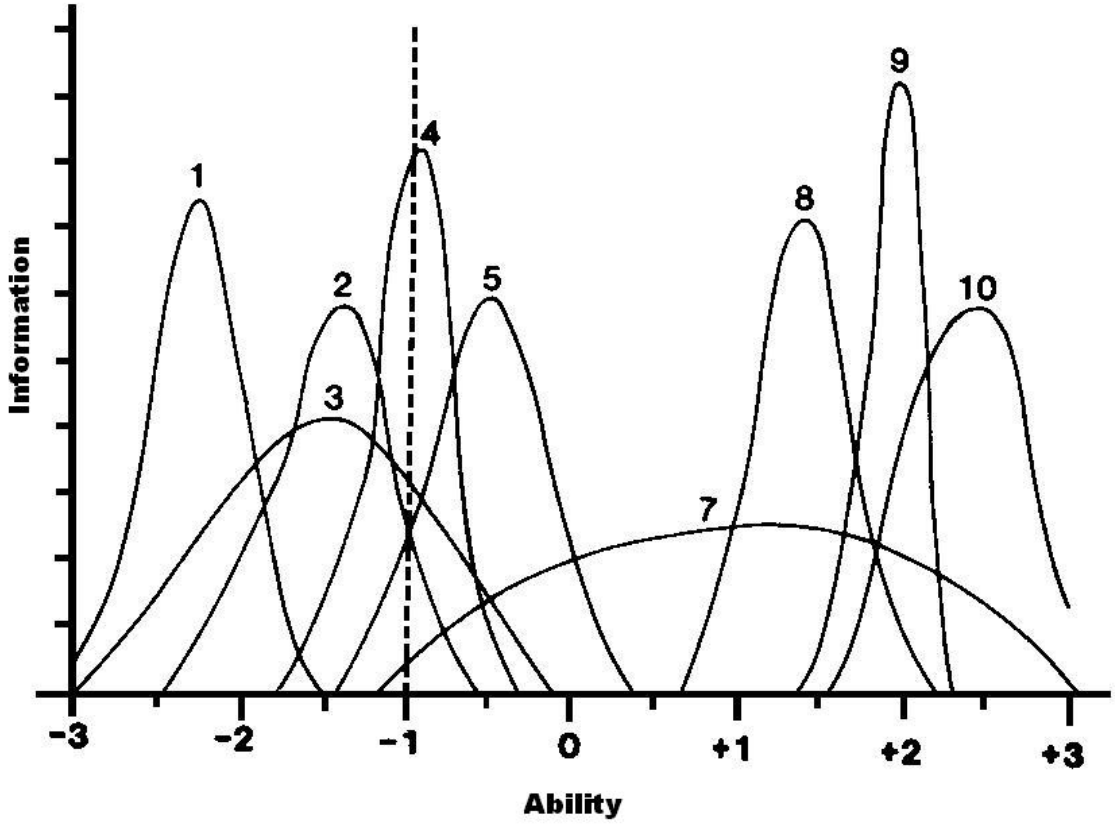

Figura 2.8: Representação das curvas de informação de 9 itens e o nível de habilidade com valor -1 (Weiss, 1985)

Depois que o item 4 foi administrado e respondido, o novo nível de habilidade é calculado, recebendo o valor de -0.5, supondo que o item 4 foi respondido corretamente (como indica a linha pontilhada da Figura 2.9). A este nível de habilidade, os itens 3, 5 e 7 fornecem diferentes níveis de informação, e como o item 5 fornece o maior nível, este é administrado.

Este processo de seleção de itens é mantido até que seja encontrado o critério de parada adotado ou os itens disponíveis se esgotem. Com o uso de um banco de itens reais, diferentes critérios de parada podem ser implementados.

Inerente a esta estratégia de seleção de itens, os métodos de pontuação também podem produzir um erro padrão após a administração de cada item, o qual serve como excelente critério de parada, de maneira que o teste pode ser continuado até que uma determinada medida de erro possa ser alcançada.

2.4.2 Método Bayesiano

O Método Bayesiano (Owen, 1975) de seleção de itens é similar ao Método da Máxima Informação, mas conceitualmente mais complicado e mais custoso de implementar (Kingsbury and Zara, 1989).

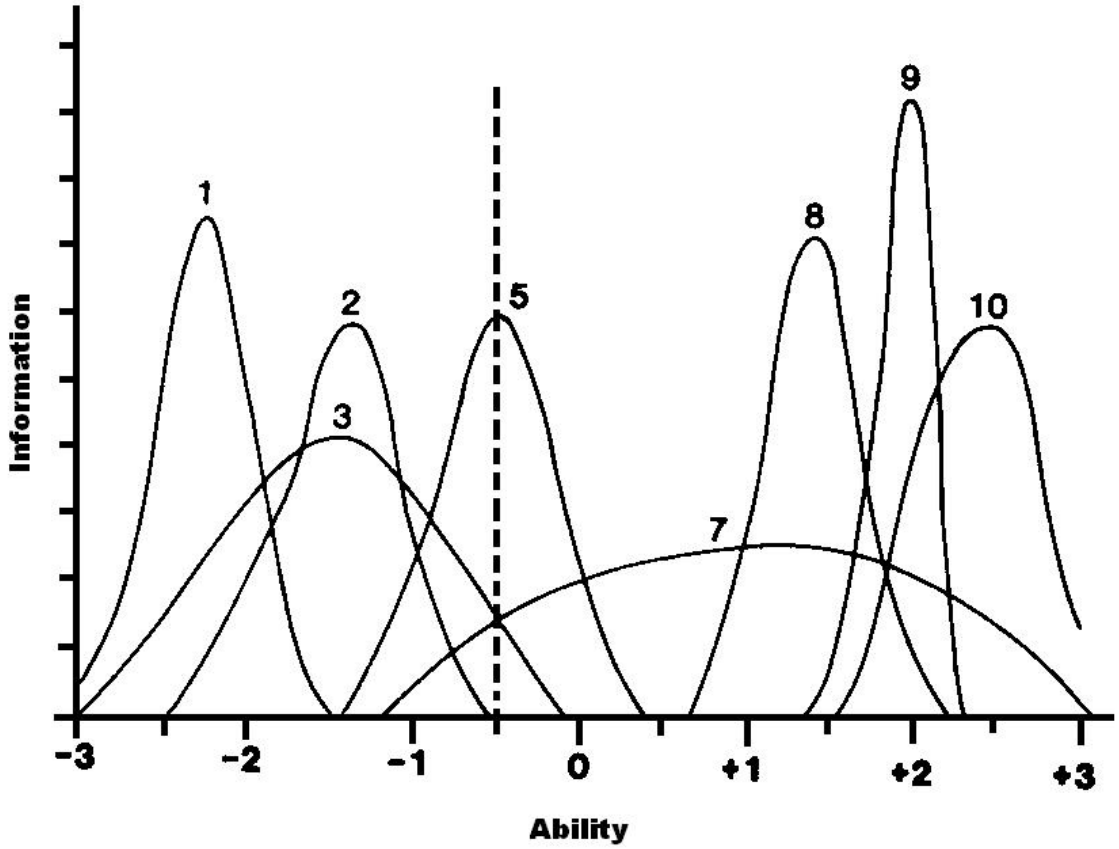

Figura 2.9: Representação das curvas de informação de 8 itens e o nível de habilidade com valor -0,5 (Weiss, 1985)

Cada indivíduo começa o teste com uma estimação inicial de habilidade e um intervalo de confiança associado a essa estimação. Eles são operacionalizados como a média e variância de uma distribuição normal da habilidade sendo medida. A medida que cada item é respondido, uma nova estimação de habilidade é calculada, usando a resposta do aluno e os valores da distribuição anterior, de maneira que uma nova distribuição de habilidade com um novo intervalo de confiança seja produzido. O Método Bayesiano de seleção escolhe um item que mais reduz esse novo valor da Variância, em outras palavras, tal variância é calculada para cada item disponível do banco, dada a habilidade corrente do indivíduo e os parâmetros dos itens, sendo escolhida a que fornece o menor valor.

As desvantagens associadas ao Método Bayesiano estão também relacionadas aos problemas de busca de itens, principalmente quando existe um grande banco de questões. Outra desvantagem que cabe salientar, diz respeito ao uso do valor da variância sobre a estimação de habilidade para selecionar o "melhor" item, pois em determinadas situações pode ocorrer uma escolha não apropriada, já que a verdadeira estimativa de habilidade não está sendo considerada.

2.4.3 Técnicas de Busca de Itens

Conforme dito anteriormente, a principal desvantagem relacionada com as abordagens de seleção de itens, tanto do Método da Máxima Informação quanto o Método Bayesiano, é a tarefa de busca de itens sobretudo quando existe um grande banco de itens. O tempo (*Delay*) existente entre a resposta do indivíduo e a administração do próximo item não pode ser longo. Portanto, Kingsbury and Zara (1989) sugerem três técnicas que podem ser utilizadas para eliminar ou diminuir o intervalo de busca entre uma questão e outra.

A primeira técnica requer que um algoritmo realize duas buscas enquanto um determinado estudante responde uma questão. Uma busca consiste em supor que o aluno acertou a questão corrente, e a outra supõe que ele a errou. Dessa forma, depois da sua resposta, o item apropriado é fornecido, e o *Delay* envolvido nesse processo é mínimo, já que o algoritmo já realizou a busca enquanto o estudante respondia a questão. Entretanto, esta técnica não é recomendada para testes em que o tempo de resposta requerido de cada questão é pequeno, pois pode aumentar o atraso entre a administração dos itens.

A segunda técnica de busca que pode diminuir o *Delay* consiste em limitar o número de itens disponíveis em dado momento do teste, considerando apenas os itens que possuem potencial para serem escolhidos de acordo com a habilidade provisória do estudante. Contudo, essa técnica pode não ser adequada quando o banco de itens é muito extenso, pois encontrar os itens "possíveis" e entre estes o mais adequado pode levar muito tempo.

A terceira técnica requer que todos os itens sejam pré-processados, antes do início do teste, criando uma tabela com os níveis de informação fornecidos, assim, o procedimento seleciona o primeiro item, que não tenha sido administrado, a partir de uma busca na tabela do item que oferece a maior informação à corrente habilidade do indivíduo.

2.5 Estimação de Parâmetros

O processo de estimação de parâmetros consiste no cálculo dos parâmetros pertencentes aos itens do banco, sendo estes referentes a um determinado modelo da Teoria de Resposta de Itens (TRI), ou ainda ao nível de habilidade θ de cada aluno durante um teste. No âmbito dos testes adaptativos informatizados, a estimação de parâmetros acontece em dois momentos distintos: o primeiro momento acontece durante a construção do banco de itens, fazendo parte do processo de calibração dos itens que irão compor o banco, antes da aplicação do teste. O segundo momento acontece durante a execução de um teste adaptativo, e corresponde ao cálculo do valor da nova habilidade de um aluno após o mesmo ter respondido a um item.

De maneira resumida, a tarefa de estimação de parâmetros baseia-se em descobrir o valor de uma determinada variável que mais se aproxima de seu valor real, dado para a execução do cálculo, uma amostra de valores aleatórios regidos por uma função ou modelo de distribuição.

Por exemplo: seja p uma variável qualquer que se deseja estimar, e X um conjunto de amostras aleatórias $(X_1, X_2, X_3, ..., X_n)$ pertencentes a um modelo de distribuição $f(X)=p$; então a tarefa é descobrir se existe alguma estimativa $H(X_1, X_2, X_3, ..., X_n)$ que possa ser empregada para encontrar um valor para a variável p. Dessa forma, o resultado da estimação, usando o modelo definido por $f(X) = p$, é um valor $\hat{p}$ que representa a estimativa de p, a qual mais se aproxima de seu valor real (Meyer, 1973).

Para a estimação dos parâmetros dos itens e do valor da habilidade dos estudantes em um teste adaptativo são utilizados principalmente o método da Estimativa de Máxima Verossimilhança[2] (EMV) e o Método Bayesiano. A seguir veremos a descrição de cada um desses métodos.

2.5.1 Método da Estimativa de Máxima Verossimilhança (EMV)

EMV é um procedimento geral de estimação de parâmetros que, segundo Meyer (1973) e Bilmes (1998), conduz a valores de estimativas razoáveis quando aplicado.

A fim de explicar o método EMV, definimos uma função de distribuição de probabilidade (fdp), ou modelo de distribuição, da forma $f(X|\theta)$ que depende da amostra X em relação a θ, que é a variável que desejamos estimar.

Seja X um conjunto de valores amostrais $(x_1, x_2, x_3, ..., x_n)$ conhecidos. Definimos a *função de verossimilhança* L como sendo a seguinte função em relação a θ:

$$L(x_1, x_2, x_3, ..., x_n|\theta) = f(x_1|\theta)f(x_2|\theta)f(x_3|\theta)...f(x_n|\theta)$$

ou, em termos de probabilidades podemos generalizar:

$$f(x_1|\theta)f(x_2|\theta)f(x_3|\theta)...f(x_n|\theta) = P(x_1|\theta)P(x_2|\theta)P(x_3|\theta)...P(x_n|\theta)$$

A estimativa da variável θ será o valor que maximizar a função $L(x_1, x_2, x_3, ..., x_n|\theta)$. Em outras palavras, devemos calcular o valor de θ que torna a função L máxima.

Definição: *A estimativa de máxima verossimilhança de θ ($\hat{\theta}$) baseada em uma amostra aleatória $x_1, x_2, x_3, ..., x_n$ é aquele valor de θ que torna $L(x_1, x_2, x_3, ..., x_n|\theta)$ máxima, considerada como função de θ para uma dada amostra $X_1, X_2, X_3, ..., X_n$, sendo L definida pela equação $L(x_1, x_2, x_3, ..., x_n|\theta) = f(x_1|\theta)f(x_2|\theta)f(x_3|\theta)...f(x_n|\theta)$ e, é geralmente referida como estimativa MV (Meyer, 1973).*

Na maioria das tarefas de estimação de parâmetros, a variável θ representa um único valor real isolado. Entretanto, em alguns casos é necessário estimar dois ou mais valores (por exemplo, a estimação dos parâmetros do modelo da TRI), em tais casos a variável θ pode representar um vetor contendo as variáveis que se quer estimar, por exemplo, $\theta = (a, b, c)$.

Dado o fato de que devemos encontrar o valor de θ que maximiza a função L, é conveniente aplicarmos a função *ln(x)*, visto que o logaritmo natural é uma função crescente de x. Portanto, com a aplicação do *ln* sobre a função L, a equação de distribuição fica:

$$lnL(x_1, x_2, x_3, ..., x_n|\theta)$$

e alcança o valor máximo de θ para o mesmo valor que faria a função L. Por isso, em linhas gerais admitimos que θ seja um valor real, ou um vetor de valores reais, e que a função L seja uma função derivável em relação a θ, possibilitando, assim, conseguir a estimativa de máxima verossimilhança pela solução de:

$$\frac{\partial}{\partial\theta} lnL(x_1, x_2, x_3, ..., x_n|\theta) = 0.$$

o qual é conhecida como *equação de verossimilhança.*

[2]Maximum Likelihood Estimation - MLE em inglês

Se a variável θ for um vetor de valores, como por exemplo, $\theta = (\alpha, \beta)$, a equação acima deverá ser substituída pelas equações simultâneas sobre cada variável pertencente ao vetor, assim as equações parciais ficam:

$$\frac{\partial}{\partial \alpha} lnL(x_1, x_2, x_3, ..., x_n|\alpha, \beta) = 0.$$

$$\frac{\partial}{\partial \beta} lnL(x_1, x_2, x_3, ..., x_n|\alpha, \beta) = 0.$$

O Algoritmo EM (Expectation-Maximization)

O algoritmo EM pode ser entendido como um método geral para encontrar as estimativas de máxima verossimilhança dos parâmetros de um determinado modelo de distribuição $f(x|\theta)$ para uma dada amostra de valores aleatórios e, os parâmetros que se deseja estimar (Bilmes, 1998). Como descrito anteriormente, podemos assumir que o conjunto de amostras é representado por X e a variável que se deseja estimar é representado por θ.

A execução do algoritmo EM gera seqüências de estimativas θ^s, em que $s = 1, 2, 3...n$ e representa o número de iterações de cada passo do algoritmo, dado o valor inicial θ^0. Há dois passos a cada iteração: o passo E *(Expectation)* e o passo M *(Maximization)*.

No passo E o valor da *esperança condicional* sobre o modelo de distribuição $lnf(x|\theta)$ representa a distribuição condicional da variável que se deseja estimar (θ), sendo fornecidos o conjunto de amostras X, assim, considerando a variável θ (que desejamos estimar), a equação que produz o valor da *esperança condicional* a cada iteração s é descrita da seguinte forma:

$$Q(\theta|\theta^s) = E_\theta \left\{ ln \left[\prod_{i=1}^{n} f(x_i|\theta) \right] |X, \theta^s \right\}$$

com $s = 1, 2, 3, ..., n$ e sendo o valor θ^s calculado no passo anterior $(s-1)$.

O passo M (maximização) encontra valores de θ que maximizam a função de distribuição $lnf(x|\theta)$. Cada iteração s tem o objetivo de maximizar a função $Q(\theta|\theta^s)$, de maneira que as novas estimativas $\theta^{(s+1)}$ produzidas no passo M, a cada iteração s, sejam usadas no passo E da iteração $(s+1)$ e, assim sucessivamente, até a convergência do modelo ser obtida (Bilmes, 1998; Woodruff and Hanson, 1997; Hanson, 1996).

No contexto dos testes adaptativos, podemos utilizar a estimativa de verossimilhança em conjunto com o algoritmo EM, tanto para estimar os parâmetros dos itens (**a** e/ou **b** e/ou **c**) como para estimar o valor da habilidade θ do aluno. Dado este fato, é importante descrever as respectivas funções de estimação do algoritmo EM para a tarefa de estimação desses parâmetros.

Estimatição dos Parâmetros dos Itens

Esta subseção apresenta o algoritmo EM para calcular as estimativas dos parâmetros dos itens, os quais compõem um banco que será utilizado em um teste adaptativo. Tais itens possuem respostas dicotômicas, ou seja, a resposta de cada item deve ter valor 1, representando uma resposta correta, ou 0, caso seja incorreta.

Para itens dicotômicos a equação do algoritmo EM pode ser escrita em termos da Função de Resposta de um Item $P_i(\theta_j)$; que conforme visto no início deste capítulo, é a probabilidade do aluno j com nível de habilidade θ responder um item i corretamente. Dado que $\triangle$ é o conjunto de

parâmetros de um determinado item i, $\triangle_i = (a_i, b_i, c_i)$, a probabilidade de um aluno responder um dado item de forma correta é dada por: $P(\theta_j|\triangle_i)$ e de forma incorreta: $Q(\theta_j|\triangle_i) = 1 - P(\theta_j|\triangle_i)$.

Assim, considerando um vetor de respostas U, sendo que $U_i = 1$ representa uma resposta correta e $U_i = 0$ uma incorreta, o modelo de distribuição dado o nível de habilidade θ e os parâmetros de um item é:

$$f(U_{ij}|\theta_j, \triangle_i) = P(\theta_j|\triangle_i)^{U_{ij}} \cdot Q(\theta_j|\triangle_i)^{1-U_{ij}}$$

$$f(U_{ij}|\theta_j, \triangle_i) = \prod_{i=1}^{I} P(\theta_j|\triangle_i)^{U_{ij}} \cdot Q(\theta_j|\triangle_i)^{1-U_{ij}}$$

e a equação do algoritmo EM para o passo E é definida como:

$$Q(\triangle|\triangle^s) = E\left\{ ln\left[\prod_{i=1}^{J} f(U_{ij}|\theta_j, \triangle_i)\right] |\theta_j, U_{ij}, \triangle^s \right\}$$

com $s = 1, 2, 3, ..., n$ representando o número de iterações.

Estimação de Habilidade θ

Existem diferentes formas de estimar a habilidade inicial de um estudante. Entre elas estão uma simples que consiste em atribuir um valor aleatório entre -1.0 e 1.0 a cada indivíduo, ou ainda um valor *default* e a partir daí, o ajuste da habilidade se dá a partir das respostas obtidas durante o teste.

Outro método adotado na estimação da habilidade inicial é a Estimativa de Máxima Verossimilhança. Este método estima a habilidade a partir do fornecimento de uma amostra comum de itens a cada estudante, o qual os responde, e baseado na quantidade de respostas certas e erradas dessa amostra o valor de θ é calculado. Suponha que um determinado aluno recebeu uma amostra de cinco itens que devem ser respondidos, sendo que seu vetor de respostas foi o seguinte:

$$U = (u_1, u_2, u_3, u_4, u_5) = (10110)$$

em que um valor de $U = 1$ representa um resposta correta e o valor de $U = 0$ representa uma resposta incorreta. Dado o vetor de respostas, é possível estimar a habilidade inicial de um estudante usando a seguinte equação:

$$\theta_0 = ln(\frac{r}{(n-r)})$$

em que r é quantidade de respostas corretas e n a quantidade de itens propostos.

No exemplo acima, o estudante acertou 3 questões ($r = 3$) de um total de 5 ($n = 5$) e sua habilidade inicial vale:

$$\theta_0 = ln(\frac{3}{(5-3)}) = 0.4054$$

e com este valor de $\theta = 0.4054$ é possível iniciar o processo de teste (Hambleton and Swaminathan, 1985).

Outra maneira de estimar o valor da habilidade inicial também pode ser empregada de acordo com o conhecimento prévio dos alunos, como utilizado no CBAT-2 (Huang, 1996). Quando o

ambiente na qual a avaliação está inserida já está padronizado, com as características pessoais mais ou menos padrão, é possível por meio de informações empíricas, tais como histórico escolar, experiências vividas, currículo, etc. traçar um perfil do candidato de maneira que essa análise possa ser quantificada e representar a habilidade inicial para que o teste adaptativo seja iniciado.

O método da Estimativa de Máxima Verossimilhança também pode ser utilizado para estimar a habilidade de um indivíduo enquanto o teste está ocorrendo, ou seja, a medida que cada estudante responde um dado item, e dependendo de sua resposta é possível estimar sua habilidade para que outro item seja selecionado e o processo de teste continue até o critério de parada ser encontrado.

Da mesma forma a probabilidade de uma resposta correta, dado um nível de habilidade θ é: $P(U_i = 1|\theta)$ e incorreta $P(U_i = 0|\theta)$. Assim, dada toda a seqüência de respostas de um candidato em um teste, o cálculo da probabilidade de respostas corretas e incorretas é dado por:

$$P(U_i|\theta) = P(U_i = 1|\theta)^{U_i} \cdot P(U_i = 0|\theta)^{1-U_i}$$

$$P(U_i|\theta) = P_i^{U_i} \cdot (1 - P_i)^{1-U_i}$$

$$P(U_i|\theta) = P_i^{U_i} \cdot Q_i^{1-U_i}$$

em que $Q_i = 1 - P_i$.

Dessa maneira, se um estudante com nível de habilidade θ responde n itens, o vetor de respostas desse estudante pode ser representado por: $U_i = (U_1, U_2, U_3, ..., U_n)$ e a probabilidade de suas respostas será: $P(U_1, U_2, U_3, ..., U_n|\theta)$. Generalizando, pode ser definido que dada a habilidade θ e as n respostas de um estudante, a função de verossimilhança (*Likelihood Function*) expressa como:

Seja

$$P(U_1, U_2, U_3, ..., U_n|\theta) = P(U_i|\theta), P(U_2|\theta), P(U_3|\theta), \ldots, P(U_n|\theta)$$

$$'' = \prod_{i=1}^{n} P(U_i|\theta)$$

$$'' = \prod_{i=1}^{n} P_i^{U_i} \cdot Q_i^{(1-U_i)}$$

$$L(u_1, u_2, u_3, \ldots, u_n|\theta) = \prod_{i=1}^{n} P_i^{U_i} \cdot Q_i^{(1-U_i)}$$

o que significa dizer que quando $u_i = 1$ o termo Q_i torna-se 0 (zero) e quando $u_i = 0$ o termo P_i assume valor 0 (zero), saindo da fórmula.

Por exemplo, supondo que o vetor de respostas de cinco itens de um determinado aluno seja o seguinte:

$$u = (10110)$$

a função de verossimilhança fica:

$$L(u|\theta) = P_1 \cdot Q_2 \cdot P_3 \cdot P_4 \cdot Q_5$$

Dessa maneira, inserindo a função crescente ln, a função de Estimativa de Máxima Verossimilhança (MLE) pode ser expressa como:

$$ln(L(u|\theta)) = lnP_1 + lnQ_2 + lnP_3 + lnP_4 + lnQ_5$$

que, generalizando, a função MLE assume a forma a seguir:

$$ln(L(u|\theta)) = \sum_{i=1}^{n} [u_i \cdot ln(P_i) + (1 - u_i) \cdot ln(Q_i)]$$

O cálculo da nova estimativa de habilidade θ, depois de um estudante responder um dado item, pode ser dada por meio da seguinte equação: $\theta_{i+1} = \theta_i - h$ na qual o valor h é dado por:

$$h = \frac{[\frac{d}{d\theta} \cdot ln(L(u|\theta))]}{[\frac{d^2}{d\theta^2} \cdot ln(L(u|\theta))]}$$

ou, em outras palavras, o valor h pode ser visto como a divisão entre a primeira derivada e segunda derivada da função de Máxima Verosimilhança.

Para estimar o nível de habilidade θ de um determinado aluno j utilizando o algoritmo EM, basta alterar a equação do passo E, descrita na subseção anterior da seguinte maneira:

$$Q(\theta|\theta^s) = E\left\{ ln\left[\prod_{i=1}^{n} f(U_{ij}|\theta_j, \triangle_i)\right] |\triangle_i, \theta^s \right\}$$

com $s = 1, 2, 3, ..., n$ representando o número de iterações.

Método Bayesiano de Estimativa de Habilidade

Quando qualquer informação prévia sobre a distribuição de habilidade dos estudantes é disponível antes do início do teste, o Método Bayesiano de estimação de habilidade pode fornecer resultados significativos (Hambleton and Swaminathan, 1985). Dado que a habilidade de um aluno é representada por θ_a, onde a representa cada aluno variando de *1, 2, ... N*, é possível considerar, baseado nessas informações prévias, que o valor da habilidade seja um valor aleatório dentro de um intervalo que deve ser especificado. Por exemplo, os projetistas do teste acreditam que uma pequena porção de estudantes tenham habilidades fora de um intervalo, então a maioria dos estudantes tem níveis de habilidade em torno de uma média. Assim, pode ser indicado que a habilidade obedece a uma distribuição normal dada por:

$$\theta_a \sim N(\mu, \phi)$$

sendo que $N(\mu, \phi)$ denota uma distribuição normal com média μ e variância ϕ ,de maneira que seus valores precisam ser especificados. Segundo Owen (1975), no contexto dos testes adaptativos esses valores podem ser $\mu = 0$ e $\phi = 1$.

Assume-se que a distribuição de habilidade de θ_a pode ser a função logística $f(\theta)$:

$$f(\theta) = \frac{\exp(\theta)}{[1 + \exp(\theta)]^2}$$

O teorema de Bayes é o principal componente do Método Bayesiano de estimação de parâmetros e relata as probabilidades condicionais:

$$P(B|A) = P(A|B) \cdot P(B)/P(A)$$

No contexto da estimação de habilidade, A pode ser considerada como θ_a e B como um conjunto de respostas observadas em n itens e é denominado como u. Então a equação das probabilidades condicionais pode ser transformada em:

$$P(\theta_a|u) = P(u|\theta_a)P(\theta_a)/P(u)$$

2.6 Procedimentos de Pontuação

Vistos os principais métodos de estimação de habilidade, é necessário estabelecer critérios que estabelecem a verdadeira pontuação obtida por um estudante em um teste. Os procedimentos de pontuação, então, tem a responsabilidade converter a habilidade estimada durante todo teste em pontos convencionais, que permitem classificar o estudante após a avaliação.

As diferenças entre a escala de distribuição de habilidade e a verdadeira pontuação do estudante em um determinado teste, dependem, entre outras coisas, do nível de dificuldade que cada teste administrado representou, de acordo com a habilidade desenvolvida pelo indivíduo. Dessa forma, não é possível realizar uma conversão direta, sendo necessário considerar o nível de dificuldade do itens fornecidos ao estudante e as variações de habilidade (quando acontece um acerto ou erro) ocorridas durante um teste. Por exemplo, se alguém aplicasse dois testes medindo a mesma habilidade para um mesmo grupo de estudantes, não sendo necessariamente paralelos, duas diferentes habilidades, e consequentemente suas pontuações, seriam distintas; pois os itens disponibilizados, os níveis de dificuldade e as habilidades estimadas durante o teste foram diferentes. Ou seja, é preciso considerar os itens fornecidos, bem como seus parâmetros, e as habilidades calculadas, para encontrar de maneira justa a pontuação final do aluno.

A pontuação verdadeira possui uma interessante relação com a habilidade θ de um indivíduo. Segundo Hambleton and Swaminathan, (1985), a pontuação observada (obtida durante um teste) de um estudante é definida como r, baseada no somatório dos n itens apresentados em um teste. Então:

$$r = \sum_{i=1}^{n} U_i$$

onde U_i representa a resposta de um item i e pode ser 1 ou 0 (zero).

Em linhas gerais, a medida de pontuação verdadeira é dada pelo somatório dos itens que foram respondidos corretamente durante um teste, dividido pelo total de itens, e pode ser representada pela equação:

$$\pi = \frac{1}{n} \cdot \sum_{i=1}^{n} U_i$$

Entretanto, esse método não considera, tanto os parâmetros dos itens disponibilizados no teste, quanto as habilidades individuais de cada indivíduo, já que apenas as questões certas serão somadas. Então, é necessário que a equação acima considere as características dos itens fornecidos para um estudante com nível de habilidade θ. Assim, a equação torna-se:

$$\pi|\theta = \frac{1}{n} \cdot \sum_{i=1}^{n} (U_i|\theta)$$

e visto que Ui é uma variável que assume valores 1 ou 0, segue:

$$(U_i|\theta) = (U_i = 1)P_i(U_i = 1|\theta) + (U_i = 0)P_i(U_i = 0|\theta)$$

$$(U_i|\theta) = 1 \cdot P_i(U_i = 1|\theta) + 0 \cdot P_i(U_i = 0|\theta)$$

$$(U_i|\theta) = P_i(U_i = 1|\theta)$$

$$(U_i|\theta) = P_i(\theta)$$

que é a Função de Resposta de um Item dado o modelo de resposta adotado. Então, a função de conversão do nível de habilidade de um estudante ao final de uma avaliação para a sua pontuação verdadeira é definida da seguinte forma:

$$\pi|\theta = \pi = \frac{1}{n} \cdot \sum_{i=1}^{n} P_i(\theta)$$

Segundo Hambleton and Swaminathan, (1985), esta função é conhecida como ***Função Característica de um Teste*** (*Test Characteristic Function*) e sua curva no gráfico bidimensional é referida como ***Curva Característica de um Teste*** (*Test Characteristic Curve*) e é utilizada no teste adaptativo CBAT-2 (Huang, 1996) e no sistema SIETTE (Rios et al., 1998) para o cálculo da pontuação verdadeira.

Capítulo 3

Conclusão

A teoria dos testes adaptativos, englobando seus modelos de resposta, as abordagens de seleção, métodos de estimação de habilidade, e a calibração e organização do conteúdo do banco de itens, parece ser a forma mais justa e honesta de avaliar um indivíduo num processo educativo. No entanto, a prática e a implementação desses modelos dependem do acerto e cuidado de muitos detalhes inerentes a este processo. Assim, é necessário um esforço conjunto e coordenado, com o envolvimento de vários especialistas de diferentes áreas (estatísticos, analistas e educadores), além da necessidade boas ferramentas (computadores e aplicativos), para promover uma boa avaliação. Não podemos esquecer é claro que todas as atividades necessárias para a criação e implantação de uma avaliação adaptativa possui um custo, relativamente alto, já que abrange várias especialidades e equipamentos.

Entretanto, como pôde ser visto neste relatório, a complexidade dos modelos estatísticos da Teoria de Resposta de Itens, incluindo os cálculos matemáticos dos métodos de estimação de parâmetros e seleção itens, favoreceu o desenvolvimento de programas dedicados especialmente a essa tarefa, facilitando assim a implantação dos testes adaptativos em um cenário de avaliação. Por sua vez, o desenvolvimento de tais programas propiciou o uso mais intenso do computador nesses ambientes, acelerando o processo de implantação de uma avaliação adaptativa. Como exemplo podemos citar o programa XCALIBRE (www.assess.com) que implementa os principais modelos de estimação de parâmetros dos itens para a calibração de um banco de itens. Como principais exemplos de programas de aplicação de testes adaptativos podemos citar o **FastTest Professional** (www.assess.com) e o **CATGlobal System** (www.catinc.com), os quais reúnem os principais métodos seleção de itens e estimação de habilidade pertencentes à TRI.

O trabalho desenvolvido por Oliveira (2002) mostra as principais diretrizes para a criação de um teste adaptativo, destacando principalmente os processos de criação e calibração de um banco de itens, voltado para o Exame de Proficiência de Inglês (EPI) do ICMC[1] , além da modelagem de um sistema adaptativo, considerando os métodos e modelos envolvidos neste domínio.

[1]Disponível no endereçco http://www.nilc.icmc.sc.usp.br/staff/capteap/

Bibliografia e Referências

(Baker, 1992) Baker, F. (1992). *Item Response Theory Parameter Estimation Techniques.* MARCEL DEKKER, INC.

(Bilmes, 1998) Bilmes, J. A. (1998). A gentle tutorial of the em algotithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report TC-97-021, International Computer Science Institute.

(Hambleton and Swaminathan, 1985) Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Kluwer Nijhoff Publishing.

(Huang, 1996) Huang, S. X. (1996). A content-balanced adaptive testing algorithm for computer-based training systems. In *ITS'96 3th International Conference in ITS*, Montreal.

(Kingsbury and Zara, 1989) Kingsbury, G. G. and Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied measurement in education*, 2(4):16.

(Meyer, 1973) Meyer, P. L. (1973). *Probabilidade: aplicações à estatística.* Ao Livro Técnico S.A.

(Olea et al., 1999) Olea, J., Ponsoda, V., and Prieto, G. (1999). *Tests Informatizados Fundamentos y Aplicaciones.* Ediciones Pirámede.

(Oliveira, 2002) Oliveira, L. H. M. (2002). Testes adaptativos sensíveis ao conteúdo do banco de itens: Uma aplicação em exame de proficiência em inglês para pós-graduação. Master's thesis, Universidade de São Paulo - USP São Carlos-SP. Dissertação de Mestrado.

(Owen, 1975) Owen, R. J. (1975). A baysesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70(350).

(Rios et al., 1998) Rios, A., de la Cruz, J. L. P., and Conejo, R. (1998). Siette: Intelligent evaluation system using tests for teleeducation. In *ITS'98 4th International Conference in ITS*, Texs.

(Rudner, 1998) Rudner, M. L. (1998). An on-line, interactive, computer adaptative testing mini-tutorial. [www.ericae.net/scripts/cat/catdemo.html] acessado em (14/02/2001).

(Weiss, 1985) Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and clinical Psychology*, 53(6):15.