Modelos para Testes com Respostas Dicotômicas com Principal Enfoque em Teoria de Resposta de Item

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida por Márcia Milena Pivatto e aprovada pela Comissão Julgadora.

Campinas, 09 de Julho de 1.992

Clarice Azevedo de Luna Freire

Profa. Dra. Clarice Azevedo de Luna Freire

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Título de Mestre em Estatística

UNIVERSIDADE ESTADUAL DE CAMPINAS
DEPARTAMENTO DE ESTATÍSTICA - IMECC

Modelos para Testes com Respostas Dicotômicas com Principal Enfoque em Teoria de Resposta de Item

Márcia Milena Pivatto

Orientação
Profa. Dra. Clarice Azevedo de Luna Freire

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência de Computação da Universidade Estadual de Campinas para obtenção de título de Mestre em Estatística
Campinas - S.P.
1992

Agradecimentos

Gostaria de agradecer aos professores do departamento de estatística pela contribuição à minha formação. Em especial a professora Clarice, pela orientação, dedicação e amizade, que foram essenciais para a realização deste trabalho. A professora Eugenia Charnet pela sugestão do tema e orientação inicial. E ao professor Jonathan Biele e ao amigo Aloisio F. Ribeiro pela dedicação ao ler os originais. A Maria T. Albanese pela simpatia e por ter nos cedido o exemplo referenciado neste trabalho.

Aos funcionários do departamento pela colaboração. Aos amigos que conquistei durante o mestrado, pelo companheirismo e apoio.

A minha mãe, a minha avó pelo amor e carinho que sempre me dedicaram. A Fernando pelo companheirismo e compreensão, principalmente nas horas mais difíceis.

Conteúdo

1 INTRODUÇÃO **1**

2 TEORIA CLÁSSICA DE TESTES **4**

2.1 Introdução 4

2.2 Fundamentos da Teoria Clássica 5

2.3 Estimação da Fidedignidade e Validade de um Teste 14

2.4 Conclusão 15

3 TEORIA DE RESPOSTA DE ITEM **16**

3.1 Introdução 16

3.2 Função de Resposta de Item 17

3.3 Estimação da Habilidade e dos Parâmetros de Itens 19

3.3.1 Métodos da Máxima Verossimilhança 20

3.3.2 Procedimento Bayesiano 23

3.3.3 Estimação pelo Método do Mínimo Logito Qui-Quadrado 24

3.4 Intervalo de Confiança Assintótico para a Habilidade 25

3.5 Sugestão de Birnbaum para a Construção de um Teste 26

3.6 Conclusão 27

4 ESTIMAÇÃO DA FUNÇÃO DE RESPOSTA DE ITEM 29

4.1 Introdução 29

4.2 Procedimento de Baker 30

4.2.1 Primeiro Estágio: Estimação dos Parâmetros de Itens segundo o Método do Mínimo Logito Qui-Quadrado 30

4.2.2 Segundo Estágio: Estimação das Habilidades dos Indivíduos segundo o Método de Máxima Verossimilhança 36

4.3 Simulações 37

4.3.1 Descrição do planejamento das simulações 39

4.3.2 Resultados das simulações 41

4.4 Intervalos de Confiança Bootstrap para a Habilidade 53

4.5 Ilustração 57

4.6 Conclusão 61

A MÉTODO DO "SCORING" DE FISHER **62**

B PROGRAMA PARA ESTIMAÇÃO DE PARÂMETROS DA FUNÇÃO DE RESPOSTA DE ITEM **66**

Lista de Figuras

3.1 Função de resposta de item - modelo logístico 18

3.2 (a) curva de informação alvo (sólida) e as curvas de informação para testes com: 3, 7, 11 e 15 itens respectivamente; (b) curvas de informação dos itens utilizados para aproximar a curva de informação alvo 28

4.1 Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker padrão, de um teste "mediano", segundo a primeira fase de simulações . . . 45

4.2 Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker padrão, de um teste "fácil", segundo a primeira fase de simulações 46

4.3 Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker padrão, de um teste "difícil", segundo a primeira fase de simulações 47

4.4 Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker padrão, de um teste "misturado", segundo a primeira fase de simulações . . 48

4.5 Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker modificado, de um teste "misturado", segundo a primeira fase de simulações . . 49

4.6 Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker modificado, de um teste "misturado" em indivíduos gerados segundo distribuição truncada em $\theta = 3$, segundo a terceira fase de simulações 50

4.7 Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker modificado, de um teste "fácil" em indivíduos gerados segundo distribuição truncada em $\theta = 3$, segundo a terceira fase de simulações 51

4.8 Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker modificado, de um teste "difícil" em indivíduos gerados segundo distribuição truncada em $\theta = 3$, segundo a terceira fase de simulações 52

4.9 Resultados obtidos pelo método bootstrap (paramétrico), 300 amostras bootstrap, a partir de amostra de 200 indivíduos gerados de uma distribuição normal padrão e 60 itens de um teste "misturado" 56

4.10 Intervalos de confiança para as habilidades dos indivíduos do exemplo apresentado como ilustração (bootstrap paramétrico, $k_b = 300$) 60

Lista de Tabelas

4.1 Esquema das três fases de simulações 40

4.2 Resultados referentes à primeira fase de simulações, com o objetivo de comparar os procedimentos de Baker padrão e o modificado (correlações e erros quadráticos médios (EQM) observados entre parâmetros simulados e estimativas) 42

4.3 Resultados referentes à segunda fase de simulações, com o objetivo de verificar a sensibilidade do procedimento de Baker modificado com relação ao número de indivíduos e itens (correlações e erros quadráticos médios (EQM) observados entre parâmetros simulados e estimativas); testes medianos . . . 43

4.4 Resultados referentes à terceira fase de simulações, com o objetivo de verificar efeito do procedimento de Baker modificado com relação a distribuições com acumulo de indivíduos com habilidades altas e baixas (correlações e erros quadráticos médios (EQM) observados entre parâmetros simulados e estimativas) 44

4.5 Estimativas referentes aos parâmetros de itens a e b, obtidas pelos métodos de máxima verossimilhança marginal (MVM) e de Baker 58

4.6 Frequências observadas dos padrões de resposta às sete questões; ordenação dos padrões de resposta segundo o método usado por Knott et. al. (MVM) e segundo o método de Baker modificado (Baker) 59

Capítulo 1

INTRODUÇÃO

Existem dois tipos principais de testes: os compostos por itens com respostas abertas e os dicotômicos. Aos itens com respostas abertas são normalmente atribuídas notas que podem, por exemplo, variar entre 0 e 10, já aos itens dicotômicos é simplesmente considerado se a resposta é certa ou errada. Os testes de múltipla escolha são exemplos de testes dicotômicos.

O maior problema dos testes educacionais (psicométricos, etc) reside no fato de que, normalmente, somente uma aplicação de um mesmo teste pode ser efetuada a uma mesma pessoa. Respostas associadas a várias replicações podem ser afetadas por efeitos de fadiga, fatores psicológicos, mudanças por aprendizagem, etc. Muitas vezes deseja-se fazer inferências individuais, sobre o grupo de indivíduos, ou ainda estudar a eficácia do teste. Como por exemplo, em um teste para avaliar o QI de indivíduos deseja-se fazer inferências sobre cada indivíduo, ou sobre o grupo de indivíduos, e ainda saber se teste é eficiente para avaliar o QI dos indivíduos.

Neste trabalho duas metodologias, encontradas na literatura, desenvolvidas para o estudo de testes educacionais são apresentadas. A chamada teoria clássica de testes é baseada principalmente em correlações entre as notas de um teste e as de outro teste padrão, ou

em correlações entre as notas de subtestes de um mesmo teste. A outra metodologia, denominada teoria de resposta de item (IRT - Item Response Theory), descreve itens dicotômicos de um teste através da função de resposta de item. A função de resposta de item define a probabilidade de acerto ao item por indivíduo dotado de certa habilidade. Esta função compreende parâmetro referente ao indivíduo e parâmetros referentes ao item. A teoria clássica avalia o teste como um todo, ou descreve o comportamento do grupo de indivíduos em relação ao teste que lhes foi aplicado. A teoria de resposta de item disseca o teste em parâmetros de itens e parâmetro referente ao indivíduo, com a finalidade de poder avaliar o indivíduo ou ainda construir testes a partir de itens com parâmetros previamente conhecidos.

A teoria clássica é a mais antiga, e vem sendo estudada por autores como Kuder e Richardson (1937) e Rulon (1939). Gulliksen (1950) publicou "Theory of Mental Test", que deu origem a publicações como Lord e Novick (1968), Carmines e Zeller (1981), Vianna (1982) e outros. O Capítulo 2 apresenta o modelo usado na teoria clássica, assim como alguns procedimentos de estimação mais comumente utilizados.

Vários autores, por exemplo, Lord (1953, 1980) e Stocking e Pearlman (1989), analisam aspectos da teoria de resposta de item, que é o principal enfoque deste trabalho. Modelos desta teoria são apresentados no Capítulo 3, assim como alguns procedimentos de estimação para os parâmetros da função de resposta de item. No Apêndice A é encontrado o desenvolvimento teórico de um procedimento utilizado em alguns destes procedimentos de estimação.

A teoria clássica apresenta a vantagem de ser pouco complexa, uma vez que a fidedignidade e a validade de um teste são correlações entre as notas obtidas em testes. Problemas referentes à estimação das correlações são abordados no Capítulo 2. Entretanto, o principal enfoque neste trabalho é a teoria de resposta de item, cujos modelos mais elaborados conduzem a resultados mais ricos em informações. O inconveniente desta teoria é a necessidade

de se estimar um número relativamente grande de parâmetros.

Na teoria de resposta de item uma suposição fundamental é de que a probabilidade de acerto a um item depende da habilidade do indivíduo e da natureza do item. Em muitos estudos de simulações, documentados na literatura, são geradas amostras pseudo-aleatórias de indivíduos com habilidades provenientes de distribuição normal. Além disso, algumas vezes o modelo básico da teoria de resposta de item é modificado de forma a tornar-se necessária a definição de uma função de densidade de probabilidade para a habilidade do indivíduo, quando, geralmente, a densidade do modelo normal é escolhida. Neste estudo, simulações foram planejadas onde as habilidades dos indivíduos foram definidas através da geração de amostras pseudo-aleatórias de distribuições normais, e também de outras distribuições (não normais), apresentadas no Capítulo 4. Observou-se que em alguns trabalhos eram simulados testes compostos exclusivamente por itens fáceis, exclusivamente por itens difíceis ou exclusivamente por itens moderados. Neste trabalho, testes compostos por itens com uma maior variação possível de dificuldade foram os que apresentaram melhores resultados. Nas simulações utilizou-se o procedimento de Baker, descrito detalhadamente neste mesmo capítulo. Este procedimento foi escolhido por ser menos honeroso computacionalmente, aliado aos fatos de exigir um menor número de suposições e de ter sido verificado por vários autores que o procedimento de Baker resulta em estimativas similares às obtidas pelos procedimentos usuais. Uma modificação, proposta neste trabalho, em etapa básica do procedimento de Baker (1987), apresentou melhores estimativas nas situações simuladas que o procedimento de Baker original. É investigado o método bootstrap (Efron, 1979) para a construção de intervalos de confiança para os parâmetros da função de resposta de item, sendo de maior interesse a estimação dos parâmetros referentes aos indivíduos. Um conjunto de dados analisado por Knott, Albanese e Galbraith (1991), referente a uma pesquisa de opinião, é usado como ilustração. Esta ilustração vem destacar a utilização desta teoria em

testes que não sejam somente os educacionais.

Para a realização das simulações foi elaborado um programa em linguagem IML (Interactive Matrix Language), parte integrante do SAS (Statistical Analysis System), para a estimação de parâmetros da função de resposta de item. Este programa é apresentado no Apêndice B.

Capítulo 2

TEORIA CLÁSSICA DE TESTES

2.1 Introdução

Algumas vezes, deseja-se observar uma variável, mas esta não pode ser observada diretamente e sim através de outra variável. Esta variável de interesse é comumente chamada de variável latente. No caso de testes educacionais, o escore em um teste pode ser decomposto como a soma de uma constante ("escore verdadeiro") e uma variável aleatória, que representa a união de vários fatores aleatórios que venham a interferir no momento da aplicação do teste. Para exemplificar, pode-se supor que um professor deseja saber o "escore verdadeiro" de cada aluno (variável latente) em um determinado teste, mas, o que ele observa são os escores destes alunos no teste, os "escores observados". Os escores observados não são necessariamente iguais aos escores verdadeiros, mas eles representam uma maneira indireta de se observar os "escores verdadeiros". O professor gostaria de saber se este teste está conseguindo avaliar o "escore verdadeiro", e ainda saber se o teste é comparável a um outro teste elaborado anteriormente. Na abordagem da teoria clássica de testes, tais medidas de adequação de um teste são, em geral, obtidas por correlações entre os escores de um teste e os de outro teste padrão, ou, por correlações entre os escores de subtestes de um mesmo

teste.

O objetivo deste capítulo é apenas mostrar alguns aspectos da teoria clássica, baseados em trabalhos como os de Carmines e Zeller (1981), Lord e Novick (1968) e Vianna (1982). Neste capítulo são apresentados de forma organizada os fundamentos da teoria clássica de testes através de uma notação unificada. Alguns conceitos importantes como fidedignidade e validade de um teste, contidos na Secção 2.2, são ferramentas utilizadas para avaliar um teste na teoria clássica. A Secção 2.3 apresenta algumas formas mais conhecidas para encontrar estimativas da fidedignidade e da validade, respectivamente.

Uma observação é a de que a teoria aqui apresentada pode ser aplicada a testes que não sejam dicotômicos.

2.2 Fundamentos da Teoria Clássica

Inicialmente são definidos a decomposição do escore (nota) observado de uma pessoa em um teste, e alguns resultados oriundos desta decomposição.

Seja $\mathcal{P}$ uma população de pessoas.

Definição 1 *Para uma pessoa "a", pertencente à população $\mathcal{P}$, o escore correspondente ao desempenho num determinado teste, X_a, é o resultado da soma de uma constante (positiva e finita), τ_a, e uma variável aleatória, ϵ_a,*

$$X_a = \tau_a + \epsilon_a,$$

onde o índice "a" é usado para referenciar a pessoa específica da população $\mathcal{P}$.

Além disso, o valor esperado da variável aleatória X_a é definido como τ_a, denominado

escore verdadeiro:

$$E_a(X_a) = \tau_a.$$

A constante τ_a é um valor que pode ser interpretado como a média de escores observados em um número infinito de aplicações (hipotéticas) do mesmo teste à mesma pessoa.

Resultado 1 *Em consequência à definição 1, o valor esperado de* ϵ_a *é zero:*

$$E_a(\epsilon_a) = 0. \tag{2.1}$$

□

Definição 2 *Seja* X *a variável aleatória que representa o escore correspondente ao desempenho de uma pessoa, selecionada ao acaso de* $\mathcal{P}$*, num determinado teste. Desta forma,* $X = T + \epsilon$*, onde* T *é a variável escore verdadeiro, e* ϵ *é a variável escore erro.*

Resultado 2 *O valor esperado de* ϵ *é zero, e, consequentemente, o valor esperado de* X *é igual ao valor esperado de* T*, onde* X*,* T *e* ϵ *seguem a definição 2.*

Prova

$$E(\epsilon) = E[E(\epsilon/\textit{seleção de pessoa } a)] = E(E_a(\epsilon_a)) = 0, \textit{ por (2.1)} \tag{2.2}$$

$$E(X) = E(T + \epsilon) = E(T). \tag{2.3}$$

□

No modelo clássico algumas suposições com respeito às variáveis definidas são formuladas:

1. $0 < \sigma^2(X), \sigma^2(T), \sigma^2(\epsilon) < \infty$, onde $\sigma^2(X)$ é definido como a variância da variável X.

2. As variáveis aleatórias T e ϵ são não correlacionadas:

$$\rho(T, \epsilon) = 0, \tag{2.4}$$

onde, $\rho(X, Y)$ é a correlação entre X e Y.

A fidedignidade de um teste avalia o grau com que o escore observado representa o escore verdadeiro.

Definição 3 *A fidedignidade de um teste, $\rho^2(X,T)$, é definida como o quadrado da correlação entre o escore observado e o escore verdadeiro de uma pessoa selecionada ao acaso de $\mathcal{P}$.*

Resultado 3 *A fidedignidade de um teste, cujo escore é dado por $X = T + \epsilon$, de uma pessoa, selecionada ao acaso de $\mathcal{P}$, pode ser expressa como*

$$\rho^2(X,T) = \frac{\sigma^2(T)}{\sigma^2(X)}. \tag{2.5}$$

Prova:

Sendo $X = T + \epsilon$,

$$\begin{aligned}
\sigma(X,T) &= E[(T+\epsilon)T] - E(T+\epsilon)E(T) \\
&= \left[E(T^2) - E^2(T)\right] + E(\epsilon T) - E(\epsilon)E(T) \\
&= \sigma^2(T) + \sigma(\epsilon, T) \\
&= \sigma^2(T), \text{ por (2.4)},
\end{aligned}$$

$$\sigma(X,T) = \sigma^2(T) \Rightarrow \rho^2(X,T) = \frac{[\sigma^2(T)]^2}{\sigma^2(X)\sigma^2(T)} = \frac{\sigma^2(T)}{\sigma^2(X)}.$$

□

Ainda, por (2.4), $\sigma^2(X) = \sigma^2(T) + \sigma^2(\epsilon)$; assim, pode-se expressar a fidedignidade de um teste por:

$$\rho^2(X,T) = \frac{\sigma^2(T)}{\sigma^2(T) + \sigma^2(\epsilon)}, \tag{2.6}$$

ficando explícito que quanto menor for a variância de ϵ maior será a fidedignidade do teste.

Outras suposições adicionais às apresentadas anteriormente para o modelo clássico são apresentadas abaixo. Os índices, aqui utilizados se referem a diferentes testes.

1. Os escores erros associados a dois testes distintos são não correlacionados:

$$\rho(\epsilon_1, \epsilon_2) = 0. \tag{2.7}$$

2. O escore erro e o escore verdadeiro associados a dois testes distintos são não correlacionados:

$$\rho(\epsilon_1, T_2) = 0. \tag{2.8}$$

A fidedignidade de um teste, conforme expressa em (2.5) e (2.6), é uma função da variável aleatória latente T. Como o escore verdadeiro, T, não pode ser observado, torna-se necessário a definição de testes com características bem específicas, de modo que, na presença de tais tipos de testes, a fidedignidade seja uma função de variáveis observáveis. Esses testes são denominados testes paralelos e sua definição é dada abaixo.

Definição 4 *Sejam os escores de dois testes, aplicados a uma mesma pessoa, selecionada ao acaso da população $\mathcal{P}$, expressos por $X = T + \epsilon$ e $X' = T' + \epsilon'$, segundo a definição 2. Estes dois testes são denominados paralelos se:*

$$T = T' \quad e \tag{2.9}$$

$$\sigma^2(\epsilon) = \sigma^2(\epsilon'). \tag{2.10}$$

Dois testes paralelos são construídos com o objetivo de medir o mesmo escore verdadeiro e suas fidedignidades são iguais, pois as variâncias dos escores erros são iguais. No exemplo citado por Carmines e Zeller (1981, pág. 32) são apresentados os seguintes itens da escala de Rosenberg (1965) para avaliar a auto-estima:

- Eu acho que tenho boas qualidades.
- Eu me sinto, pelo menos, em mesmo plano que as outras pessoas.

Uma mesma pessoa deveria responder a esses dois itens da mesma forma. Estes dois itens são considerados paralelos, pelos autores. Um teste é paralelo a um segundo teste se este teste é composto por itens paralelos aos itens do segundo teste.

Resultado 4 *Se dois testes, aplicados a uma pessoa, selecionada ao acaso de $\mathcal{P}$, com escores $X = T + \epsilon$ e $X' = T' + \epsilon'$ são paralelos, então*

$$\begin{aligned} E(X) &= E(X') \quad e \\ \sigma^2(X) &= \sigma^2(X'). \end{aligned} \tag{2.11}$$

Prova:

$$\begin{aligned} X' = T' + \epsilon' \Rightarrow \quad E(X') &= E(T') + E(\epsilon') \\ &= E(T), \text{ por (2.9) e (2.2)} \\ &= E(X), \text{ por (2.3)}; \end{aligned}$$

$$X = T + \epsilon \Rightarrow \sigma^2(X) = \sigma^2(T) + \sigma^2(\epsilon), \text{ por (2.4)}$$
$$= \sigma^2(T') + \sigma^2(\epsilon'), \text{ por (2.9) e (2.10)}$$
$$= \sigma^2(X'), \text{ por (2.4)}.$$

□

Resultado 5 *A fidedignidade de um teste, aplicado a uma pessoa, selecionada ao acaso de* $\mathcal{P}$*, cujo escore é dado por* $X = T + \epsilon$*, é igual à correlação entre o escore deste teste e o escore* $X' = T' + \epsilon'$ *de um outro teste, paralelo ao primeiro,*

$$\rho^2(X, T) = \rho(X, X').$$

Prova:

Assumindo, sem perda de generalidade, que

$$E(X) = E(X') = E(T) = 0,$$

$$\begin{aligned} \rho(X, X') &= \frac{E(XX')}{\sigma(X)\sigma(X')} \\ &= \frac{E[(T+\epsilon)(T+\epsilon')]}{\sigma(X)\sigma(X')}, \text{ por (2.9)} \\ &= \frac{E(T^2) + E(T\epsilon') + E(T\epsilon) + E(\epsilon\epsilon')}{\sigma(X)\sigma(X')}; \end{aligned}$$

sendo $E(T\epsilon) = E(T\epsilon') = E(\epsilon\epsilon') = 0$, em consequência a (2.7) e (2.8), e usando o fato de que $\sigma^2(X) = \sigma^2(X')$, pela hipótese dos testes serem paralelos (Resultado 4),

$$\begin{aligned} \rho(X, X') &= \frac{\sigma^2(T)}{\sigma^2(X)} \\ &= \rho^2(X, T), \text{ pelo Resultado 3.} \end{aligned}$$

□

Definição 5 *Um teste, aplicado a uma pessoa, selecionada ao acaso de $\mathcal{P}$, composto por itens agrupados em n subtestes, onde o escore do i-ésimo subteste é expresso como $X_i = T_i + \epsilon_i$, $i = 1, 2, \ldots, n$, conforme a definição 2, pode ser decomposto em*

$$X = \sum_{i=1}^{n} X_i,$$
$$T = \sum_{i=1}^{n} T_i,$$
$$\epsilon = \sum_{i=1}^{n} \epsilon_i,$$

sendo $X = T + \epsilon$ o escore total do teste.

A fidedignidade de um teste é também definida através da decomposição do próprio teste em dois subtestes paralelos (se esta decomposição for possível).

Resultado 6 *Considere os escores de dois testes paralelos, X e X', aplicados a uma pessoa, selecionada ao acaso de $\mathcal{P}$. Seja o teste, cujo escore é X, decomposto em dois subtestes paralelos, X_1 e X_2. E seja o teste, cujo escore é X', decomposto em dois subtestes paralelos, cujos escores são X_1' e X_2', sendo os quatro subtestes, com escores X_1, X_2, X_1' e X_2', paralelos. Então, a fidedignidade do teste cujo escore é X pode ser expressa como,*

$$\frac{2\rho(X_1, X_2)}{1 + \rho_(X_1, X_2)}. \tag{2.12}$$

Prova:

Sendo X e X' escores de testes paralelos, tem-se, por definição (2.11):

$$\sigma^2(X) = \sigma^2(X'), \tag{2.13}$$

e, pelo Resultado 5, a fidedignidade do teste cujo escore é X, é dada por $\rho(X, X')$.

Como X_1, X_2, X_1', X_2' são escores de subtestes paralelos, tem-se, por definição:

$$\sigma^2(X_1) = \sigma^2(X_2) = \sigma^2(X_1') = \sigma^2(X_2'), \tag{2.14}$$

e pelo resultado 5, tem-se que:

$$\rho(X_1, X_2) = \rho(X_i, X_j'), \text{ para } i = 1,2 \text{ e } j = 1,2,$$

implicando em

$$\sigma(X_1, X_2) = \sigma(X_i, X_j'), \text{ para } i = 1,2 \text{ e } j = 1,2. \tag{2.15}$$

Assim,

$$\rho(X, X') = \frac{\sigma(X, X')}{\sigma^2(X)}, \text{ por (2.13)},$$

e, expressando X e X' por $X_1 + X_2$ e $X_1' + X_2'$, respectivamente,

$$\begin{aligned}
\rho(X, X') &= \frac{\sigma(X_1, X_1') + \sigma(X_1, X_2') + \sigma(X_2, X_1') + \sigma(X_2, X_2')}{\sigma^2(X_1) + \sigma^2(X_2) + 2\sigma(X_1, X_2)} \\
&= \frac{4\sigma(X_1, X_2)}{2\sigma^2(X_1) + 2\rho(X_1, X_2)\sigma^2(X_1)}, \text{ por (2.14) e (2.15)} \\
&= \frac{4\rho(X_1, X_2)\sigma^2(X_1)}{2\sigma^2(X_1) + 2\rho(X_1, X_2)\sigma^2(X_1)}, \text{ por (2.14)} \\
&= \frac{2\rho(X_1, X_2)}{1 + \rho(X_1, X_2)}
\end{aligned}$$

□

A quantidade (2.12) é denominada fórmula de Spearman-Brown para testes de comprimento duplo.

Uma cota inferior (que não é função de variáveis latentes) para a fidedignidade de um teste pode ser definida mesmo quando não existe paralelismo entre os subtestes de um mesmo teste.

Resultado 7 *Sejam $X_1, X_2, \ldots, X_n$ os escores de subtestes, expressos como $X_i = T_i + \epsilon_i$, onde $X = \sum_{i=1}^{n} X_i$ é o escore (nota) total do teste, aplicado a uma pessoa, selecionada ao acaso de $\mathcal{P}$. Então, uma cota inferior para a fidedignidade do teste é dada por*

$$\frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma^2(X_i)}{\sigma^2(X)}\right].$$

Prova:

Pela não negatividade da variância, para $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, n$:

$$\sigma^2(T_i) + \sigma^2(T_j) \geq 2\sigma(T_i, T_j).$$

Logo,

$$\sum_{i=1}^{n}\sum_{i=1_{i\neq j}}^{n}\left[\sigma^2(T_i) + \sigma^2(T_j)\right] \geq 2\sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}\sigma(T_i, T_j). \tag{2.16}$$

Como

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}[\sigma^2(T_i) + \sigma^2(T_j)] &= \sum_{i=1}^{n}\left[n\sigma^2(T_i) + \sum_{j=1}^{n}\sigma^2(T_j)\right] \\ &= n\sum_{i=1}^{n}\sigma^2(T_i) + n\sum_{j=1}^{n}\sigma^2(T_j) \\ &= 2n\sum_{i=1}^{n}\sigma^2(T_i), \qquad (2.17)\end{aligned}$$

e

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}[\sigma^2(T_i) + \sigma^2(T_j)] &= \sum_{i=1}^{n}\sum_{j=1_{i=j}}^{n}[\sigma^2(T_i) + \sigma^2(T_j)] \\ &+ \sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}[\sigma^2(T_i) + \sigma^2(T_j)] \\ &= 2\sum_{i=1}^{n}\sigma^2(T_i) + \sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}[\sigma^2(T_j) + \sigma^2(T_j)], \qquad (2.18)\end{aligned}$$

tem-se

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}[\sigma^2(T_i)+\sigma^2(T_j)] &= \sum_{i=1}^{n}\sum_{j=1}^{n}[\sigma^2(T_i)+\sigma^2(T_j)] - 2\sum_{i=1}^{n}\sigma^2(T_i), \text{ por (2.18)} \\ &= 2n\sum_{i=1}^{n}\sigma^2(T_i) - 2\sum_{i=1}^{n}\sigma^2(T_i), \text{ por (2.17)} \\ &= 2(n-1)\sum_{i=1}^{n}\sigma^2(T_i).\end{aligned}$$

Consequentemente

$$\sum_{i=1}^{n}\sigma^2(T_i) = \frac{1}{2(n-1)}\sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}[\sigma^2(T_i)+\sigma^2(T_j)],$$

e

$$\sum_{i=1}^{n}\sigma^2(T_i) \geq \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}\sigma(T_i,T_j), \text{ por (2.16)} \tag{2.19}$$

Ainda,

$$\sigma^2(T) = \sigma^2(\sum_{i=1}^{n}T_i) = \sum_{i=1}^{n}\sigma^2(T_i) + \sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}\sigma(T_i,T_j)$$

implica,

$$\sum_{i=1}^{n}\sigma^2(T_i) = \sigma^2(T) - \sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}\sigma(T_i,T_j). \tag{2.20}$$

Substituindo $\sum_{i=1}^{n}\sigma^2(T_i)$ conforme expresso em (2.20), na desigualdade em (2.19):

$$\sigma^2(T) \geq \sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}\frac{\sigma(T_i,T_j)}{n-1} + \sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}\sigma(T_i,T_j),$$

e assim,

$$\sigma^2(T) \geq \frac{n}{n-1}\sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}\sigma(T_i,T_j). \tag{2.21}$$

Para completar a prova,

$$\sigma^2(X) - \sum_{i=1}^{n}\sigma^2(X_i) = \sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}\sigma(X_i,X_j)$$

(2.22) $$= \sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n} \sigma(T_i, T_j),$$

(2.23) em consequência à (2.7) e (2.8).

Substituindo o lado esquerdo da equação (2.22) no lado direito da inequação (2.21), tem-se

$$\sigma^2(T) \geq \frac{n}{n-1}\left[\sigma^2(X) - \sum_{i=1}^{n}\sigma^2(X_i)\right],$$

e dividindo ambos os lados da inequação acima por $\sigma^2(X)$,

$$\frac{\sigma^2(T)}{\sigma^2(X)} \geq \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma^2(X_i)}{\sigma^2(X)}\right].$$

□

Denomina-se "coeficiente α" a cota inferior para a fidedignidade de um teste (Cronbach, 1951):

(2.24) $$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma^2(X_i)}{\sigma^2(X)}\right].$$

Kuder e Richardson (1937) abordam casos especiais para este coeficiente. Um deles refere-se aos escores X_i, variáveis aleatórias com espaço amostral $\{0, 1\}$, com probabilidades $1 - p_i$ e p_i, respectivamente, $i = 1, 2, \ldots, n$. Nestas condições o coeficiente α se reduz à "fórmula 20 de Kuder-Richardson", ou KR-20:

(2.25) $$\alpha_{(20)} = \frac{n}{n-1}(1 - \frac{\sum p_i(1-p_i)}{\sigma^2(X)}).$$

Uma simplificação de (2.25) pode ser feita, considerando $p_i = p$, para todo i, denominada "fórmula 21 de Kuder-Richardson", ou KR-21:

(2.26) $$\alpha_{(21)} = \frac{n}{n-1}[1 - \frac{np(1-p)}{\sigma^2(X)}].$$

Uma outra maneira de se avaliar um teste é através da validade. A validade de um teste começa no momento em que se pensa em elaborar o teste e subexiste durante todo o processo

de elaboração, aplicação, correção e interpretação dos resultados. Um teste não é válido de modo geral, mas sim em relação a determinado propósito. Instituições como a American Psychological Association (APA), American Educational Research Association (AERA) e o National Councial on Measurement (NCME) distinguem três tipos fundamentais de validade: a critério-relacionada (concorrente e preditiva), a de conteúdo e a de construto.

Do ponto de vista estatístico, são interessantes as validades critério-relacionada e a de construto, uma vez que a de conteúdo é um parecer contendo o julgamento de diferentes examinadores. A validade critério-relacionada pode ser dividida em dois tipos de validades: a validade concorrente e a validade preditiva. A validade concorrente é a correlação entre os escores de um teste e os escores de um critério, obtidos simultaneamente, ou melhor ao mesmo tempo, e que medem o mesmo desempenho. A validade preditiva é dada pela correlação entre os escores de um teste e outras medidas do desempenho (critério), obtidas independentemente do primeiro teste, em um tempo futuro. A validade de construto é obtida pela correlação entre escores obtidos de critérios compatíveis com as hipóteses que dizem respeito a conceitos que se deseja avaliar. Sem distinguir entre os tipos critério-relacionada e construto, uma definição de validade é dada a seguir.

Definição 6 *A validade de um teste, aplicado a uma pessoa, selecionada ao acaso de $\mathcal{P}$, cujo escore é dado por X, em relação a um segundo teste padrão, cujo escore é dado por Y, é definida como a correlação entre os escores X e Y.*

2.3 Estimação da Fidedignidade e Validade de um Teste

Como foi visto na secção anterior, a fidedignidade e validade são correlações. Neste caso, os estimadores destas correlações podem ser definidos pelas correlações amostrais entre dois conjuntos de escores observáveis. Existem várias abordagens para a definição de testes ou subtestes paralelos, estratégia necessária para a estimação da fidedignidade de um teste. Nessas várias abordagens é crucial o questionamento da suposição de paralelismo. Alguns desses procedimentos de estimação são: método do teste reteste, método das formas paralelas e métodos de análise interna, que serão apresentados a seguir.

Na abordagem do método do teste-reteste, o mesmo teste é aplicado à mesma pessoa após um certo período de tempo. Considerando as duas administrações do teste como paralelas, o coeficiente de correlação entre os escores, obtidos pelas duas aplicações, pode ser considerado um estimador da fidedignidade do teste. Algumas objeções são feitas a este método no que tange a suposição de paralelismo entre o teste e o reteste. Em algumas situações alguns fatores podem contribuir no escore erro do reteste, como por exemplo: problemas de saúde, fadiga, flutuações de memória e compreensão . Ou ainda, as respostas do reteste não são independentes da primeira aplicação. Então as respostas do reteste podem ser influenciadas por lembranças, por discussões entre estudantes e ainda pelo esforço individual ou coletivo em aprender o material no intervalo entre os testes. Se o intervalo entre os testes é longo, os escores erros podem ser confundidos com mudanças reais resultantes de aprendizado. A falta de interesse da parte dos estudantes também pode resultar em um segundo teste com um resultado mais pobre que o primeiro. Nos casos citados acima, o teste e o reteste não podem ser considerados paralelos.

O método das formas paralelas é um método bastante utilizado, o qual requer a aplicação

de dois testes diferentes, pressupostos paralelos. O coeficiente de correlação entre os escores destes dois testes é um estimador da fidedignidade. O maior problema deste método está na dificuldade em se construir formas paralelas de testes.

Os métodos de análise interna necessitam de apenas uma aplicação do teste e podem ser, basicamente, subdivididos em dois grupos: os que utilizam subtestes paralelos (método das metades) e os que não utilizam (coeficiente α). No método das metades o total de itens é dividido em duas partes, de forma a poder considerar os dois subtestes paralelos. Estes dois subtestes são usados como duas fontes independentes de escores para a estimação da fidedignidade, através da fórmula de Spearman-Brown (2.12). Deve-se tomar cuidado ao se construir os dois subtestes, para se poder garantir o paralelismo entre eles. Na ausência de paralelismo, pelo menos a cota inferior para a fidedignidade pode ser estimada, com base em uma única aplicação do teste. Um estimador desta cota inferior é obtido através do coeficiente α (2.24), usando estimadores das variâncias dos escores do teste e dos subtestes que compõem o teste.

É importante observar que muitas vezes não temos uma amostra aleatória de indivíduos, e sim um conjunto de indivíduos que deverão ser avaliados pelo teste. Assim, a fidedignidade estimada pode não refletir somente características do teste, mas também características do grupo de indivíduos em que o teste foi aplicado. Então, quanto mais representativo da população for o grupo de indivíduos ao qual se aplica o teste, menor a dependência do coeficiente de fidedignidade com o mesmo. Se este coeficiente for obtido de um grupo de indivíduos que não representa a população de indivíduos este coeficiente irá refletir o comportamento do teste para este grupo de indivíduos.

2.4 Conclusão

A teoria clássica de testes busca avaliar um teste como um todo, usando ferramentas baseadas em coeficientes de correlação. Uma dessas ferramentas é a fidedignidade, que pode ser calculada através de testes ou subtestes paralelos ou ao menos, na ausência de paralelismo, uma cota inferior para a fidedignidade pode ser encontrada.

Esta metodologia pode ser aplicada a testes compostos de itens com respostas abertas, assim como a testes compostos por itens dicotômicos.

Capítulo 3

TEORIA DE RESPOSTA DE ITEM

3.1 Introdução

Neste capítulo é descrita a função de resposta de item que define a probabilidade de acerto a um item de um teste por um indivíduo dotado de certa habilidade. Uma suposição razoável é que esta função é não decrescente em relação à habilidade do indivíduo. Na secção (3.2) são apresentados os modelos mais comumente utilizados para descrever esta função. A função de resposta de item compreende parâmetros referentes ao item e parâmetro referente ao indivíduo. Por exemplo, o modelo denominado logístico de dois parâmetros (que será definido na próxima secção) é composto por parâmetro referente à dificuldade do item e parâmetro referente à inclinação da curva no ponto de inflexão, além de parâmetro representante da habilidade de cada indivíduo. Na secção (3.3) são apresentados alguns métodos de estimação mais comumente utilizados para a estimação dos parâmetros da função de resposta de item. Ao estimar estes parâmetros se está fazendo inferência sobre a natureza do indivíduo, e também sobre o tipo de questão (item) em um determinado teste, como por exemplo, quanto à dificuldade do item. Na secção (3.4) é apresentado intervalo de confiança assintótico para o parâmetro referente ao indivíduo. E finalmente na secção (3.5) é apresentada a sugestão de

Birnbaum (1968) para a construção de um teste composto de itens, cujos parâmetros sejam conhecidos previamente.

A teoria de resposta de item (Item Response Theory ou IRT) vem sendo estudada por vários autores, por exemplo Lord e Novick (1968), Lord (1980) e Stocking e Pearlman (1989). Tais autores enfocaram o caso de testes com itens binários, ou seja, testes cujas respostas aos itens são classificadas como certas ou erradas. Os testes de múltipla escolha são exemplos de testes com itens binários. Neste trabalho é apenas abordado o caso de testes com itens binários. O objetivo deste capítulo é apresentar a teoria de resposta de item.

3.2 Função de Resposta de Item

A função de resposta de item tem como objetivo descrever matematicamente a probabilidade de acerto a um item, envolvendo nesta descrição parâmetros inerentes ao item e parâmetro inerente ao examinando. No que se refere ao examinando, assume-se que acertar a resposta de um item depende somente de uma única habilidade específica, denotada por θ; Bock e Aitkin (1981) abordaram modelos onde mais de um tipo de habilidade foram consideradas. Para um determinado item a função de resposta é definida como a probabilidade, $P(\theta)$, de resposta correta. É razoável assumir que $P(\theta)$ é uma função não decrescente em θ, ou seja, quanto maior a habilidade do indivíduo maior a probabilidade de acertar a um item.

O modelo mais simplificado para descrever esta probabilidade é dado por

$$P(\theta) = \frac{1}{\exp\{-(\theta - b)\}},$$

onde b é chamado parâmetro de dificuldade do item. Este modelo é denominado modelo de Rash (Wright, 1977).

Outro modelo comumente adotado é dado por

$$P(\theta) = c + \frac{1-c}{1+\exp\{-1,7a(\theta-b)\}}, \tag{3.1}$$

onde a ,b e c são parâmetros que caracterizam o item. Este modelo é denominado modelo logístico de três parâmetros. Pode-se observar que o modelo de Rash nada mais é que o modelo logístico com apenas um parâmetro. O parâmetro c, número real pertencente ao intervalo $[0,1)$, representa a probabilidade de uma pessoa com habilidade muito baixa responder ao item corretamente; é o chamado parâmetro de adivinhação. Particularmente, c é igual a zero, se um item não pode ser respondido corretamente por adivinhação. A curva logística (3.1) tem um único ponto de inflexão que ocorre em $\theta = b$ e a tangente à curva, no ponto $\theta = b$, é proporcional ao valor de a. O parâmetro b, número real, parâmetro de locação que determina a posição na escala da habilidade θ onde ocorre o ponto de inflexão, é chamado de parâmetro de dificuldade. O parâmetro a, número real não negativo, é proporcional à inclinação da curva no ponto de inflexão, b, representa o poder de discriminção do item. A Figura (3.1) apresenta uma função de resposta de item de um modelo logístico de três parâmetros.

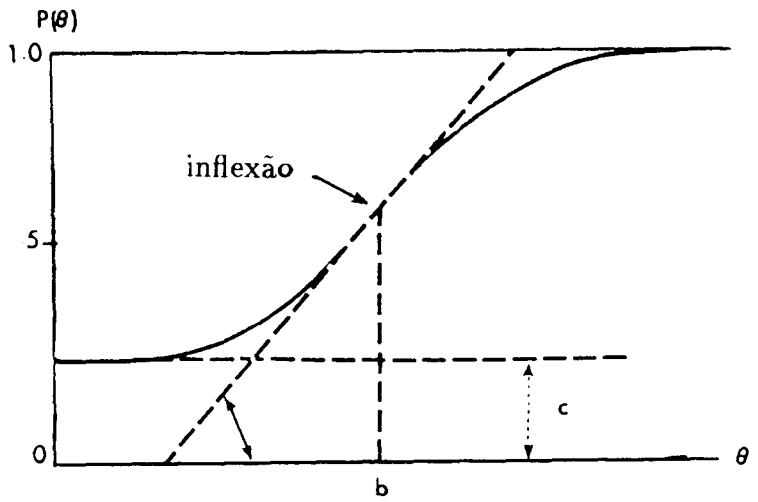
Uma forma alternativa para a curva de resposta de item é

$$P(\theta) = c + (1-c)\int_{-\infty}^{a(\theta-b)} \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-t^2}{2}\right\} dt, \tag{3.2}$$

onde c representa a probabilidade de uma pessoa com habilidade muito baixa responder corretamente ao item, b é o nível de habilidade no ponto de inflexão da curva, e a é proporcional à inclinação da curva no ponto de inflexão, b. Este modelo é denominado normal de três parâmetros.

Existem na literatura outros modelos derivados dos modelos logístico e normal de três parâmetros. Por exemplo, o logístico de dois parâmetros dado por Birnbaum (1968), ou seja,

Figura 3.1: Função de resposta de item - modelo logístico



o modelo (3.1) com $c = 0$, onde supõem-se que não existe adivinhação. Existem também modelos não paramétricos como o citado por Sijtsma e Molenar (1987), também conhecido como modelo de Mokken, onde nenhuma forma matemática de função de resposta de item é especificada.

Lord (1980, pág. 17) analisou resultados da aplicação de teste composto por 150 itens, onde cada item apresentava cinco escolhas, apenas uma escolha correspondendo à resposta correta ao item. Para cada item foi construído um gráfico bidimensional, onde para cada total de acertos no teste (eixo horizontal) foi indicada a proporção de acertos do item em questão, entre as pessoas que obtiveram este número total de acertos (eixo vertical). Um total de 103.275 examinandos foi submetido ao teste. Lord observa que dentre os 150 gráficos somente seis não revelavam claramente uma função não decrescente da proporção de acertos ao item em relação ao total de acertos no teste, e nestes casos as distorções encontradas eram pequenas. Este estudo sugere a adequacidade da suposição de que a probabilidade de acerto

a um item é uma função não decrescente em relação à habilidade do indivíduo.

3.3 Estimação da Habilidade e dos Parâmetros de Itens

Nesta secção, considera-se um teste composto por n itens binários, aplicado a m indivíduos. Seja U_{ij} a variável aleatória indicadora de acerto ao item i pelo j-ésimo indivíduo, dotado de habilidade θ_j, onde $j = 1, 2, \ldots, m$ e $i = 1, 2, \ldots, n$. Usando índices para referência ao item i e indivíduo j, as probabilidades definidas em (3.1) e (3.2) passam a ser denotadas por $P_i(\theta_j)$. No modelo logístico tem-se

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp\{-1,7a_i(\theta_j - b_i)\}}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m \tag{3.3}$$

Portanto, a função de probabilidade de U_{ij} é dada por

$$P(U_{ij} = u_{ij} \mid \theta_j, a_i, b_i, c_i) = Pi(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m \tag{3.4}$$

para $u_{ij} = 0$ ou 1, onde $Q_i(\theta_j) = 1 - P_i(\theta_j)$, ou seja, o modelo de probabilidade de U_{ij} é Bernoulli, com parâmetro $P_i(\theta_j)$.

Birnbaum (1968, pág 389) e Lord (1980, pág. 19) argumentam que uma suposição adicional bastante razoável é a independência entre as variáveis associadas às respostas de um indivíduo a diferentes itens, condicionada a habilidade θ. Sob a suposição de independência, a função de probabilidade conjunta de $U_{1j}, U_{2j}, \ldots,$ e U_{nj} (indicadores de acertos aos vários itens, pelo j-ésimo indivíduo), no ponto $u_{1j}, u_{2j}, \ldots,$ e u_{nj}, é dada por

$$P(u_{1j}, u_{2j}, \ldots, u_{nj} \mid \theta_j, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}}, \tag{3.5}$$

$$j = 1, 2, \ldots, m$$

para $u_{ij} = 0$ ou 1, e considerando que as variáveis associadas a indivíduos diferentes são mutuamente independentes,

$$P[(\mathbf{U_1}, \mathbf{U_2}, \ldots, \mathbf{U_m}) = (\mathbf{u_1}, \mathbf{u_2}, \ldots \mathbf{u_m}) \mid \theta, \mathbf{a}, \mathbf{b}, \mathbf{c}] = \prod_{i=1}^{n} \prod_{j=1}^{m} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}}, \tag{3.6}$$

para $u_{ij} = 0$ ou 1, onde, $\mathbf{a} = (a_1, a_2, \ldots, a_n)'$; $\mathbf{b} = (b_1, b_2, \ldots, b_n)'$; $\mathbf{c} = (c_1, c_2, \ldots, c_n)'$, $\theta = (\theta_1, \theta_2, \ldots, \theta_m)'$, $\mathbf{U_j} = (U_{1j}, U_{2j}, \ldots, U_{nj})$ e $\mathbf{u_j} = (u_{1j}, u_{2j}, \ldots, u_{nj})$.

Os procedimentos mais conhecidos para a estimação dos parâmetros de uma função de resposta de item são descritos a seguir.

3.3.1 Métodos da Máxima Verossimilhança

Os métodos da máxima verossimilhança são os mais conhecidos e foram os primeiros a serem utilizados para estimar os parâmetros de funções de resposta de item. Nesta secção são apresentados dois procedimentos que utilizam a função de verossimilhança das respostas dos indivíduos aos itens, a fim de se obter os estimadores dos parâmetros da função de resposta de item.

Sendo a função de probabilidade das respostas das variáveis indicadoras de acertos dos m indivíduos aos n itens dada por (3.6), a função de verossimilhança pode ser expressa como

$$L(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c} \mid \mathbf{u_1}, \mathbf{u_2}, \ldots \mathbf{u_m}) = \prod_{j=1}^{m} \prod_{i=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}}, \tag{3.7}$$

onde, por simplicidade, P_{ij} e Q_{ij} são usados ao invés de $P_i(\theta_j)$ e $Q_i(\theta_j)$. O logarítmo natural desta função é dado por

$$l = \sum_{j=1}^{m} \sum_{i=1}^{n} [u_{ij} \ln P_{ij} + (1 - u_{ij}) \ln Q_{ij}]. \tag{3.8}$$

Se γ representar θ_j, ou a_i, ou b_i, ou c_i, a derivada do logarítmo da função de verossimilhança, com respeito a γ, é dada por:

$$\begin{aligned}
\frac{\partial l}{\partial \gamma} &= \sum_{j=1}^{m}\sum_{i=1}^{n}[u_{ij}\frac{P'_{ij}}{P_{ij}} - (1-u_{ij})\frac{P'_{ij}}{Q_{ij}}] \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n}[u_{ij}\frac{P'_{ij}}{P_{ij}} - \frac{P'_{ij}}{Q_{ij}} + u_{ij}\frac{P'_{ij}}{Q_{ij}}] \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n}[\frac{P'_{ij}}{P_{ij}Q_{ij}}(u_{ij}Q_{ij} - P_{ij} + u_{ij}P_{ij})] \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n}[\frac{P'_{ij}}{P_{ij}Q_{ij}}(u_{ij} - P_{ij})], \qquad (3.9)
\end{aligned}$$

onde, $P'_{ij} = \frac{\partial P_{ij}}{\partial \gamma}$. Uma forma explícita para a expressão P'_{ij} pode ser dada quando se especifica a forma de P_{ij}, equação (3.1) ou (3.2).

Lord (1980) cita que os estimadores de máxima verossimilhança para os parâmetros da função de resposta de item são as soluções do sistema abaixo:

$$\sum_{i=1}^{n}\frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}}\frac{\partial P_{ij}}{\partial \theta_j} = 0 \quad (j = 1, 2, \ldots, m) \qquad (3.10)$$

$$\sum_{j=1}^{m}\frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}}\frac{\partial P_{ij}}{\partial a_i} = 0 \quad (i = 1, 2, \ldots, n) \qquad (3.11)$$

$$\sum_{j=1}^{m}\frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}}\frac{\partial P_{ij}}{\partial b_i} = 0 \quad (i = 1, 2, \ldots, n) \qquad (3.12)$$

$$\sum_{j=1}^{m}\frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}}\frac{\partial P_{ij}}{\partial c_i} = 0 \quad (i = 1, 2, \ldots, n) \qquad (3.13)$$

As equações dadas acima são equações não lineares em θ, **a**, **b** e **c** (Lord, 1980). Formas para obter soluções deste sistema são procedimentos iterativos. Neste sistema,com $m + 3n$ equações e $m + 3n$ incógnitas, as estimativas dos parâmetros de itens dependem das estimativas das habilidades e as estimativas das habilidades dependem das estimativas dos parâmetros de item. Lord (1980, pág 181) argumenta que, quando estimativas dos

parâmetros de itens e parâmetros de indivíduos são estimados simultaneamente, tais estimativas não convergem para os parâmetros verdadeiros se o número de indivíduos é grande.

Procedimento: "Máxima Verossimilhança Conjunta"

Birnbaum (1968) propõe um procedimento iterativo para a estimação de parâmetros da função de resposta de item, que pode ser utilizado tanto para o modelo logístico (3.2) quanto para o modelo normal (3.1).

No primeiro estágio são utilizados valores iniciais para as habilidades dos indivíduos, com a finalidade de estimar os parâmetros de itens, encontrando a solução de (3.11), (3.12) e (3.13). Estes valores iniciais são definidos pelo autor como a inversa da distribuição acumulada da normal padrão da proporção de acertos do indivíduo no teste. Em um segundo estágio os parâmetros de itens são supostos conhecidos, utilizando as estimativas obtidas no primeiro estágio, com a finalidade de obter estimativas das habilidades de cada indivíduo, através de (3.10). Estes dois estágios são repetidos até a convergência por algum critério, não definido pelo autor. O programa LOGIST, desenvolvido por Wingerbky e Lord (1973), utililiza um procedimento similar ao descrito por Birnbaum para o modelo logístico de três parâmetros.

Procedimento: "Máxima Verossimilhança Marginal"

Bock e Lieberman (1970) descrevem uma forma alternativa para estimar os parâmetros referentes a itens do modelo normal de dois parâmetros, ou seja, (3.1) com $c = 0$. Definem dois estágios, onde primeiramente estimam-se os parâmetros de itens, e depois são estimadas as

habilidades dos indivíduos. No primeiro estágio utilizaram a função de distribuição marginal de $(U_{1j}, U_{2j}, \ldots, U_{nj})$, considerando a habilidade com uma densidade de probabilidade $g(\theta)$. Como a função de distribuição conjunta de $(U_{1j}, U_{2j}, \ldots, U_{nj}, \theta)$ é dada por (3.5), a distribuição marginal de $(U_{1j}, U_{2j}, \ldots, U_{nj})$ seria dada por

$$P\left[(U_{1j}, U_{2j}, \ldots, U_{nj}) = (u_{1j}, u_{2j}, \ldots, u_{nj}) \mid \mathbf{a}, \mathbf{b}\right] =$$
$$\int_{-\infty}^{+\infty} P[(U_{1j}, U_{2j}, \ldots, U_{nj}) = (u_{1j}, u_{2j}, \ldots, u_{nj}) \mid \theta, \mathbf{a}, \mathbf{b}] \;\; g(\theta) \;\; d\theta.$$
$$j = 1, 2, \ldots, m$$

Neste caso, a função de probabilidade conjunta de $(\mathbf{U_1}, \mathbf{U_2}, \ldots, \mathbf{U_m})$ seria dada por

$$P[(\mathbf{U_1}, \mathbf{U_2}, \ldots, \mathbf{U_m}) = (\mathbf{u_1}, \mathbf{u_2}, \ldots \mathbf{u_m}) \mid \theta, \mathbf{a}, \mathbf{b}] =$$
$$\prod_{j=1}^{m} P[(U_{1j}, U_{2j}, \ldots, U_{nj}) = (u_{1j}, u_{2j}, \ldots, u_{nj}) \mid \mathbf{a}, \mathbf{b}],$$

observando-se que esta função não depende de θ. A

função de verossimilhança correspondente a este modelo é

$$L(\mathbf{a}, \mathbf{b} \mid \mathbf{u_1}, \mathbf{u_2}, \ldots \mathbf{u_m}) =$$
$$\prod_{j=1}^{m} P[(U_{1j}, U_{2j}, \ldots, U_{nj}) = (u_{1j}, u_{2j}, \ldots, u_{nj}) \mid \mathbf{a}, \mathbf{b}]. \tag{3.14}$$

Os estimadores de máxima verossimilhança dos parâmetros de itens são as soluções do sistema de $2n$ equações, construído por derivações do logarítmo da equação (3.14), com respeito aos parâmetros de itens. Essas equações são não lineares em relação aos parâmetros de itens e, desta forma, é necessário um procedimento iterativo para encontrar as soluções deste sistema.

No segundo estágio as estimativas dos parâmetros de itens obtidas no primeiro estágio são utilizadas para obter as habilidades, $\theta_1, \theta_2, \ldots, \theta_m$ dos indivíduos, através de (3.10).

Bock e Aitkin (1981) apontam que, computacionalmente, o método de Bock e Lieberman requer no processo iterativo, para n itens, a geração e inversão (4 ou 5 vezes) de uma matriz $2n \times 2n$, onde cada elemento desta matriz corresponde a uma soma de 2^n termos. Logo, concluem que o método não é prático para $n > 12$. Outro problema é a dificuldade em especificar uma distribuição a priori para habilidade ($g(\theta)$). Bock e Aitkin reformularam este procedimento definindo $g(\theta)$ através da distribuição empírica do número de acertos de cada indivíduo, livrando o método da suposição arbitrária sobre a distribuição da habilidade da população amostrada, além disso, afirmam que o método tem solução viável para qualquer número de itens. Tal reformulação deu origem ao programa BILOG.

Uma das generalizações mais importantes do método de máxima verossimilhança marginal é a adaptação à suposição de multidimensionalidade para a habilidade do indivíduo. No entanto, esta suposição acarretaria a estimação de um número muito maior de parâmetros, uma vez que o número de parâmetros referentes a indivíduos seria agora a dimensão da habilidade, multiplicada pelo número de indivíduos.

Yen (1987) compara as estimativas obtidas pelos programas LOGIST (máxima verossimilhança) e BILOG (máxima verossimilhança marginal). Em simulações de testes com diferentes número de itens aplicados a diferentes números de indivíduos, Yen gerou habilidades através de amostras pseudo-aleatórias de distribuições normais e de distribuições não normais. As habilidades geradas por distribuições não normais eram uma mistura de duas normais, com o objetivo de conseguir uma certa assimetria. Yen conclui através das situações simuladas que o programa LOGIST é usualmente mais rápido que o BILOG (pelo tempo em CPU). As estimativas das habilidades obtidas pelo BILOG eram quase sempre mais precisas que as obtidas pelo LOGIST. As estimativas dos parâmetros de itens obtidas pelos dois programas eram na maioria dos casos igualmente precisas.

3.3.2 Procedimento Bayesiano

Swaminathan e Gifford (1986) desenvolveram um procedimento Bayesiano para a estimação da função de resposta de item para o modelo logístico de três parâmetros, (3.3).

A distribuição conjunta a posteriori $f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c} \mid \mathbf{u})$ pelo teorema de Bayes é dada por

$$f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c} \mid \mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_m}) \propto P(\mathbf{U_1}, \mathbf{U_2}, \ldots, \mathbf{U_m} = \mathbf{u_1}, \mathbf{u_2}, \ldots \mathbf{u_m} \mid \theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) \times f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}), \tag{3.15}$$

onde $P(\mathbf{U_1}, \mathbf{U_2}, \ldots, \mathbf{U_m} = \mathbf{u_1}, \mathbf{u_2}, \ldots \mathbf{u_m} \mid \theta, \mathbf{a}, \mathbf{b}, \mathbf{c})$ é dado por (3.6), e $f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c})$ é a distribuição conjunta a priori de θ, **a**, **b**, **c**.

Assumindo que θ, **a**, **b**, **c** são independentes,

$$f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) = f_1(\theta) f_2(\mathbf{a}) f_3(\mathbf{b}) f_4(\mathbf{c}).$$

Swaminathan e Gifford sugerem que θ e b sejam normalmente distribuídos, a tenha distribuição Qui-Quadrado e c tenha distribuição Beta. Os estimadores modais (Lindley e Smith, 1972) são os pontos que maximizam (3.15). No processo de definição destes estimadores forma-se um sistema de $m + 3n$ equações, funções dos parâmetros a serem estimados. O sistema obtido é não linear em θ, **a**, **b**, **c**, os autores optaram por um procedimento iterativo como o de Newton-Rapson.

Swaminathan e Gifford (1986) fizeram um estudo de simulações em que compararam as estimativas Bayesianas com estimativas de máxima verossimilhança obtidas pelo programa LOGIST (Wingerbky e Lord, 1973). Foram simuladas aplicações de testes com números diferentes de itens, aplicados a diferentes números de indivíduos. As habilidades dos indivíduos (θ) e os parâmetros b foram gerados através de amostras pseudo-aleatórias de distribuições

normais. Os parâmetros c e a foram gerados através de amostras pseudo-aleatórias de distribuições uniformes. Para estas simulações, as distribuições a priori para θ e b foram consideradas uniformes. Os autores observaram que para tais simulações as estimativas Bayesianas tiveram um melhor comportamento.

3.3.3 Estimação pelo Método do Mínimo Logito Qui-Quadrado

Baker, em 1987, verificando o procedimento desenvolvido por Berkson (1944) para a estimação de parâmetros de uma curva logística para ensaios biológicos com a utilização do método do mínimo logito qui-quadrado, observou que este método poderia ser aplicado no contexto de teoria de resposta de item. Em resumo, Baker propõe a redução nos parâmetros referentes a indivíduos numa primeira fase. Nesta primeira fase se objetiva estimar os parâmetros referentes a itens, através do método do mínimo logito qui-quadrado. Assim, seria necessário uma segunda fase para estimação das habilidades, onde Baker propõe a utilização do método de máxima verossimilhança, supondo os parâmetros de itens conhecidos, utilizando as estimativas destes parâmetros de itens obtidas na primeira fase.

Uma vantagem do método do mínimo logito qui-quadrado é a existência de uma forma fechada, em cada iteração, para os estimadores dos parâmetros de itens.

A combinação do método do mínimo logito qui-quadrado com o método de máxima verossimilhança, em duas fases distintas, idealizada por Baker, será referenciada neste trabalho como "procedimento de Baker". O procedimento de Baker será descrito com um maior detalhamento no próximo capítulo.

Baker comparou estimativas obtidas pelo seu procedimento com estimativas obtidas através do programa LOGIST (máxima verossimillhança) através de simulações. Para tais

simulações Baker variou o número de indivíduos e o número de itens e escolheu os parâmetros referentes a indivíduos gerando amostras pseudo-aleatórias de distribuição normal. Gerou o parâmetro a de distribuições uniformes e b de normais, construindo testes com itens classificados como fáceis, testes com itens classificados como difíceis e testes com itens classificados como medianos. Observou que seu procedimento requeria cerca de um terço do tempo computacional do programa LOGIST (comparando os tempos em CPU). Comparou também os coeficientes de correlação e os erros quadráticos médio entre as estimativas e parâmetros correspondentes, verificando que os dois procedimentos produziram resultados similares. Baker observou também que o parâmetro de discriminação do item, a, não foi na maioria dos casos bem estimado (verificando que os coeficientes de correlação entre parâmetro e estimativa eram baixos), tanto nas estimativas obtidas pelo programa LOGIST quanto as obtidas pelo procedimento de Baker.

3.4 Intervalo de Confiança Assintótico para a Habilidade

Supondo os parâmetros dos itens conhecidos, a função de probabilidade de $U_{11}, U_{21}, \ldots, U_{n1}$ (variáveis aleatórias correspondentes às respostas aos itens de um teste pelo primeiro indivíduo) é apenas função do parâmetro θ_1, sob essa condição de conhecimento dos valores dos parâmetros de itens. No Apêndice A é definido um estimador de máxima verossimilhança, $\hat{\theta}_1$, para o parâmetro θ_1. Neste mesmo apêndice é mostrado que a informação de Fisher de $U_{11}, U_{21}, \ldots, U_{n1}$ para θ_1 é dada por

$$I(\theta_1) = \sum_{i=1}^{n} \frac{P_{i1}'^2}{P_{i1}Q_{i1}}. \tag{3.16}$$

Pela desigualdade de Cramer-Rao, dado um estimador não-viciado $\hat{\theta}_1$ de θ_1, tem-se sob condições de regularidade que

$$Var(\hat{\theta}_1) \geq \frac{1}{I(\theta_1)}.$$

Sob condições de regularidade, Cramer (1946,pág. 500) mostrou que, se $\hat{\theta}_1$ for estimador de máxima verossimilhança de θ_1, $\hat{\theta}_1$ tem assintoticamente distribuição Normal com média θ_1 e variância $1/I(\theta_1)$.

Como as condições de regularidade são satisfeitas pelos modelos (3.1) ou (3.2) (Birnbaum, 1968 pág 457), Birnbaum propõe um intervalo de confiança assintótico de nível $(1-\alpha)100\%$ de confiança, para o parâmetro θ_1:

$$\left[\hat{\theta}_1 - \frac{z(\alpha/2)}{\sqrt{I(\hat{\theta}_1)}}, \hat{\theta}_1 + \frac{z(\alpha/2)}{\sqrt{I(\hat{\theta}_1)}}\right],$$

onde, $z(\alpha/2)$ é o valor da função inversa da normal padrão no ponto $(1-\alpha/2)$. Como o intervalo de confiança é assintótico, é razoável quando o número de itens é grande.

3.5 Sugestão de Birnbaum para a Construção de um Teste

Birnbaum (1968) sugeriu um procedimento para a construção de um teste. Este procedimento, no entanto, exige que os itens sejam "previamente calibrados", ou seja, que já sejam conhecidos os parâmetros de itens. A informação de Fisher de $U_{11}, U_{21}, \ldots, U_{n1}$ para θ_1 é dada por (3.16). Como a informação de Fisher de $U_{11}, U_{21}, \ldots, U_{n1}$ para θ_1 é composta pela soma das informações de cada U_{i1} de θ_1, $i = 1, 2, \ldots, n$, tem-se

$$I(\theta_1, U_{i1}) = \frac{P'^2_{i1}}{P_{i1}Q_{i1}}.$$

A função de informação de Fisher definida em (3.16), para um θ genérico é denominada função de informação do teste,

$$I(\theta) = \sum_{i=1}^{n} \frac{P_{ij}'^2}{P_{ij}Q_{ij}}.$$

A contribuição de cada item, denominada função de informação de item, é dada por

$$I(\theta, U_{ij}) = \frac{P_{ij}'^2}{P_{ij}Q_{ij}}.$$

O procedimento de Birnbaum, ilustrado na Figura (3.2), é o seguinte:

1. Decide-se pela precisão desejada para a estimação de cada nível de habilidade. Isto é, qual a forma desejada da função de informação do teste. Não esquecendo que esta função de informação é inversamente proporcional à variância assintótica do estimador da habilidade de um indivíduo. Esta curva é chamada de "curva de informação alvo".

2. Selecionam-se itens "previamente calibrados" e através da função de informação do teste, composta pela soma das funções de informação dos itens "calibrados", e a função de informação do teste é comparada com a curva de informação alvo. O objetivo é compor um teste cuja função de informação se aproxime da curva de informação alvo.

3. Acumulativamente, adicionam-se curvas de informação de itens, obtendo-se sempre a curva final do teste composta pelos itens já selecionados.

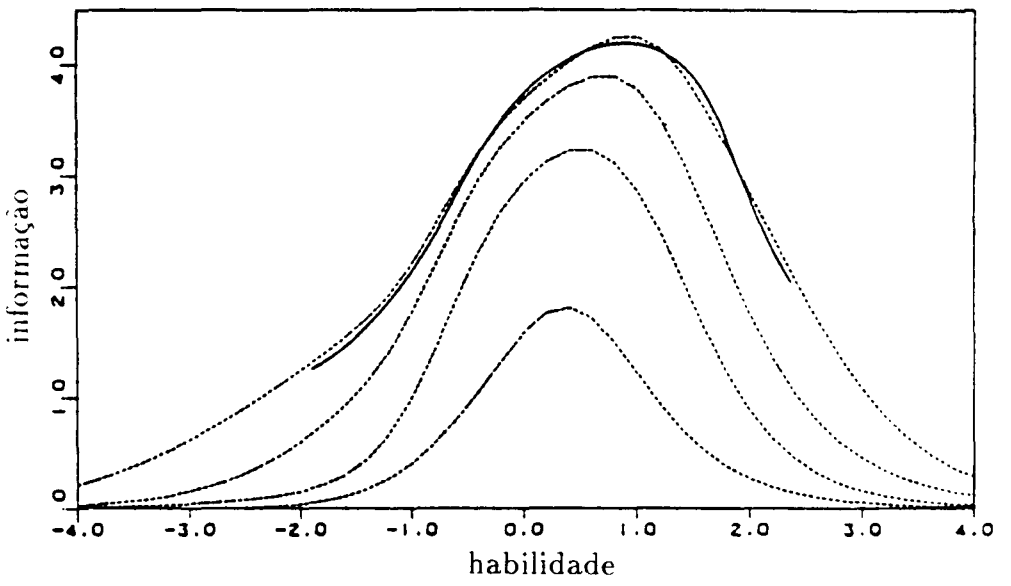
4. Continua-se até que a curva de informação do teste se aproxime satisfatoriamente da curva de informação alvo.

3.6 Conclusão

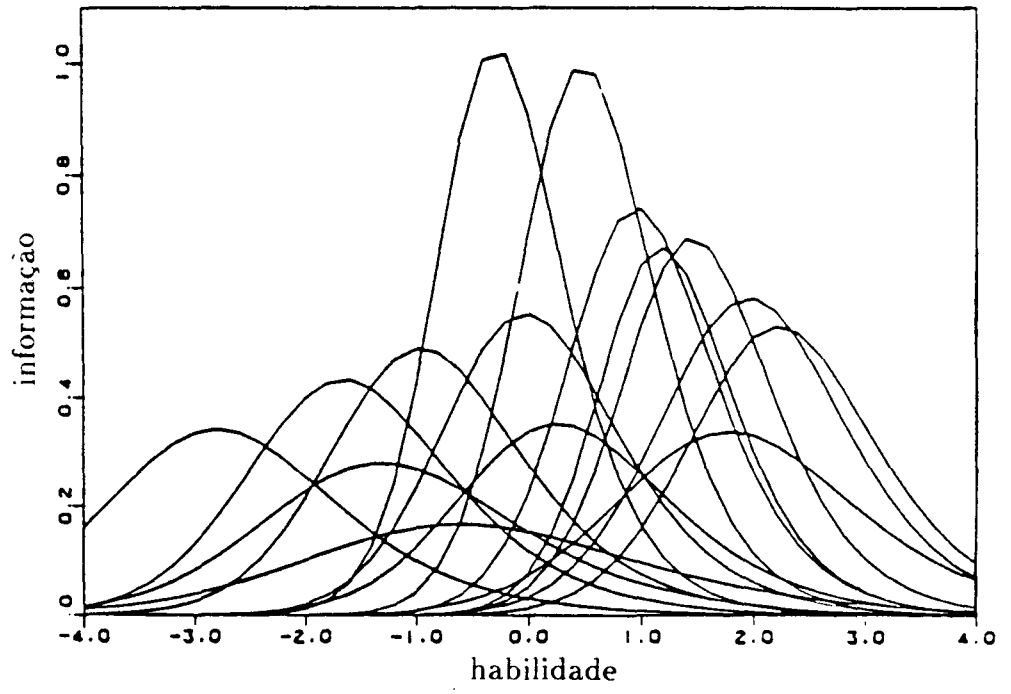
O objetivo primordial da teoria de resposta de item é fornecer elementos para inferência sobre um teste composto por itens, além de inferências sobre os indivíduos aos quais o teste é

Figura 3.2: (a) curva de informação alvo (sólida) e as curvas de informação para testes com: 3, 7, 11 e 15 itens respectivamente; (b) curvas de informação dos itens utilizados para aproximar a curva de informação alvo

(a)



(b)



aplicado, tendo como base amostra aleatória de respostas associadas a uma única aplicação do teste ao grupo de indivíduos. Uma dificuldade é o grande número de parâmetros que devem ser estimados. Por exemplo, para o ajuste do modelo logístico de três parâmetros, em um teste com n itens aplicado a m indivíduos, devem ser estimados $3n + m$ parâmetros. Outro fato importante é que a teoria de resposta de item possibilita a construção de um teste a partir de itens "previamente calibrados".

É importante observar que ao contrário da teoria clássica que avalia o teste como um todo, a teoria de resposta de item observa o teste segundo parâmetros de itens e parâmetros de indivíduos. Assim, a teoria de resposta de item enfatiza as características dos itens que compõem o teste, ao invés da simples análise das notas resultantes do mesmo.

Capítulo 4

ESTIMAÇÃO DA FUNÇÃO DE RESPOSTA DE ITEM

4.1 Introdução

A teoria de resposta de item determina para cada item, função de resposta de item que descrevem a probabilidade de acerto a um determinado item de um teste por um indivíduo dotado de certa habilidade (θ). Estas funções descrevem parâmetros relativos ao item e parâmetro relativo ao indivíduo. No Capítulo 3 alguns modelos para estas funções foram descritos e apresentados vários procedimentos de estimação. Este capítulo se restringe ao modelo logístico de dois parâmetros sendo escolhido um procedimento de estimação formulado por Baker (1937) para ser usado nas simulações. Portanto a probabilidade de um indivíduo com habilidade θ acertar um item é definida por

$$P(\theta) = \frac{1}{1 + \exp[-1,7a(\theta - b)]}. \tag{4.1}$$

Na secção 4.2 é descrito com detalhes o procedimento formulado por Baker. Este procedimento, ao invés de abordar o problema complexo de estimar simultaneamente todos os

parâmetros relativos aos itens e relativos aos indivíduos, desenvolve dois estágios de estimação. Na secção (4.3.1) é descrito o planejamento das simulações. Estas simulações tiveram como objetivo analisar este procedimento, em situações diversas como: habilidades de indivíduos geradas de amostras pseudo-aleatórias de distribuições normais e não normais. Foram simuladas situações onde o teste era composto por itens de dificuldades variadas como: com somente itens fáceis, ou com somente itens difíceis, ou com somente itens medianos, ou ainda testes compostos por uma mistura de itens fáceis, difíceis e medianos. A secção (4.4) apresenta um breve estudo sobre intervalos de confiança bootstrap para os parâmetros da função de resposta de item. A secção (4.5) apresenta uma ilustração.

4.2 Procedimento de Baker

Dado um teste composto de n itens dicotômicos aplicado a m indivíduos, supondo o modelo logístico de dois parâmetros, o procedimento de Baker define a estimação destes $2n+m$ parâmetros em dois estágios. No primeiro estágio existe uma redução dos parâmetros relativos aos indivíduos de m para k parâmetros e a finalidade é estimar os parâmetros relativos aos itens. Neste primeiro estágio, para se obter as estimativas dos parâmetros de itens, é utilizado o método do mínimo logito qui-quadrado (Bishop e outros, 1975). No segundo estágio, as estimativas dos parâmetros de itens obtidas no primeiro estágio são utilizadas para obter as estimativas dos parâmetros relativos aos indivíduos e para tal é utilizado o método de máxima verossimilhança.

O objetivo desta secção é descrever o procedimento formulado por Baker (1987) com um maior detalhamento que no artigo original.

4.2.1 Primeiro Estágio: Estimação dos Parâmetros de Itens segundo o Método do Mínimo Logito Qui-Quadrado

O primeiro estágio do procedimento de Baker tem como objetivo a estimação de parâmetros referentes aos itens (a_i e b_i, $i = 1, 2, \ldots, n$). Uma simplificação é feita: os m indivíduos são divididos em k grupos, sendo que Baker em seu artigo não sugere como deve ser feita esta divisão. Supõe-se que todos os indivíduos pertencentes ao l-ésimo grupo possuem a mesma habilidade $\tilde{\theta}_l$, $l = 1, 2, \ldots, k$. Sejam f_l o número de indivíduos no l-ésimo grupo, e r_{il} o número de respostas corretas para o i-ésimo item, no l-ésimo grupo. Logo, a proporção de respostas corretas ao i-ésimo item, observadas no l-ésimo grupo, é dada por $p_{il} = r_{il}/f_l$ e $q_{il} = 1 - p_{il}$ é a proporção observada de respostas incorretas, pois os itens são dicotômicos. Assim, para o i-ésimo item do teste, tem-se a seguinte tabela $k \times 2$:

	item i	
grupos	*acertos*	*erros*
1	r_{i1}	$f_1 - r_{i1}$
2	r_{i2}	$f_2 - r_{i2}$
⋮	⋮	⋮
k	r_{ik}	$f_k - r_{ik}$

De acordo com o modelo logístico de dois parâmetros (4.1), a probabilidade de acerto ao i-ésimo item, por qualquer indivíduo do l-ésimo grupo, é dada por

$$P_i(\tilde{\theta}_l) = \frac{1}{1 + \exp[-1.7a_i(\tilde{\theta}_l - b_i)]}, \quad i = 1, 2, \ldots, n, \quad l = 1, 2, \ldots, k \tag{4.2}$$

e este modelo pode ser reparametrizado, definindo $a_i = \lambda_i/1,7$ e $b_i = -\xi_i/\lambda_i$:

$$P_i(\tilde{\theta}_l) = \frac{1}{1 + \exp[-(\xi_i + \lambda_i\tilde{\theta}_l)]},$$

e consequentemente,

$$Q_i(\tilde{\theta}_l) = \frac{\exp[-(\xi_i + \lambda_i\tilde{\theta}_l)]}{1 + \exp[-(\xi_i + \lambda_i\tilde{\theta}_l)]}.$$

Sob este modelo, a frequência esperada de acertos do i-ésimo item, por indivíduos do l-ésimo grupo, é dada por $f_l P_i(\tilde{\theta}_l)$.

A partir deste ponto, $P_i(\tilde{\theta}_l)$ é denotado por P_{il} e $Q_i(\tilde{\theta}_l)$ é denotado por Q_{il}.

Baseada nas diferenças entre as frequências de acertos e erros, observadas e esperadas, para o i-ésimo item, em todos k grupos de indivíduos, pode ser definida a seguinte estatística:

$$\begin{aligned}
\chi_i^2 &= \sum_{l=1}^{k}\left\{\frac{(r_{il} - f_l P_{il})^2}{f_l P_{il}} + \frac{(f_l - r_{il} - f_l Q_{il})^2}{f_l Q_{il}}\right\} \\
&= \sum_{l=1}^{k}\left\{\frac{(f_l p_{il} - f_l P_{il})^2}{f_l P_{il}} + \frac{(f_l q_{il} - f_l Q_{il})^2}{f_l Q_{il}}\right\} \\
&= \sum_{l=1}^{k}\left\{\frac{f_l(p_{il} - P_{il})^2}{P_{il}} + \frac{f_l(q_{il} - Q_{il})^2}{Q_{il}}\right\} \\
&= \sum_{l=1}^{k}\left\{\frac{f_l(p_{il} - P_{il})^2}{P_{il}} + \frac{f_l[1 - p_{il} - (1 - P_{il})]^2}{1 - P_{il}}\right\} \\
&= \sum_{l=1}^{k}\frac{f_l(1 - P_{il})(p_{il} - P_{il})^2 + f_l P_{il}[P_{il} - p_{il}]^2}{P_{il}(1 - P_{il})} \\
&= \sum_{l=1}^{k}\frac{f_l(p_{il} - P_{il})^2(1 - P_{il} + P_{il})}{P_{il}(1 - P_{il})} \\
&= \sum_{l=1}^{k}\frac{f_l(p_{il} - P_{il})^2}{P_{il}(1 - P_{il})} \\
&= \sum_{l=1}^{k}\frac{f_l(p_{il} - P_{il})^2}{P_{il}Q_{il}}. \qquad (4.3)
\end{aligned}$$

Uma vez definida a estatística χ_i^2, ela pode ser expressa em termos de quantidades denominadas "logito". O logito de P_{il} é definido por $\ln(\frac{P_{il}}{Q_{il}})$. Usando L_{il} para denotar o logito de P_{il},

$$L_{il} = \ln\left(\frac{1 + \exp[-(\xi_i + \lambda_i\tilde{\theta}_l)]}{\{1 + \exp[-(\xi_i + \lambda_i\tilde{\theta}_l)]\}\exp[-(\xi_i + \lambda_i\tilde{\theta}_l)]}\right)$$

$$= \ln[\exp(\xi_i + \lambda_i\tilde{\theta}_l)]$$
$$= \xi_i + \lambda_i\tilde{\theta}_l.$$

Seja $l_{il} = \ln(\frac{p_{il}}{q_{il}})$ o logito observado. Valores de $p_{il} = 0$ ou $p_{il} = 1$ resultam em indefinições. Para contornar este fato, Baker sugere a definição de $p_{il} = 1 - \frac{1}{2f_l}$, quando $p_{il} = 1$, e $p_{il} = \frac{1}{2f_l}$, quando $p_{il} = 0$.

Portanto, pode-se aproximar o logito, definido anteriomente, pela fórmula de Taylor, até o termo de primeira ordem, no ponto p_{il}, uma vez que, $\frac{\partial L_{il}}{\partial P_{il}} = \frac{1-P_{il}}{P_{il}}\frac{1}{(1-P_{il})^2} = \frac{1}{P_{il}(1-P_{il})}$ é bem definida em qualquer vizinhança de p_{il}. Assim,

$$\begin{aligned} L_{il} &\cong l_{il} + \frac{P_{il} - p_{il}}{1!}\frac{\partial L_{il}}{\partial P_{il}}\,|_{P_{il}=p_{il}} \\ &\cong l_{il} + (P_{il} - p_{il})\frac{1}{p_{il}(1-p_{il})} = l_{il} + \frac{(P_{il} - p_{il})}{p_{il}q_{il}}, \end{aligned}$$

portanto, $(p_{il} - P_{il})$ pode ser aproximado por

$$(p_{il} - P_{il}) \cong (l_{il} - L_{il})p_{il}q_{il}.$$

Desta forma,

$$(p_{il} - P_{il})^2 \cong (l_{il} - L_{il})^2 p_{il}^2 q_{il}^2,$$

ou ainda,

$$(p_{il} - P_{il})^2 \cong P_{il}Q_{il}(l_{il} - L_{il})^2 p_{il}q_{il}. \tag{4.4}$$

Substituindo (4.4) em (4.3) tem-se,

$$\begin{aligned} \chi_i^2 &= \sum_{l=1}^{k} \frac{f_l P_{il}Q_{il}(l_{il} - L_{il})^2 p_{il}q_{il}}{P_{il}Q_{il}} \\ &= \sum_{l=1}^{k} f_l p_{il}q_{il}(l_{il} - L_{il})^2 \\ &= \sum_{l=1}^{k} w_{il}(l_{il} - L_{il})^2, \end{aligned} \tag{4.5}$$

onde $w_{il} = f_l p_{il} q_{il}$.

A equação (4.5) pode ser escrita em forma matricial por

$$\chi_i^2 = [\mathbf{l}_i - \Theta\boldsymbol{\Gamma}_i]'\mathbf{W}_i^{-1}[\mathbf{l}_i - \Theta\boldsymbol{\Gamma}_i],$$

onde,

$$\mathbf{l}_i = \begin{bmatrix} l_{i1} \\ l_{i2} \\ \vdots \\ l_{ik} \end{bmatrix}$$

$$\Theta = \begin{bmatrix} 1 & \tilde{\theta}_1 \\ 1 & \tilde{\theta}_2 \\ \vdots & \vdots \\ 1 & \tilde{\theta}_k \end{bmatrix}$$

$$\boldsymbol{\Gamma}_i = \begin{bmatrix} \xi_i \\ \lambda_i \end{bmatrix}$$

$$\mathbf{W}_i^{-1} = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{ik} \end{bmatrix}.$$

Assim, supondo as habilidades dos grupos $\tilde{\theta}_l$ conhecidas, o vetor $\boldsymbol{\Gamma}_i$ contém os parâmetros do i-ésimo item a serem estimados. O objetivo é obter estimador para $\boldsymbol{\Gamma}_i$ minimizando χ_i^2. Tecnicamente, o problema é idêntico ao de se encontrar estimadores da regressão ponderada de $\mathbf{l}_i$ em Θ. Segundo o método dos mínimos quadrados, o valor de $\boldsymbol{\Gamma}_i$ que minimiza a expressão χ_i^2 é dado por

(4.6)
$$\begin{aligned} \hat{\boldsymbol{\Gamma}}_i &= [\Theta'\mathbf{W}_i^{-1}\Theta]^{-1}\Theta'\mathbf{W}_i^{-1}\mathbf{l}_i \\ &= \begin{bmatrix} \hat{\xi}_i \\ \hat{\lambda}_i \end{bmatrix} \end{aligned}$$

e,

$$Var(\hat{\boldsymbol{\Gamma}}_i) = [\Theta' \mathbf{W}_i^{-1} \Theta]^{-1}, \quad i = 1, 2, \ldots, n, \tag{4.7}$$

obtendo assim uma forma fechada para a estimação dos parâmetros de itens. Ou seja, se as habilidades de cada grupo de indivíduos são conhecidas (ou utilizando as estimativas como valores verdadeiros) as estimativas dos parâmetros de cada item são obtidas diretamente.

Todo o desenvolvimento exposto acima pode ser reformulado, similarmente, analisando unicamente os indivíduos de um mesmo grupo. Para o l-ésimo grupo de indivíduos tem-se a seguinte tabela $n \times 2$:

	grupo l	
itens	*acertos*	*erros*
1	r_{1l}	$f_l - r_{1l}$
2	r_{2l}	$f_l - r_{2l}$
$\vdots$	$\vdots$	$\vdots$
n	r_{nl}	$f_l - r_{nl}$

e para o l-ésimo grupo de indivíduos tem-se a seguinte estatística baseada na diferença entre as frequências, observadas e esperadas, de acertos e erros do l-ésimo grupo, em todos os n itens,

$$\begin{aligned} G_l^2 &= \sum_{i=1}^{n} \left\{ \frac{(r_{il} - f_l P_{il})^2}{f_l P_{il}} + \frac{(r_{il} - f_l Q_{il})^2}{f_l Q_{il}} \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{(f_l p_{il} - f_l P_{il})^2}{f_l P_{il}} + \frac{(f_l q_{il} - f_l Q_{il})^2}{f_l Q_{il}} \right\} \\ &= \sum_{i=1}^{n} f_l p_{il} q_{il} (l_{il} - L_{il})^2 \\ &= \sum_{i=1}^{n} w_{il} (l_{il} - \xi_i - \lambda_i \tilde{\theta}_l)^2. \end{aligned} \tag{4.8}$$

A expressão (4.8) pode ser reduzida à seguinte notação matricial,

$$G_l^2 = [\Upsilon_l - \lambda\tilde{\theta}_l]'\mathbf{W}_l^{-1}[\Upsilon_l - \lambda\tilde{\theta}_l],$$

onde,

$$\Upsilon_l = \begin{bmatrix} l_{1l} - \xi_1 \\ l_{2l} - \xi_2 \\ \vdots \\ l_{nl} - \xi_n \end{bmatrix},$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix},$$

e

$$\mathbf{W}_l^{-1} = \begin{bmatrix} w_{1l} & 0 & \cdots & 0 \\ 0 & w_{2l} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{nl} \end{bmatrix}.$$

Assim, supondo os parâmetros de itens conhecidos (ou usando as estimativas de λ e ξ como se fossem valores verdadeiros) tem-se o seguinte estimador para a habilidade $\hat{\theta}_l$ do l-ésimo grupo:

$$\hat{\theta}_l = (\lambda'\mathbf{W}_l^{-1}\lambda)^{-1}\lambda'\mathbf{W}_l^{-1}\Upsilon_l. \tag{4.9}$$

Usando os resultados (4.6) e (4.9), Baker propõe o seguinte processo iterativo:

1. São estabelecidos valores iniciais para as habilidades dos grupos de indivíduos, $\tilde{\theta}_1^{(0)}, \tilde{\theta}_2^{(0)}, \ldots, \tilde{\theta}_k^{(0)}$. Baker não explicitou a definição destes valores iniciais.

2. Usando os valores iniciais para as habilidades dos grupos de indivíduos, como se fossem os valores verdadeiros, estimam-se os parâmetros de itens, λ_i e ξ_i, por (4.6), separadamente para cada um dos itens ($i = 1, 2, \ldots, n$).

3. Usando as estimativas dos parâmetros de itens, $\lambda_1, \lambda_2, \ldots, \lambda_n, \xi_1, \xi_2, \ldots, \xi_n$, obtidas em (2) como se fossem valores verdadeiros, estimam-se as habilidades dos grupos de indivíduos, $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k$, por (4.9), separadamente para cada um dos grupos de indivíduos.

4. Repetem-se (2) e (3) até algum critério de parada. Baker define como critério de parada que as diferenças das estimativas dos parâmetros de itens entre uma iteração e outra, obtidas por (4.6), e as diferenças entre uma iteração e outra das variâncias (dos estimadores dos parâmetros de itens) estimadas dadas por (4.7), não sejam maiores que $0,05$.

No segundo estágio são usadas as estimativas dos parâmetros de itens obtidas na última iteração. As estimativas das habilidades dos grupos de indivíduos não são utilizadas no segundo estágio.

4.2.2 Segundo Estágio: Estimação das Habilidades dos Indivíduos segundo o Método de Máxima Verossimilhança

No primeiro estágio são estimados os parâmetros de itens
$\lambda_1, \lambda_2, \ldots, \lambda_n$, $\xi_1, \xi_2, \ldots, \xi_n$. Como estes parâmetros se relacionam diretamente com
$a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$, pela reparametrização definida na secção 4.2.1, Baker propõe que estas estimativas sejam utilizadas como valores verdadeiros de $a_1, a_2, \ldots, a_n$, $b_1, b_2, \ldots, b_n$, para se obter as estimativas dos parâmetros associados aos indivíduos pelo método de máxima verossimilhança.

A função de verossimilhança (Capítulo 3), para o modelo logístico de dois parâmetros,

supondo os parâmetros de itens conhecidos, é dada por

$$L(\theta \mid \mathbf{u}) = \prod_{i=1}^{n}\prod_{j=1}^{m} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}}$$

para $u_{ij} = 0$ ou 1, onde,

$$\theta = (\theta_1, \theta_2, \cdots, \theta_m)',$$
$$\mathbf{u} = \begin{bmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nm} \end{bmatrix}.$$

O logarítmo natural da função de verossimilhança é dado por

$$l = ln[L(\theta \mid \mathbf{u})] = \sum_{i=1}^{n}\sum_{j=1}^{m}[u_{ij} \ln P_{ij} + (1 - u_{ij}) \ln Q_{ij}].$$

Logo, o estimador de máxima verossimilhança da habilidade do j-ésimo indivíduo é dado pela solução de:

$$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n} \frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}} \frac{\partial P_{ij}}{\partial \theta_j} = 0,$$

onde

$$\frac{\partial P_{ij}}{\partial \theta_j} = \frac{1,7a_i}{\exp[1,7a_i(\theta_j - b_i)] + 2 + \exp[-1,7a_i(\theta_j - b_i)]},$$
$$i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

Como não existe uma forma fechada para tal solução, Baker propõe o procedimento iterativo de Newton-Raphson, utilizando valores iniciais (não indicou quais) para as habilidades de cada indivíduo.

4.3 Simulações

Simulações são utilizadas para verificar a adequabilidade do procedimento de Baker em algumas situações que se acredita poderem ser encontradas na prática. Foram simulados

testes com somente itens fáceis, somente itens difíceis ou somente itens medianos ou ainda testes compostos por itens de dificuldades variadas. Os testes com itens de dificuldades variadas são aqui denominados "testes misturados", uma vez que são compostos por quantidades iguais de itens fáceis, difíceis e itens medianos. Baker (1987) considera, em suas simulações, testes difíceis aqueles com parâmetros de dificuldade (b) em média iguais a 1, medianos em média iguais a 0 e fáceis em média iguais a -1, e variâncias pequenas, utilizando-se da distribuição normal e parâmetros de discriminação (a) com valores uniformes entre 0,5 e 1,5. Autores como, Swaminathan e Gifford (1986) e Yen (1987) utilizam em suas simulações padrões de testes similares aos de Baker. Ainda, as habilidades dos indivíduos são amostras pseudo-aleatórias de distribuições não normais e normais. Neste trabalho também utilizadas simulações para comparar as estimativas dos parâmetros obtidas através do procedimento de Baker com estimativas obtidas por uma modificação neste mesmo procedimento, aqui denominado "procedimento de Baker modificado".

Como foi visto nas secções anteriores o procedimento de Baker requer uma estimação de parâmetros em dois estágios. No primeiro estágio é utilizado o método do mínimo logito qui-quadrado, onde existe uma redução no número de parâmetros referentes aos indivíduos (θ). Os m indivíduos aos quais é aplicado o teste são agrupados em k grupos homogêneos com a finalidade de estimar os parâmetros referentes aos itens (a e b). Estas habilidades são desconhecidas, tornando necessária a utilização de um valor inicial. A habilidade de um indivíduo é definida em uma escala que varia de $-\infty$ a $+\infty$ pelos modelos aqui apresentados. Bock e Aitkin (1981) e Birnbaum (1968) utilizam o inverso da função de distribuição acumulada da normal padrão da proporção de acertos de cada indivíduo. Este valor inicial também foi utilizado para as simulações apresentadas neste trabalho. Assim, o valor inicial para a habilidade do j-ésimo indivíduo é definido como:

$$\theta_j^{(0)} = \phi^{-1}\left(\sum_{i=1}^{n} \frac{u_{ij}}{n}\right), \quad j = 1, 2, \ldots, m, \tag{4.10}$$

onde ϕ^{-1} é o inverso da função de distribuição acumulada da normal padrão da proporção de acertos de do j-ésimo indivíduo. Quando o indivíduo acerta a todos os itens o valor inicial de $\theta_j^{(0)} = 3$ é atribuído e quando erra todos os itens, $\theta_j^{(0)} = -3$, uma vez que são valores de habilidades bastante altas e bastante baixas, respectivamente (estando à distância de três desvios padrões da média do modelo normal padrão). Para obter os k grupos de indivíduos os m indivíduos foram alocados em k intervalos de classe, para a habilidade, de amplitudes iguais, sendo a amplitude dos intervalos é dada por

$$amplitude\ \ dos\ \ intervalos = \frac{\theta_{(m)}^{(0)} - \theta_{(1)}^{(0)}}{k},$$

onde $\theta_{(j)}$ representa o j-ésimo valor ordenado de forma crescente. Assim, cada grupo de indivíduos é formado por indivíduos com aproximadamente a mesma proporção de acertos, com f_l elementos em cada grupo. A habilidade de cada grupo é dada pela média dos valores iniciais para as habilidades dos indivíduos alocados ao grupo,

$$\tilde{\theta}_1^{(0)} = \sum_{j=1}^{f_1} \frac{\theta_{(j)}^{(0)}}{f_1}, \tilde{\theta}_2^{(0)} = \sum_{j=f_1+1}^{f_2} \frac{\theta_{(j)}^{(0)}}{f_2}, \ldots, \tilde{\theta}_k^{(0)} = \sum_{j=f_1+f_2+\ldots+f_{k-1}+1}^{f_k} \frac{\theta_{(j)}^{(0)}}{f_k},$$

e os grupos com frequência zero não são considerados.

O número de grupos, k, sugerido por Baker (1987), é $k = 5$ quando $n > m$, ou seja, quando o número de itens é maior que o número de examinandos e 11 quando $m > n$, ou seja, quando o número de examinandos é maior do que o número de itens.

As duas fases do primeiro estágio do procedimento de Baker se repetem até algum critério de parada. Como critério de parada foi utilizado a seguinte regra: os valores absolutos das diferenças das estimativas dos parâmetros referentes aos itens e suas respectivas estimativas de variâncias entre uma iteração e outra (sendo as variâncias foram obtidas segundo a equação (4.7)) deveriam ser no máximo da ordem de 10^{-3}.

Depois de calculadas as estimativas dos parâmetros referentes aos itens (a e b) no primeiro estágio, o procedimento de Baker entra em seu segundo estágio. No segundo estágio as estimativas dos parâmetros referentes a itens obtidas no primeiro estágio são utilizadas para obter as estimativas das habilidades dos indivíduos, onde é utilizado o método de máxima verossimilhança. Como não existe uma forma fechada para tais estimadores, se torna necessário o uso de um procedimento iterativo. Lord (1980) sugere o método do "scoring" de Fisher (Apêndice A). Logo o processo iterativo para estimação de θ_j (j=1,2,...,m), obtido pelo método do "scoring" de Fisher, é dado por,

$$\hat{\theta}_j^{(m)} = \hat{\theta}_j^{(m-1)} + \sum_{i=1}^{n} \frac{\exp[-1,7a_i(\theta_j^{(m-1)} - b_i)] + 2 + \exp[1,7a_i(\theta_j^{(m-1)} - b_i)]}{1,7a_i}(u_{ij} - P_{ij})$$

Como um valor inicial para as habilidades dos indivíduos se torna necessário neste procedimento iterativo, utilizam-se os mesmos valores dados por (4.10). O critério de parada adotado foi o de que a maior diferença entre as estimativas obtidas entre uma iteração e outra fosse menor que 10^{-3}. Quando um indivíduo acerta ou erra todos os itens, tais estimativas de máxima verossimilhança divergem, ou seja, tendem a ∞ e $-\infty$, respectivamente, assim as estimativas das habilidades de tais indivíduos não podem ser encontradas a menos que limites sejam apontados (Lord, 1980), neste trabalho, essas estimativas não foram consideradas.

Para a realização das simulações foi necessária a criação de vários programas em linguagem IML (Iteractive Matrix Language) parte integrante do SAS (Statistical Analysis System). Esta linguagem foi escolhida, principalmente, pela confiabilidade em seus procedimentos. O programa que executa a modificação proposta ao procedimento de Baker, denominada "procedimento de Baker modificado", se encontra no Apêndice B; esta modificação restringe a criação de grupos de indivíduos com pelo menos cinco elementos.

4.3.1 Descrição do planejamento das simulações

A primeira fase de simulações visa comparar os procedimentos de Baker padrão e do modificado. Nesta fase é também verificada a sensibilidade das estimativas dos parâmetros em relação ao tipo de teste aplicado a um grupo de indivíduos com habilidades provenientes de amostra pseudo aleatória com distribuição normal padrão. Com base nos resultados obtidos nesta primeira fase de simulações é feita uma opção por um dos dois procedimentos. Uma segunda fase de simulações pretende avaliar as estimativas obtidas pelo procedimento de Baker (escolhido) em relação ao número de itens no teste e ao número de indivíduos ao qual o teste é aplicado. Em uma terceira fase de simulações, pretende-se pesquisar o efeito das estimativas dos parâmetros em aplicações de testes a grupos de indivíduos provenientes de amostras com um maior número de indivíduos com habilidades altas ou ainda com um maior número de indivíduos habilidades baixas. Na tabela 4.1 são esquematizadas tais simulações,

4.3.2 Resultados das simulações

Para estudar o comportamento das estimativas obtidas nas diferentes situações simuladas, são utilizados alguns critérios baseados em coeficientes de correlação e erros quadráticos médios, além da observação de gráficos. Os coeficientes de correlação e os erros quadráticos médios são obtidos entre parâmetros simulados e estimativas destes mesmos parâmetros. Outro critério é a correlação entre parâmetro simulado e resíduo, assim, quanto menor esta correlação mais favorável é o resultado, casos com correlação alta merecem um estudo mais detalhado. Os gráficos são basicamente diagramas de dispersão entre parâmetros simulados e estimativas destes parâmetros; estes gráficos são acrescidos de uma reta que representa a igualdade entre estimativa e parâmetro simulado.

primeira fase de simulações					
número de indivíduos	*número de itens*	*a gerados segundo*	*b gerados segundo*	*θ gerados segundo*	*classificação do teste*
200	150	$U(0,5;1,5)$	$N(-1,(0,5)^2)$	$N(0,1)$	fácil
			$N(0,(0,5)^2)$		mediano
			$N(1,(0,5)^2)$		difícil
			$N(-1,(0,5)^2)$ $N(0,(0,5)^2)$ $N(1,(0,5)^2)$		misturado
segunda fase de simulações					
50	50	$U(0,5;1,5)$	$N(0,(0,5)^2)$	$N(0,1)$	mediano
	100				mediano
	150				mediano
100	50				mediano
	100				mediano
	150				mediano
200	50				mediano
	100				mediano
	150				mediano
terceira fase de simulações					
200	150	$U(0,5;1,5)$	$N(-1,(0,5)^2)$ $N(0,(0,5)^2)$ $N(1,(0,5)^2)$	$N(2,1)^a$	misturado
			$N(1,(0,5)^2)$	$N(2,1)^a$	difícil
			$N(-1,(0,5)^2)$	$N(2,1)^a$	fácil
			$N(-1,(0,5)^2)$ $N(0,(0,5)^2)$ $N(1,(0,5)^2)$	$N(-2,1)^b$	misturado
			$N(1,(0,5)^2)$	$N(-2,1)^b$	difícil
			$N(-1,(0,5)^2)$	$N(-2,1)^b$	fácil

a: truncada em θ=3, *b*: truncada em θ=-3

Tabela 4.1: Esquema das três fases de simulações

Os resultados da primeira fase de simulações estão contidos na Tabela 4.2. Observa-se que com o emprego do procedimento de Baker modificado, na maioria das vezes, ocorreram diminuições significativas nos erros quadráticos médios, principalmente para o parâmetro habilidade (θ), uma vez que este parâmetro assume, valores entre -3 e 3. Outro fato importante é que menores correlações entre parâmetro e resíduo foram obtidas no procedimento de Baker modificado, sendo que a menor foi a do teste classificado como "misturado". Como no procedimento de Baker modificado o número de indivíduos no grupo é controlado, não existem grupos muito pequenos que seriam grupos de pouca informação. Com base nestes resultados, o procedimento de Baker modificado, passa a ser utilizado nas próximas simulações. Outros resultados desta fase de simulações podem ser enumerados abaixo:

- No caso de testes classificados como "medianos", para os dois procedimentos, ocorreu uma subestimação para pequenos valores de θ, e uma superestimação para grandes valores de θ. O mesmo ocorreu com relação ao parâmetro b. O parâmetro a, nos dois procedimentos, foi me geral subestimado. Estes fatos podem ser ilustrados pela Figura (4.1), que apresenta os resultados obtidos com o procedimento de Baker padrão, uma vez que os obtidos com o procedimento modificado são similares.

- Para testes classificados como "fáceis", para os dois procedimentos, ocorreu uma superestimação para os parâmetros θ e b. O parâmetro a foi quase sempre subestimado. A Figura (4.2) ilustra estes fatos para os resultados obtidos com o procedimento de Baker padrão, uma vez que os resultados obtidos nos dois procedimentos são similares.

- Em testes "difíceis", para os dois procedimentos, ocorreu uma subestimação para os parâmetros θ, b e a. Estes fatos podem ser observados na Figura (4.3) para os resultados obtidos com o procedimento de Baker padrão, uma vez que os resultados com o procedimento modificado são similares.

- Para a classificação de testes "misturados", com a utilização do procedimento de Baker padrão, ocorreu uma subestimação para pequenos valores de θ e uma superestimação para grandes valores de θ, o mesmo fato ocorreu com o parâmetro b. Com a utilização do procedimento de Baker modificado não ocorreram subestimações nem superestimações dos parâmetros b e θ. No caso do parâmetro a com a utilização do procedimento de Baker padrão, um menor número de subestimações foi verificado. Entretanto, com o procedimento de Baker modificado, a estimação do parâmetro a foi ainda menor. Estes fatos podem ser observados respectivamente nas Figuras (4.4) e (4.5).

Os resultados acima reafirmam a escolha do procedimento de Baker modificado neste trabalho. Pode-se observar que melhores resultados foram obtidos com a aplicação de testes classificados como "misturados", ou seja, formados por itens com dificuldades variadas. Este fato, pode ser um indicador de que em situações práticas seria melhor a aplicação de um teste com uma variabilidade maior de dificuldade, a fim de se obter melhores estimações de parâmetros de itens e principalmente de parâmetros referentes a indivíduos.

Os resultados da segunda fase de simulações estão sumarizados na Tabela 4.3. Observa-se que o aumento no número de itens é normalmente acompanhado por uma melhora no coeficiente de correlação e EQM para o parâmetro θ. Quanto maior o número de indivíduos ao qual o teste é aplicado maiores os coeficientes de correlação e menores os EQM. Pode-se verificar que os EQM dos parâmetros b para teste com 200 indivíduose 100 e 150 itens já são bastante pequenos, uma vez que este parâmetro assume, na maior parte dos casos, valores entre -3 e 3.

Os resultados da terceira fase de simulações estão contidos na Tabela (4.4), e podem ser enumerados:

Tipo de teste		*Mediano*		*Fácil*		*Difícil*		*Misturado*	
Procedimento		padr.	modif.	padr.	modif.	padr.	modif.	padr.	modif.
cor-	$a, \hat{a}$	0,823	0,835	0,747	0,735	0,779	0,786	0,636	0,664
re-	$b, \hat{b}$	0,960	0,964	0,927	0,925	0,931	0,939	0,963	0,982
la	$\theta, \hat{\theta}$	0,986	0,987	0,964	0,968	0,977	0,979	0,991	0,991
ção	$\theta, (\theta - \hat{\theta})$	-0,872	-0,674	-0,750	-0,693	-0,825	-0,513	-0,723	-0,062
	$a, \hat{a}$	0,119	0,035	0,163	0,089	0,153	0,043	0,106	0,046
EQM	$b, \hat{b}$	0,041	0,015	0,740	0,613	0,447	0,525	0,100	0,028
	$\theta, \hat{\theta}$	0,225	0,067	2,061	1,700	1,445	1,004	0,065	0,024

Tabela 4.2: Resultados referentes à primeira fase de simulações, com o objetivo de comparar os procedimentos de Baker padrão e o modificado (correlações e erros quadráticos médios (EQM) observados entre parâmetros simulados e estimativas)

- No caso de acumulo de indivíduos com habilidades altas (gerados segundo o modelo $N(2,1)$, truncado em 3) e teste "difícil", os coeficientes de correlação para todos os parâmetros foram maiores que na aplicação com teste "misturado". Entretanto, os EQM foram menores para o teste de classificação "misturado". Para o teste de classificação "difícil" ocorreu uma subestimação dos parâmetros b e θ. Com teste "fácil" os parâmetros b e θ foram superestimados, e foram obtidos os maiores EQM para todos os parâmetros.

- O caso de acumulo de indivíduos com habilidades baixas (gerados segundo o modelo $N(-2,1)$, não se utilizando valores menores do que -3) e teste fácil resultou em maiores coeficientes de correlação para todos os parâmetros, do que na aplicação de teste "misturado". Entretanto os EQM para as habilidades foram menores. Para teste "fácil" ocorreu uma superestimação dos parâmetros b e θ. Com teste "difícil" os parâmetros b e θ foram subestimados. Dos três tipos de testes neste grupo de indivíduos, o teste "difícil" foi o que resultou em menores coeficientes de correlação para

# de indivíduos		50			100			200		
# de itens		50	100	150	50	100	150	50	100	150
cor-	$a, \hat{a}$	0,611	0,704	0,603	0,816	0,803	0,717	0,903	0,843	0,835
re-	$b, \hat{b}$	0,862	0,882	0,663	0,945	0,919	0,933	0,976	0,956	0,964
la-	$\theta, \hat{\theta}$	0,961	0,984	0,986	0,959	0,976	0,981	0,965	0,982	0,987
ção	$\theta, (\theta - \hat{\theta})$	-0,814	-0,812	-0,854	-0,703	-0,795	-0,701	-0.701	-0,575	-0,674
	$a, \hat{a}$	0,215	0,227	0,366	0,070	0,124	0,107	0,026	0,024	0,035
EQM	$b, \hat{b}$	0,149	0,170	1,246	0,028	0,062	0,058	0,009	0,011	0,015
	$\theta, \hat{\theta}$	0,647	0,166	0,200	0,336	0,279	0,120	0,247	0,069	0,067

Tabela 4.3: Resultados referentes à segunda fase de simulações, com o objetivo de verificar a sensibilidade do procedimento de Baker modificado com relação ao número de indivíduos e itens (correlações e erros quadráticos médios (EQM) observados entre parâmetros simulados e estimativas); testes medianos

todos os parâmetros e maiores EQM para a maioria dos parâmetros. As Figuras (4.6), (4.7) e (4.8), apresentam os resultados para o caso de acumulo de indivíduos com habilidades altas, uma vez que o caso de acumulo de indivíduos com habilidades baixas apresenta resultados simétricos.

Os resultados das várias fases de simulações apontam que testes difíceis levam à subestimação dos parâmetros θ e b e testes fáceis levam à superestimação destes mesmos parâmetros. Os melhores resultados são obtidos quando o teste é composto por uma mistura de itens. Isto parece mostrar que as estimativas dos parâmetros de itens e das habilidades são mais sensíveis ao tipo de teste que ao grupo de indivíduos ao qual o teste foi aplicado.

θ *gerados segundo*		$N(2,1)^a$	$N(2,1)^a$	$N(2,1)^a$	$N(-2,1)^b$	$N(-2,1)^b$	$N(-2,1)^b$
Tipo de teste		*mist.*	*dif.*	*fác.*	*mist.*	*dif.*	*fác.*
cor-	$a, \hat{a}$	0,289	0,800	0,044	0,492	0,040	0,850
re-	$b, \hat{b}$	0,842	0,931	0,479	0,779	0,256	0,920
la	$\theta, \hat{\theta}$	0,962	0,987	0,808	0,952	0,851	0,983
ção	$\theta, (\theta - \hat{\theta})$	0,105	-0,228	-0,152	0,393	0, 334	-0,042
	$a, \hat{a}$	0,140	0,028	0,214	0,090	0,159	0,021
EQM	$b, \hat{b}$	0,635	0,752	1,665	0,986	8,582	0,984
	$\theta, \hat{\theta}$	0,093	0,927	2,254	0,253	0,621	1,216

a: truncada em θ=3

b: truncada em θ=-3

Tabela 4.4: Resultados referentes à terceira fase de simulações, com o objetivo de verificar efeito do procedimento de Baker modificado com relação a distribuições com acumulo de indivíduos com habilidades altas e baixas (correlações e erros quadráticos médios (EQM) observados entre parâmetros simulados e estimativas)

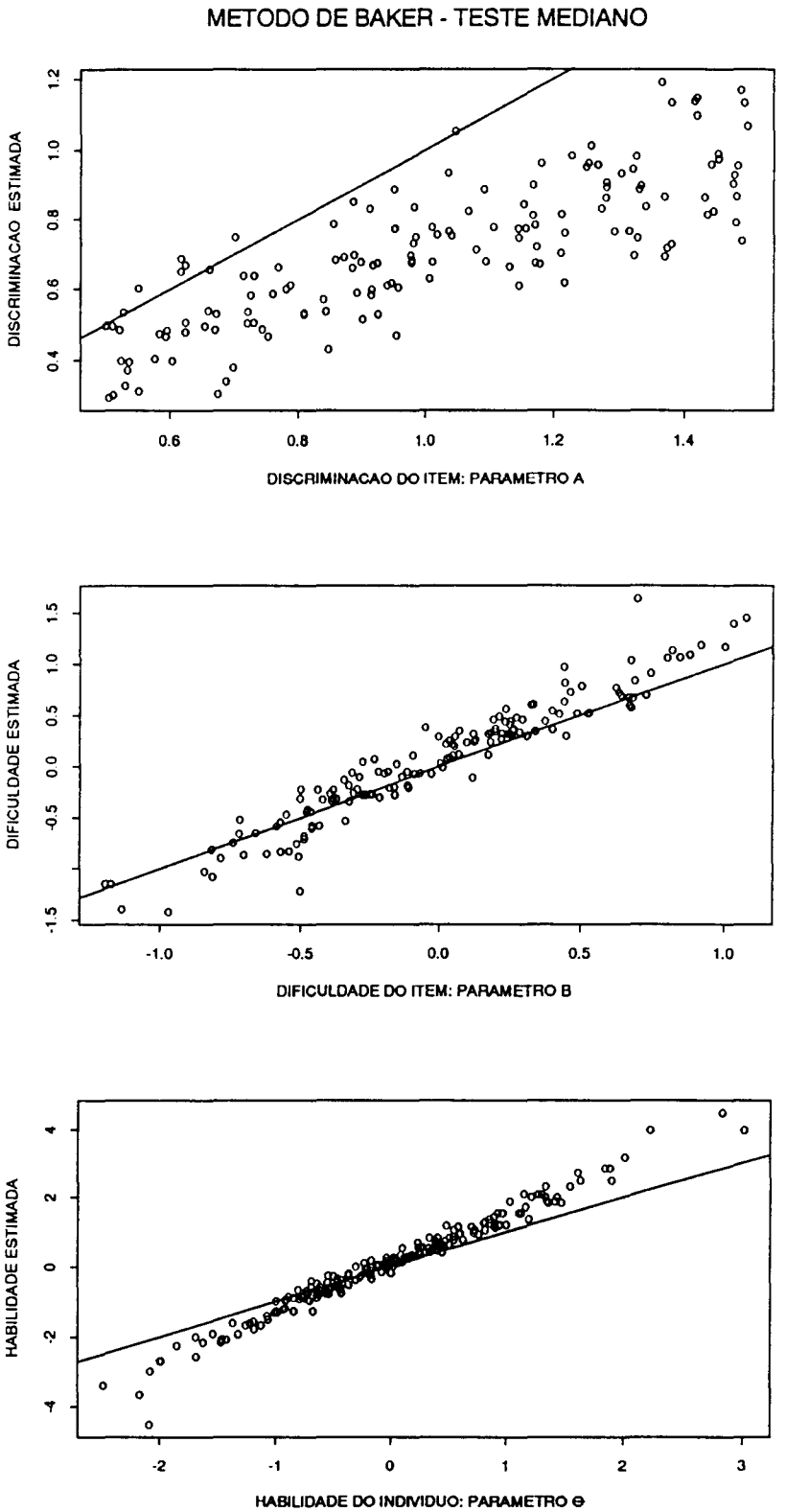


Figura 4.1: Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker padrão, de um teste "mediano", segundo a primeira fase de simulações

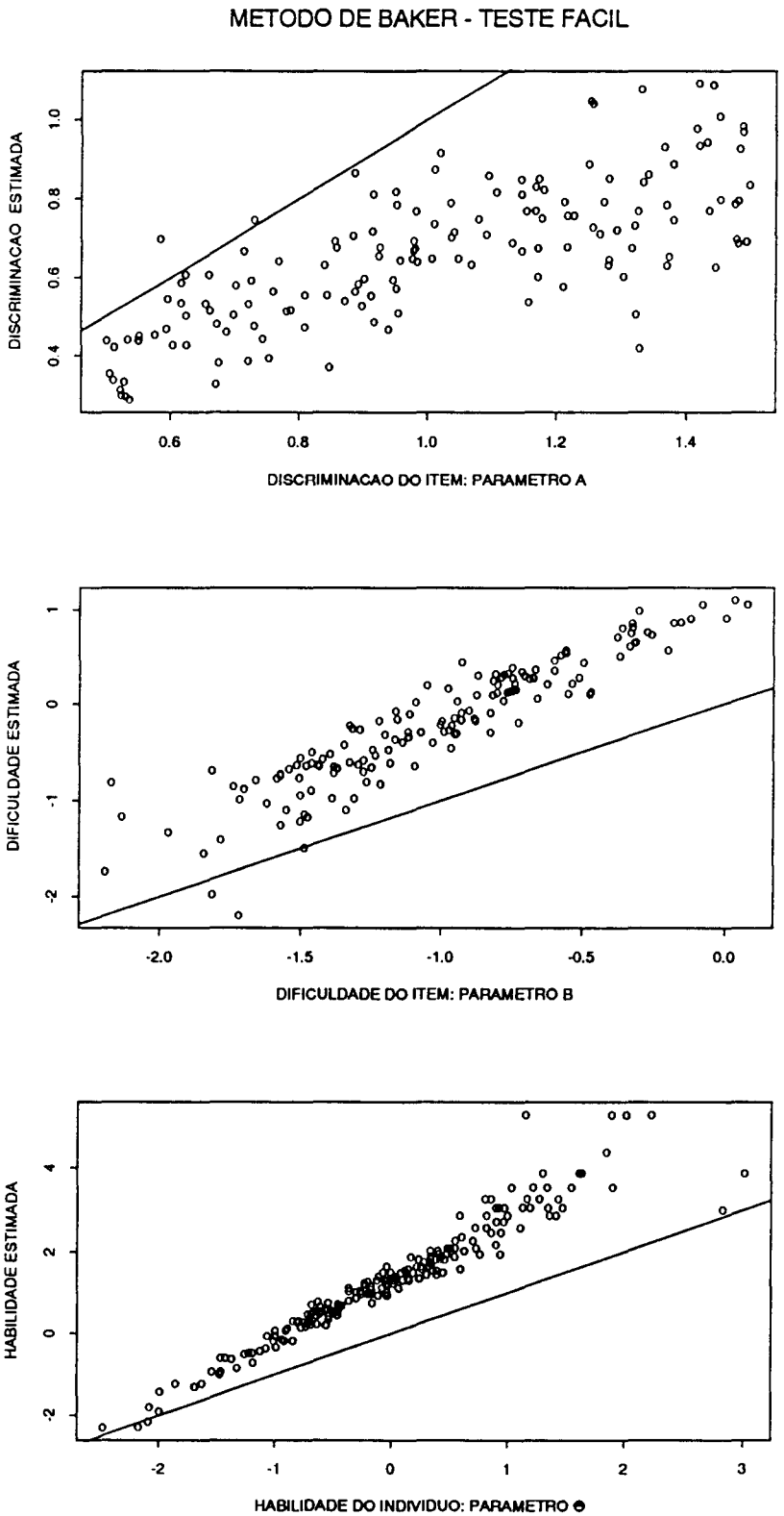


Figura 4.2: Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker padrão, de um teste "fácil", segundo a primeira fase de simulações

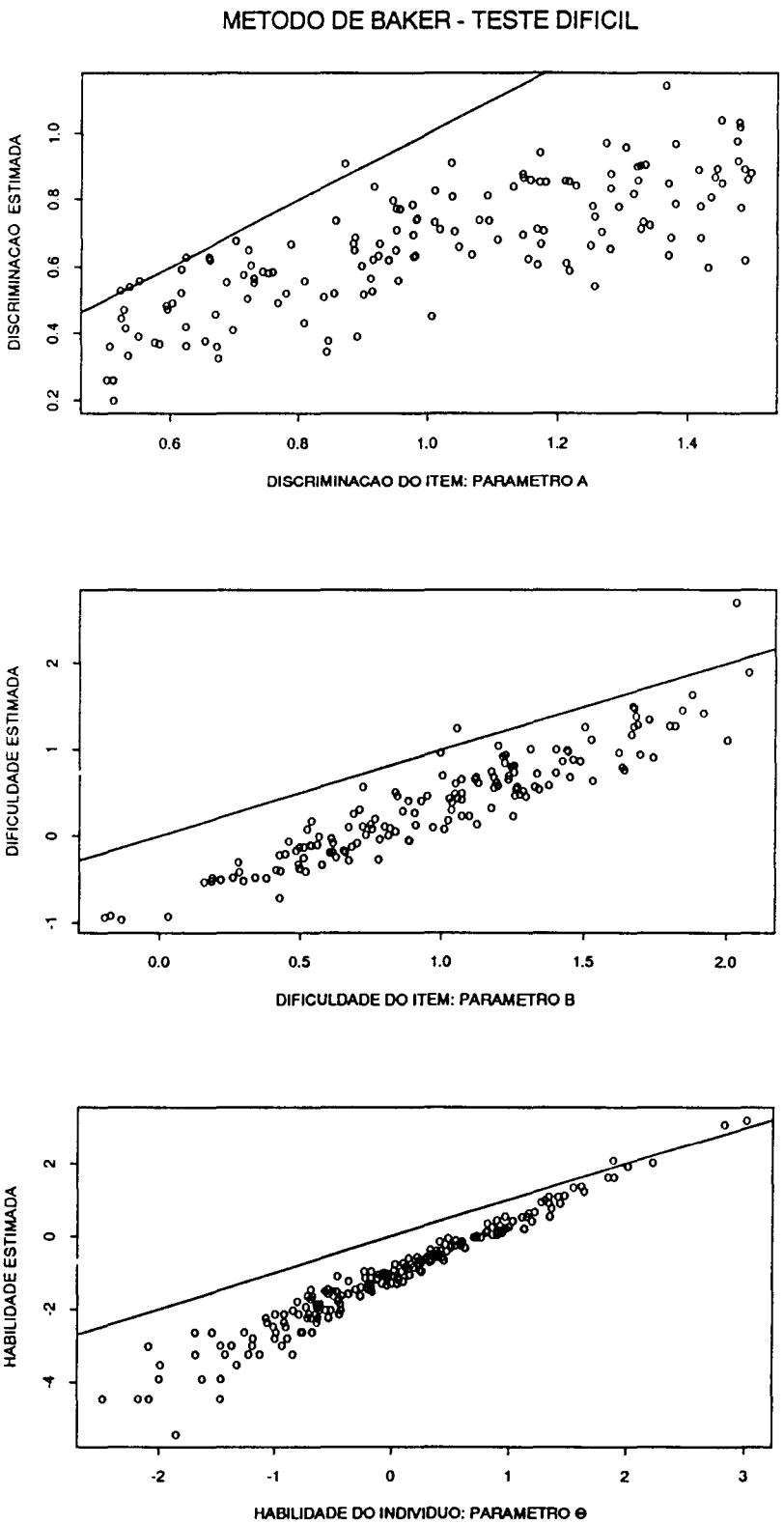


Figura 4.3: Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker padrão, de um teste "difícil", segundo a primeira fase de simulações

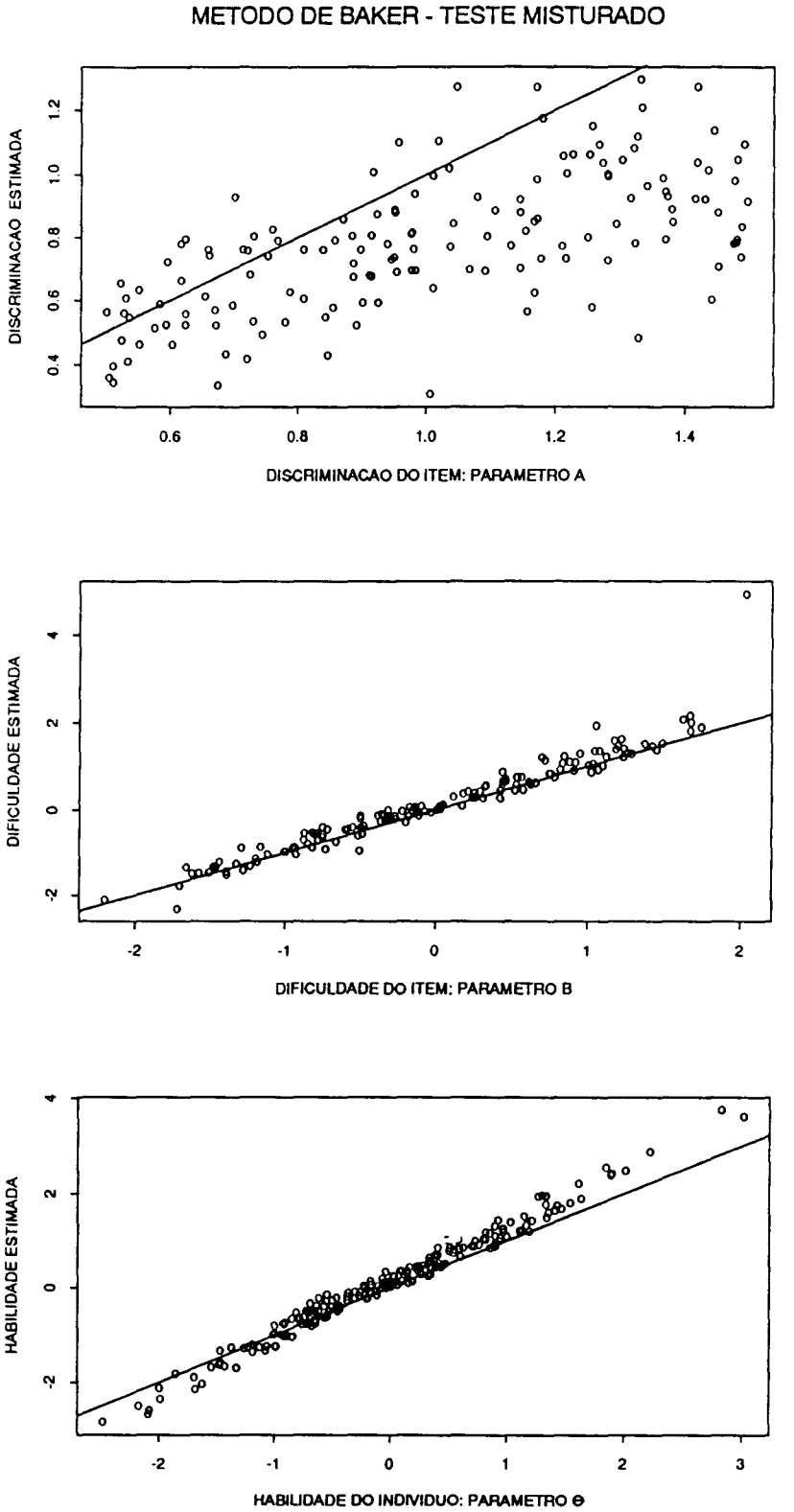


Figura 4.4: Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker padrão, de um teste "misturado", segundo a primeira fase de simulações

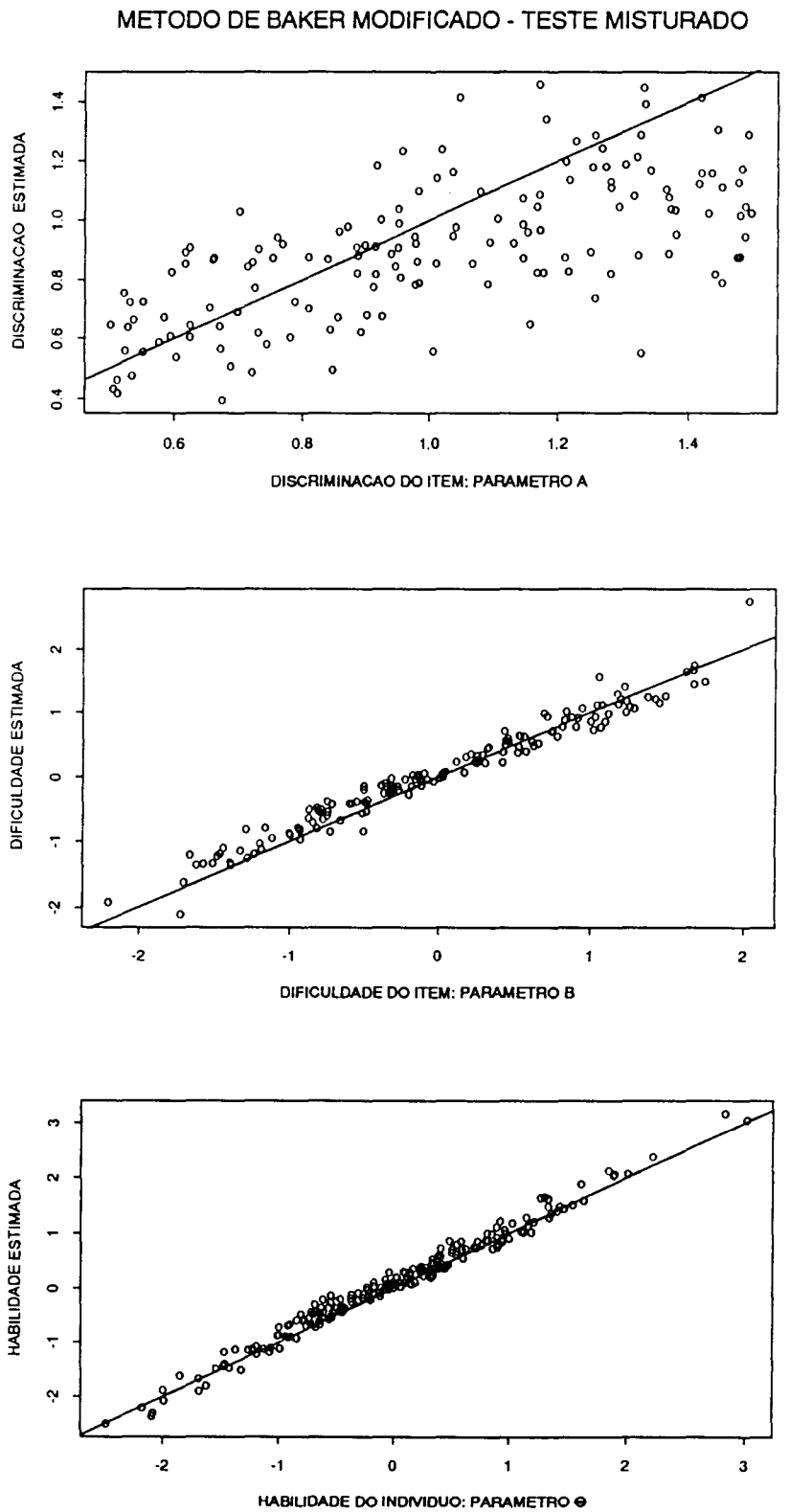


Figura 4.5: Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker modificado, de um teste "misturado", segundo a primeira fase de simulações

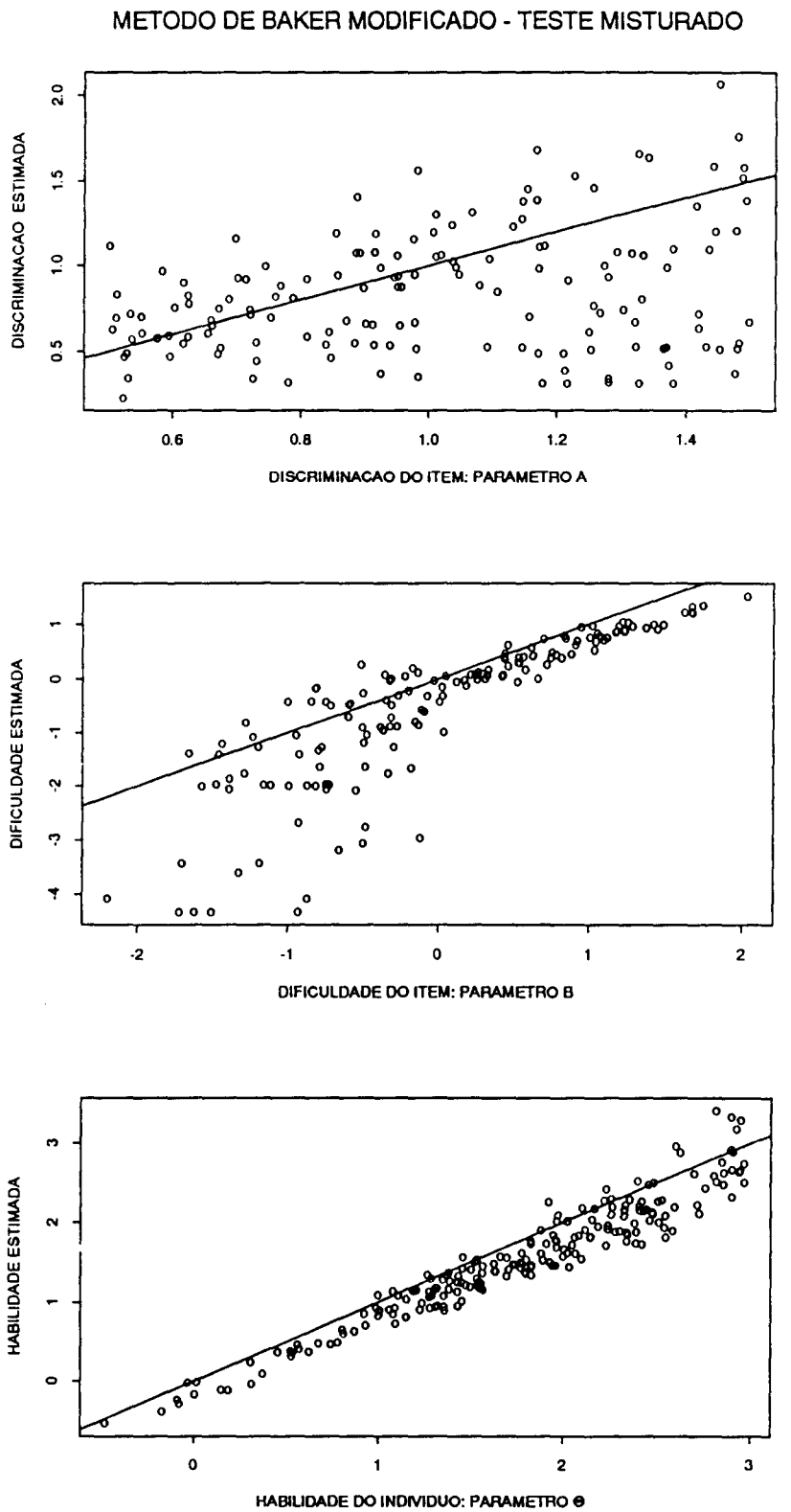


Figura 4.6: Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker modificado, de um teste "misturado" em indivíduos gerados segundo distribuição truncada em $\theta = 3$, segundo a terceira fase de simulações

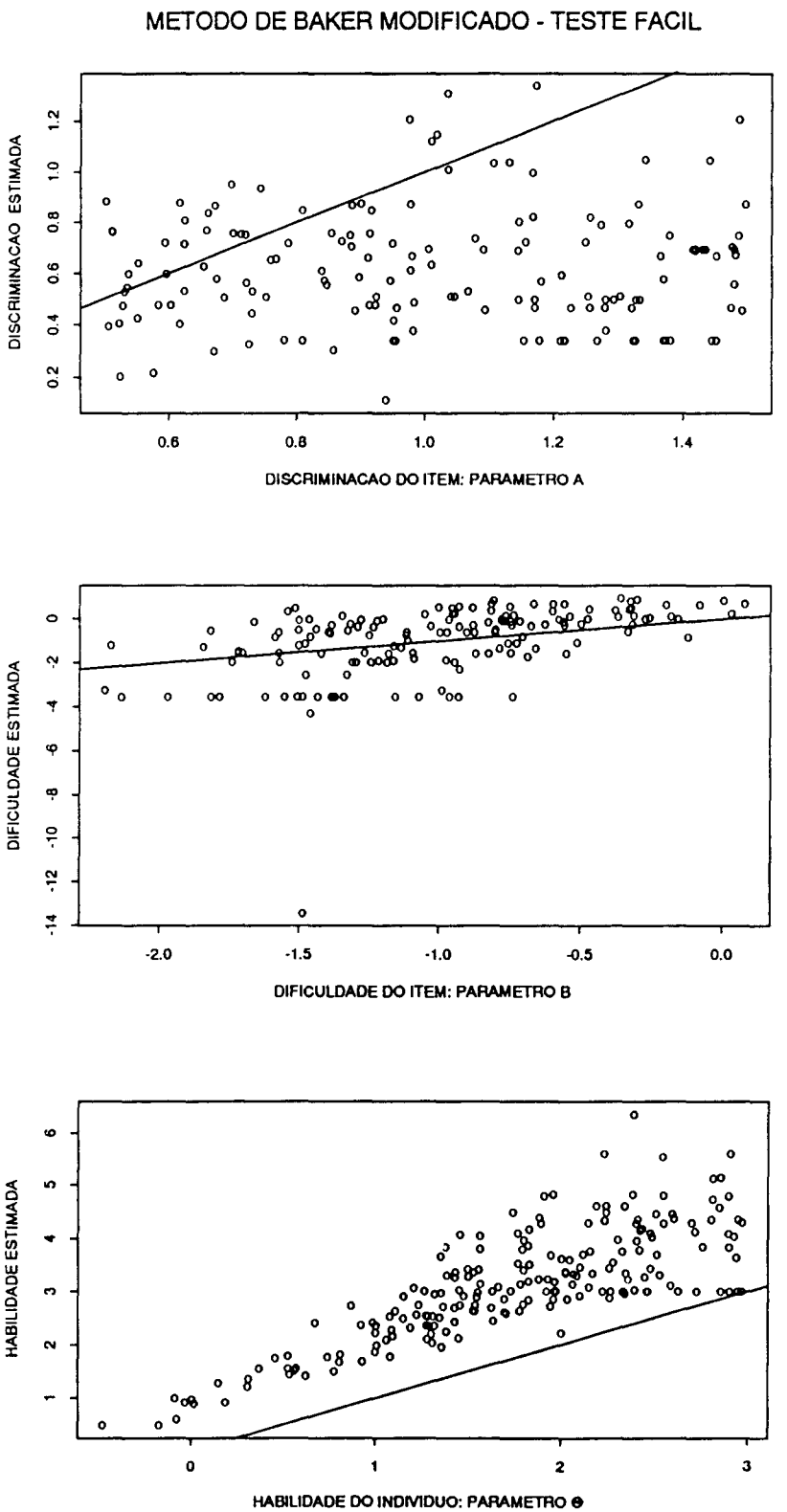


Figura 4.7: Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker modificado, de um teste "fácil" em indivíduos gerados segundo distribuição truncada em $\theta = 3$, segundo a terceira fase de simulações

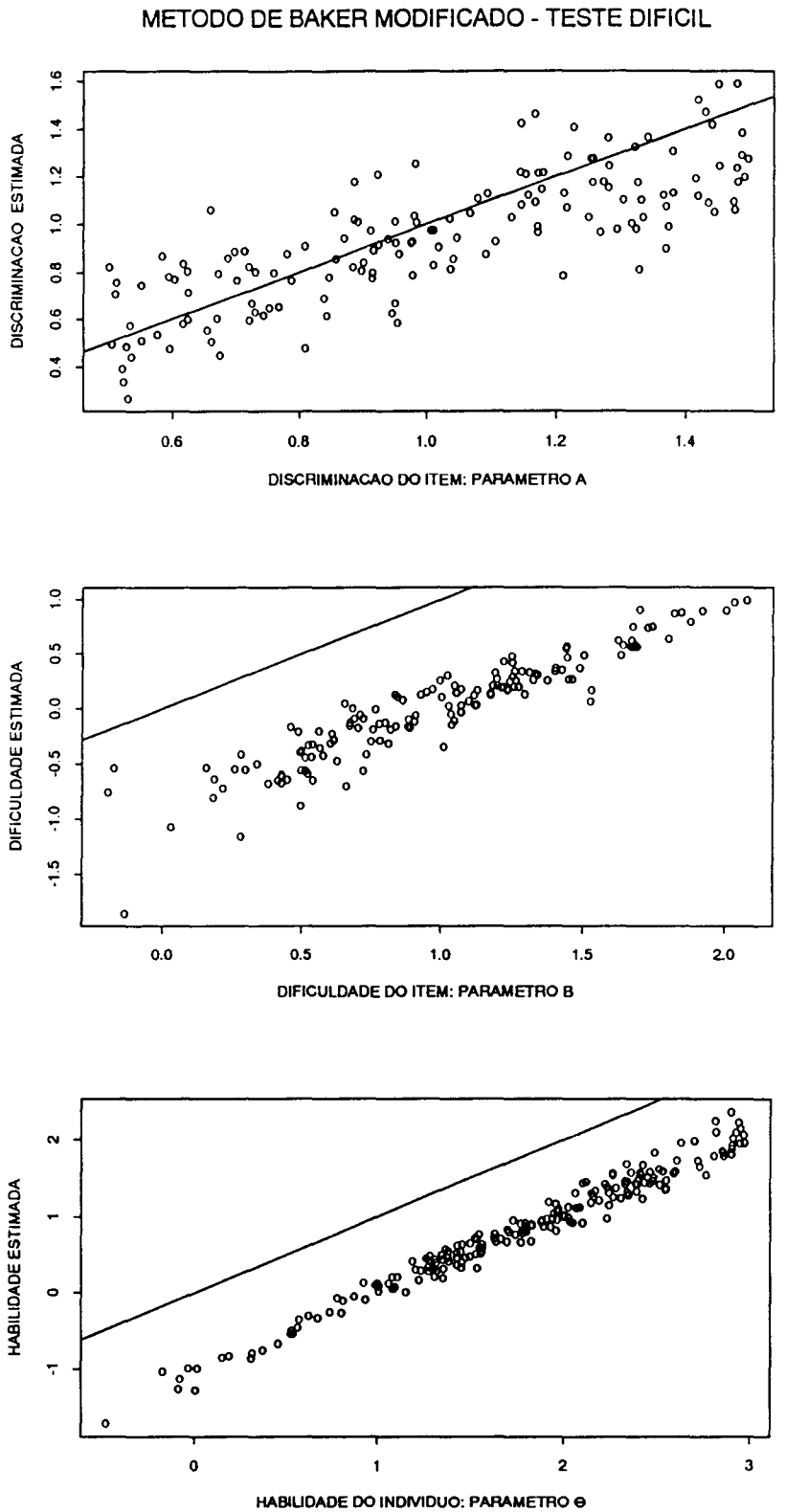


Figura 4.8: Estimativas dos parâmetros a, b e θ obtidas pelo procedimento de Baker modificado, de um teste "difícil" em indivíduos gerados segundo distribuição truncada em $\theta = 3$, segundo a terceira fase de simulações

4.4 Intervalos de Confiança Bootstrap para a Habilidade

O método bootstrap foi introduzido por Efron (1979), e vem sendo amplamente utilizado para diferentes fins. Uma possível utilização seria para a construção de intervalos de confiança. Diferentes tipos de intervalos bootstrap foram discutidos por Efron (1984). Neste trabalho serão abordados somente intervalos de confiança percentis (Efron, 1984). Os métodos para a contrução destes intervalos podem ser divididos em dois tipos básicos o não-paramétrico e o paramétrico.

Esta secção objetiva a contrução de intervalos de confiança para o parâmetro habilidade (θ) segundo métodos bootstrap, com base nos estimadores definidos pelo procedimento de Baker modificado, como uma forma alternativa ao intervalo de confiança assintótico descrito no Capítulo 3.

Dado um teste com n itens aplicado a m indivíduos, o procedimento pode ser resumido da seguinte maneira:

1. selecionam-se com mesma probabilidades m indivíduos, com reposição, da amostra original de tamanho m, obtendo-se uma matriz bootstrap:

$$\mathbf{U}^\star = (\mathbf{U}_\mathbf{1}^\star, \mathbf{U}_\mathbf{2}^\star, \ldots, \mathbf{U}_\mathbf{m}^\star).$$

sendo

$$\mathbf{U}_\mathbf{j}^\star = (U_{1j}^\star, U_{2j}^\star, \ldots, U_{nj}^\star)', \quad j = 1, 2, \ldots, m,$$

e $U_{ij}^\star$ corresponde à resposta, ao i-ésimo item do teste, obtida pelo indivíduo da j-ésima seleção bootstrap.

2. A partir de tal amostra bootstrap são obtidos no primeiro estágio do procedimento de Baker modificado as estimativas bootstrap dos parâmetros de itens,

$$\hat{a}_1^\star, \hat{a}_2^\star, \ldots, \hat{a}_n^\star \quad \text{e } \hat{b}_1^\star, \hat{b}_2^\star, \ldots, \hat{b}_n^\star.$$

3. Utilizando as estimativas bootstrap, $\hat{a}_1^\star, \hat{a}_2^\star, \ldots, \hat{a}_n^\star$ e $\hat{b}_1^\star, \hat{b}_2^\star, \ldots, \hat{b}_n^\star$, obtem-se as estimativas bootstrap das habilidades,

$$\hat{\theta}^\star = (\hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_m^\star)'.$$

4. Os passos 1 a 3 são repetidos k_b vezes.

5. As estimativas bootstrap são ordenadas em ordem crescente, obtendo-se

$$\hat{a}_1^\star(1), \hat{a}_1^\star(2), \ldots, \hat{a}_1^\star(k_b),$$

$$\hat{a}_2^\star(1), \hat{a}_2^\star(2), \ldots, \hat{a}_2^\star(k_b),$$

$$\ldots,$$

$$\hat{a}_n^\star(1), \hat{a}_n^\star(2), \ldots, \hat{a}_n^\star(k_b);$$

$$\hat{b}_1^\star(1), \hat{b}_1^\star(2), \ldots, \hat{b}_1^\star(k_b),$$

$$\hat{b}_2^\star(1), \hat{b}_2^\star(2), \ldots, \hat{b}_2^\star(k_b),$$

$$\ldots,$$

$$\hat{b}_n^\star(1), \hat{b}_n^\star(2), \ldots, \hat{b}_n^\star(k_b);$$

e

$$\hat{\theta}_1^\star(1), \hat{\theta}_1^\star(2), \ldots, \hat{\theta}_1^\star(k_b),$$

$$\hat{\theta}_2^\star(1), \hat{\theta}_2^\star(2), \ldots, \hat{\theta}_2^\star(k_b),$$

$$\ldots,$$

$$\hat{\theta}^{\star}_{m}(1), \hat{\theta}^{\star}_{m}(2), \ldots, \hat{\theta}^{\star}_{m}(k_b).$$

Assim, um intervalo de confiança bootstrap percentil para a habilidade do j-ésimo indivíduo, θ_j, de $100(1-2\alpha)\%$ de confiança, é dado por

$$\left[\hat{\theta}^{\star}(k_b.\alpha)\ ,\ \hat{\theta}^{\star}(k_b(1-\alpha))\right],$$

e os intervalos para os parâmetros de itens são obtidos similarmente.

O procedimento bootstrap paramétrico pode ser resumido nos seguintes passos:

1. Estimam-se os parâmetros a, b e θ pela amostra original através do procedimento de Baker modificado.
2. A matriz bootstrap $\mathbf{U}^{\star}$ é obtida através de simulações do modelo estimado em 1. Através desta matriz bootstrap ($\mathbf{U}^{\star}$) estimam-se os parâmetros a, b e θ pelo procedimento de Baker modificado.
3. Repete-se o passo 2 k_b vezes.
4. Procede-se conforme passo 5 dado anteriormente no procedimento não paramétrico.

Para a construção dos intervalos de confiança bootstrap foram geradas amostras com 200 indivíduos gerados de uma distribuição normal padrão e 60 itens segundo teste com classificação "misturado".

Os intervalos de 90% de confiança ($\alpha = 0,1$), não paramétricos, para os parâmetros a e b, com $k_b = 100$ amostras bootstrap, continham os parâmetros verdadeiros em 82% dos 60 intervalos construídos para os parâmetros de itens. Entretanto, os intervalos de 90%

confiança ($\alpha = 0,1$), não paramétricos, para θ se revelaram bastante pobres, pois uma porcentagem bastante alta dos 200 intervalos não continha o parâmetro verdadeiro (θ).

Intervalos de confiança percentis de $\alpha = 0,1$, paramétricos para θ se revelaram mais apropriados, 85% dos intervalos contruídos continham o parâmetro verdadeiro (para k_b = 100, 200 e 300). Os intervalos bootstrap paramétricos para amostras de tamanho $k_b = 300$ estão ilustrados na Figura 4.4, onde pode-se observar que os intervalos para as habilidades maiores e menores são os de maior amplitude (ou menor precisão). Para os parâmetros a e b os intervalos paramétricos não apresentaram bons resultados, ou seja muitos dos 60 intervalos construídos não continham o parâmetro verdadeiro.

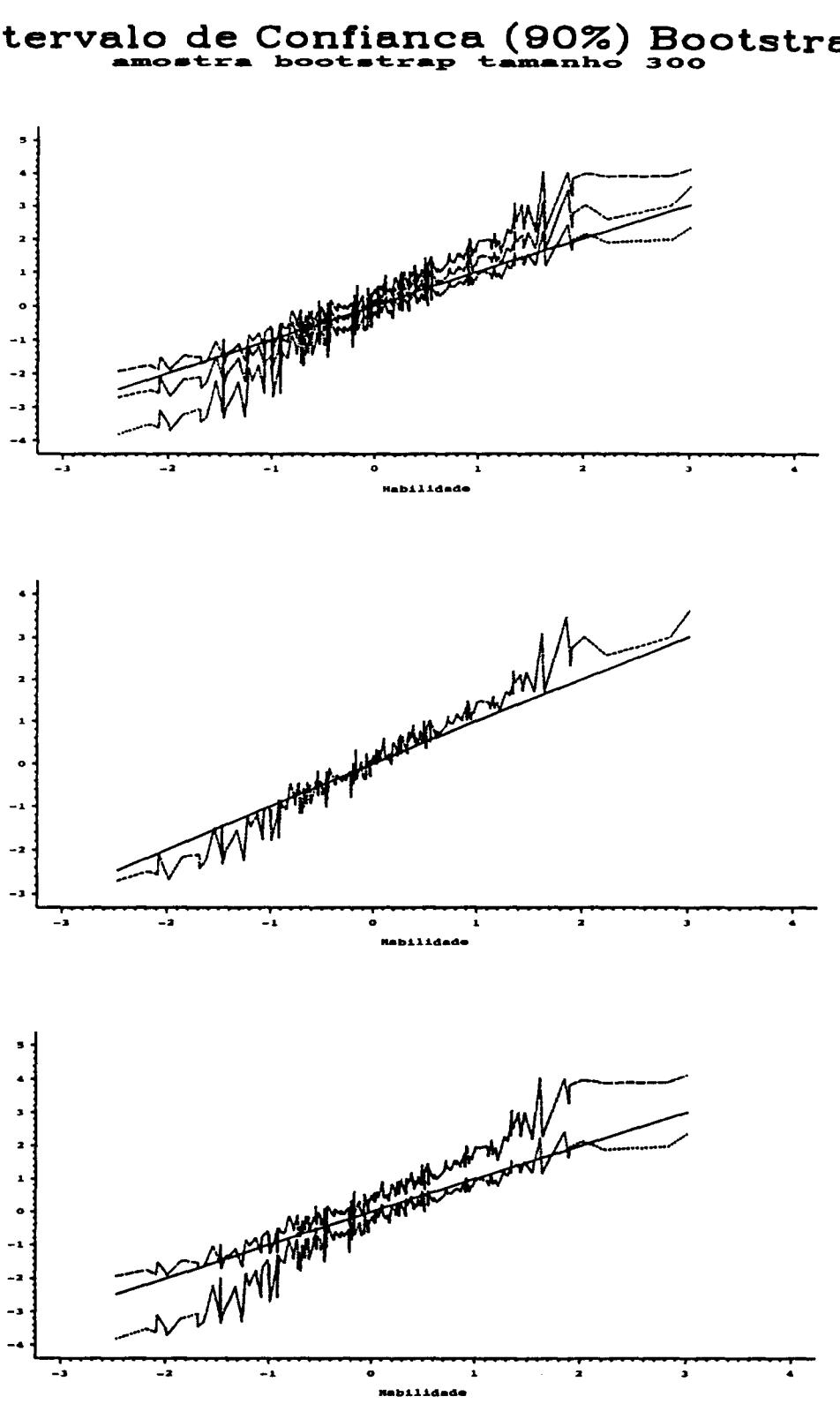


Figura 4.9: Resultados obtidos pelo método bootstrap (paramétrico), 300 amostras bootstrap, a partir de amostra de 200 indivíduos gerados de uma distribuição normal padrão e 60 itens de um teste "misturado"

4.5 Ilustração

Uma aplicação da teoria de resposta de item em uma pesquisa de opinião foi feita por Knott, Albanese e Galbraith (1991). Os dados utilizados fazem parte de uma investigação sobre atitudes sociais britânicas (McGrath e Walerton, 1986) nos anos de 1983–1986, e correspondem a uma pesquisa de 1986 sobre o aborto. Os dados desta pesquisa procuravam revelar em que situações os britânicos aprovariam o aborto. Os sete itens que compunham o questionário eram:

- A própria mulher decide não ter a criança.
- O casal acredita que não deve ter a criança.
- A mulher não é casada e não deseja se casar com o homem.
- O casal não tem recursos para sustentar mais nenhuma criança.
- Existe uma grande chance do bebê ser portador de um defeito congênito.
- A saúde da mulher é seriamente comprometida pela gravidez.
- A gravidez foi causada por um estupro.

As respostas foram categorizadas da seguinte forma: "1" se a pessoa respondeu sim, ou seja se a pessoa concorda com a realização de um aborto dentro da situação descrita pelo item, ou "0" se respondeu não.

Pode-se aplicar a teoria de resposta de item nesta situação, fazendo um paralelo com relação aos parâmetros da função de resposta de item usada em testes educacionais. O acerto ao item seria responder "sim" à questão e o erro seria responder "não". Se por

exemplo, o modelo adotado for o logístico de dois parâmetros (4.1), a habilidade pode ser interpretada como o grau com que o indivíduo é favorável ao aborto. O parâmetro b pode ser interpretado como a dificuldade que se tem em responder "sim" àquela questão; assim, uma questão é "mais difícil" quanto mais forte for a reação contra o aborto nas circunstâncias descritas.

Knott, Albanese e Galbraith utilizaram neste problema o modelo logito-probito descrito por Bartholomew (1987), que corresponde ao modelo logístico de dois parâmetros (4.2), embora considerem a habilidade (θ) como uma variável aleatória. Para obter estimadores para os parâmetros de itens (a e b), utilizam o método de máxima verossimilhança marginal (MVM) apresentado no Capítulo 3, e ao invés de estimarem θ estimam, $E(\theta \mid U)$.

Nesta secção são apresentados os resultados obtidos com a utilização do procedimento de Baker modificado, descrito neste capítulo, para este mesmo conjunto de dados. A Tabela (4.5) apresenta os resultados referentes às estimativas dos parâmetros de itens (a e b) da aplicação de Knott, Albanese e Galbraith, reparametrizados de forma a corresponderem ao modelo logístico apresentado em (4.2), juntamente com os resultados obtidos da aplicação do procedimento de Baker modificado.

As estimativas obtidas pelos dois métodos são diferentes (pela escala diferente nos dois métodos), entretanto uma ordenação decrescente em relação às estimativas dos parâmetros de dificuldade dos itens (b) se revela a mesma, a menos para os dois últimos itens que teriam sua posição invertida.

Como no artigo de Knott, Albanese e Galbraith a habilidade, θ, é considerada como uma variável aleatória, a Tabela (4.5) apresenta a ordenação do padrão de resposta obtida em relação a $E(\theta \mid U)$ e também a ordenação em relação ao parâmetro θ estimado pelo procedimento de Baker modificado. Porém esses padr oes de resposta não divergem muito,

	Método MVM		*Método Baker*	
i	a_i	b_i	a_i	b_i
1	2,335	0,300	1,188	1,448
2	2,435	-0,386	1,401	0,646
3	3,447	-0,557	1,520	0,405
4	1,847	-0,525	1,215	0,554
5	1,629	-3,234	0,922	-0,942
6	4,694	-3,323	0,751	-1,796
7	1,565	-3,585	0,845	-1,436

Tabela 4.5: Estimativas referentes aos parâmetros de itens a e b, obtidas pelos métodos de máxima verossimilhança marginal (MVM) e de Baker

como pode ser observado na Figura 4.5. Pode-se observar ainda que, por exemplo, os padrões de resposta (0000111) e (1111111) são bastantes distintos.

Com relação a tais ordenações pode-se observar uma maior diferença para os sete primeiros tipos de padrão de resposta. Tal fato pode ser creditado à inversão dos dois últimos itens em relação à ordenação segundo a dificuldade do item.

Na Figura 4.5 são apresentados intervalos de confiança bootstrap (paramétrico) para as habilidades dos indivíduos, baseados em amostra bootstrap de tamanho $k_b = 300$. Nesta figura pode-se observar que as amplitudes dos intervalos, para os maiores e menores valores de θ estimados, são as maiores. Ao se observar as estimativas dos parâmetros de itens obtidas, Tabela (4.5), pode-se notar que o "teste" é composto por itens de dificuldade variadas, que foi o tipo de teste no qual se observou as melhores estimativas nas simulações.

padrão de resposta	*frequência das respostas*	*MVM*	*Baker*
0000000	3	1	1
0000001	3	2	3
0000100	1	3	4
0000101	1	4	7
0000010	2	5	2
0000011	6	6	5
0000110	4	7	6
0000111	52	8	8
0100110	1	9	9
0001111	9	10	11
1000111	1	11	10
0100111	6	12	12
0010111	8	13	13
0101111	6	14	15
1100111	3	15	14
0011111	17	16	16
0110111	3	17	17
1101111	2	18	18
1011111	5	19	19
0111111	32	20	21
1110111	12	21	20
1111011	1	22	22
1111111	106	23	23

Tabela 4.6: Frequências observadas dos padrões de resposta às sete questões; ordenação dos padrões de resposta segundo o método usado por Knott et. al. (MVM) e segundo o método de Baker modificado (Baker)

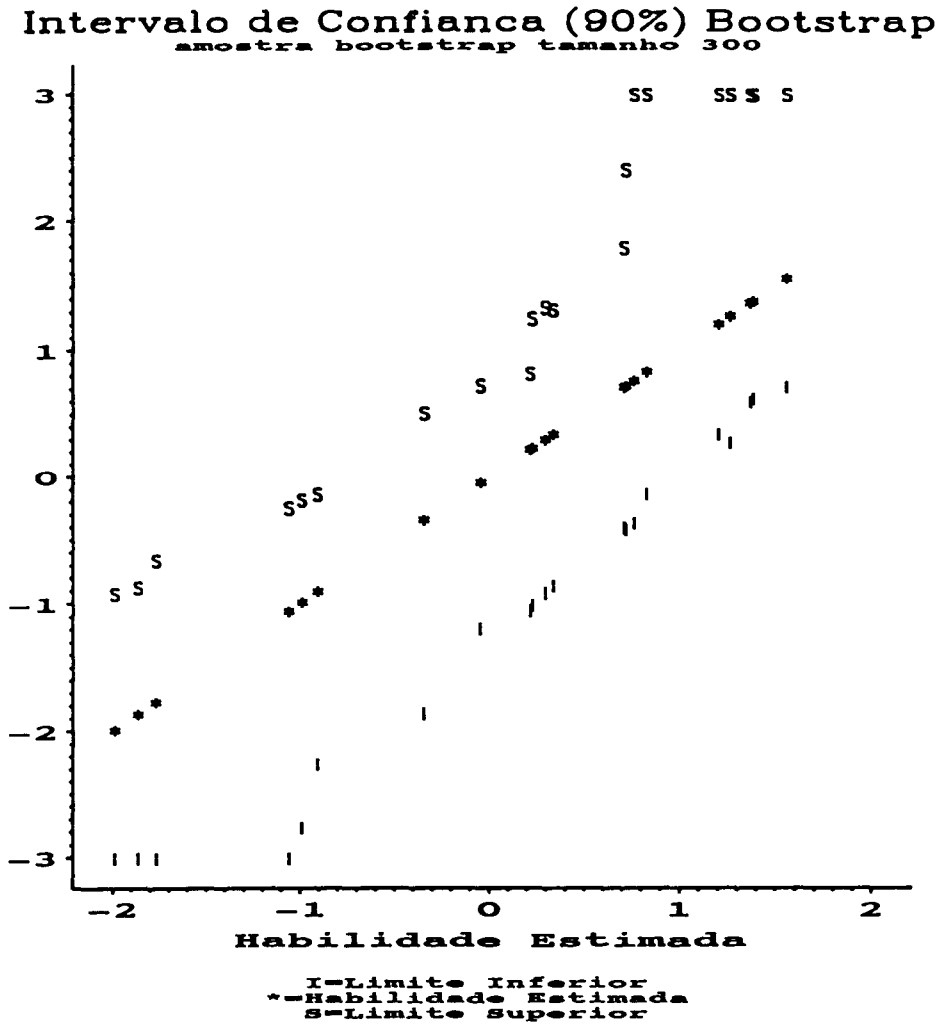


Figura 4.10: Intervalos de confiança para as habilidades dos indivíduos do exemplo apresentado como ilustração (bootstrap paramétrico, $k_b = 300$)

4.6 Conclusão

A modificação do procedimento de Baker onde foi feita a restrição para a construção de grupos com pelo menos cinco indivíduos mostrou ser bastante eficiente na obtenção das estimativas dos parâmetros de itens. Melhores estimativas foram obtidas quando os itens pertenciam à teste com classificação "misturado", o que parece mostrar que testes compostos por uma mistura heterogênea de itens, quando isto é possível, podem levar a melhores estimações de parâmetros de indivíduos e de itens. Nos intervalos de confiança, o bootstrap

paramétrico apresentou melhores resultados para o parâmetro habilidade, uma vez que este tipo de bootstrap leva em consideração informações sobre o modelo.

Apêndice A

MÉTODO DO "SCORING" DE FISHER

O objetivo deste apêndice é fornecer um pequeno resumo sobre o método do "scoring" de Fisher. Este método é utilizado para a obtenção dos estimadores de máxima verossimilhança das habilidades dos indivíduos no segundo estágio do procedimento de Baker (Capítulo 4). Os resultados aqui apresentados podem ser encontrados em Kalbfleisch (1985), e Leite e Singer (1990).

Seja uma função real g unidimensional com domínio D_g. Se g for derivável até a ordem n em alguma vizinhança de $\theta_0 \in D_g$, tem-se pela fórmula de Taylor a seguinte aproximação de g em torno de θ_0,

$$g(\theta) \simeq \sum_{k=0}^{n} \frac{g^{(k)}(\theta_0)}{k!}(\theta - \theta_0)^k, \tag{A.1}$$

para todo $\theta \in D_f$, onde $g^{(k)}(\theta_0)$ é a derivada de ordem k da função g, avaliada no ponto θ_0. Se os termos de maior ordem de (A.1) são pequenos, tem-se a seguinte aproximação,

$$g(\theta) \simeq g(\theta_0) + (\theta - \theta_0)g^{(1)}(\theta_0). \tag{A.2}$$

Deseja-se encontrar uma raiz da função g, ou seja, determinar $\theta^\star$, onde $g(\theta^\star) = 0$. Usando

a aproximação (A.2),

$$g(\theta_0) + (\theta^\star - \theta_0)g^{(1)}(\theta_0) \simeq 0,$$

e, portanto,

(A.3) $$\theta^\star \simeq \theta_0 - \frac{g(\theta_0)}{g^{(1)}(\theta_0)}.$$

Se as duas primeiras derivadas de g existem e g é monótona, o método de Newton, processo iterativo, pode ser utilizado para encontrar a raiz da função g. O método consiste na iteração de (A.3) com θ_0 substituído pelo valor de $\theta^\star$ obtido no passo anterior. Para a primeira iteração é escolhido um valor inicial para θ_0, obtendo-se, portanto, o seguinte processo iterativo,

$$\theta^{(i+1)} = \theta^{(i)} - \frac{g(\theta^{(i)})}{g^{(1)}(\theta^{(i)})},$$

e a raiz da função g é encontrada quando θ_{i+1} e θ_i não difirem muito, sob algum critério definido.

Uma modificação do Método de Newton para o caso de funções de verossimilhança é descrita por Leite e Singer (1990). Sendo $X_1, X_2, \ldots, X_n$ variáveis aleatórias independentes e identicamente distribuídas (iid), com função densidade $f(x, \theta)$ e função de verossimilhança $L(\theta \mid x_1, x_2, \ldots, x_n)$, o estimador de máxima verossimilhança de θ, $\hat{\theta}$, é a solução de

(A.4) $$g(\theta) = \frac{\partial L(\theta \mid x_1, x_2, \ldots, x_n)}{\partial \theta} = 0.$$

Usando o Método de Newton para achar a raiz de $\frac{\partial L(\theta)}{\partial \theta}$, seria necessário encontrar a derivada de segunda ordem da função de verossimilhança, $\frac{\partial}{\partial \theta}[\frac{\partial L(\theta)}{\partial \theta}] = \frac{\partial^2 L}{\partial \theta^2}$. Segundo Leite e Singer, $\frac{1}{n}\frac{\partial^2 L}{\partial \theta^2}$, no ponto θ_0, converge em probabilidade para $-I(\theta_0)$, sob condições de regularidade, quando $n \longrightarrow \infty$, onde $-I(\theta_0)$ é a informação de Fisher no ponto θ_0,

$$I(\theta_0) = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]_{\theta=\theta_0}.$$

Assim, tais autores sugerem, no método de Newton, a utilização da informação de Fisher no lugar da derivada segunda ($\frac{\partial^2 L}{\partial \theta^2}$). Esta modificação no método de Newton é denominada de método do "Scoring" de Fisher. Os autores demostram ainda que a solução do sistema (A.4) obtida por este método é um ponto de máximo da função de verossimilhança.

No segundo estágio do procedimento de Baker, supondo os parâmetros (a e b) do modelo logístico de dois parâmetros conhecidos, a função de probabilidade de $U_{1j}, U_{2j}, \ldots, U_{nj}$ (variáveis aleatórias correspondentes às respostas aos itens de um teste do j-ésimo indivíduo) é apenas função do parâmetro θ_j, e as condições de regularidade, para a utilização do método do "Scoring"de Fisher, são satisfeitas por este modelo (Birnbaum, 1968). A função de verossimilhança de $U_{1j}, U_{2j}, \ldots, U_{nj}$ neste caso é dada por

$$L(\theta_j \mid u_{1j}, u_{2j}, \ldots, u_{nj}) = \prod_{i=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{(1-u_{ij})},$$

onde $u_{ij} = 0$ ou 1. O logarítmo natural da função de verossimilhança é dado por

$$lnL = ln[L(\theta_j \mid u_{1j}, u_{2j}, \ldots, u_{nj})] = \sum_{i=1}^{n} [u_{ij} P_{ij} + (1 - u_{ij}) Q_{ij}].$$

A informação de Fisher de $U_{1j}, U_{2j}, \ldots, U_{nj}$ para θ_j é dada por

$$\begin{aligned} E\left[\left(\frac{\partial \ln L}{\partial \theta_j}\right)^2\right] &= E\left\{\left[\sum_{i=1}^{n}(U_{ij} - P_{ij})\frac{P'_{ij}}{P_{ij}Q_{ij}}\right]^2\right\} \\ &= E\left\{\left[\sum_{i=1}^{n}(U_{ij} - P_{ij})\frac{P'_{ij}}{P_{ij}Q_{ij}}\right]\left[\sum_{i=1}^{n}(U_{ij} - P_{ij})\frac{P'_{ij}}{P_iQ_i}\right]\right\} \\ &= \sum_{i=1}^{n}\sum_{l=1}^{n}\frac{P'_{ij}P'_{lj}}{P_{ij}P_{lj}Q_{ij}Q_{lj}}E[(U_{ij} - P_{ij})(U_{lj} - P_{lj})]. \end{aligned}$$

Sob a suposição de independência entre os itens de um teste, e usando $E(U_{ij} \mid \theta_j) = P_{ij}$ e $Var(U_{ij} \mid \theta_j) = P_{ij}Q_{ij}$, por (3.1), tem-se

$$E\left[\left(\frac{\partial \ln L}{\partial \theta_j}\right)^2\right] = \sum_{i=1}^{n}\frac{P'^2_{ij}}{P^2_{ij}Q^2_{ij}}Var(U_{ij} \mid \theta_j) = \sum_{i=1}^{n}\frac{P'^2_{ij}}{P^2_{ij}Q^2_{ij}}P_{ij}Q_{ij} = \sum_{i=1}^{n}\frac{P'^2_{ij}}{P_{ij}Q_{ij}},$$

onde $P_{ij} = 1/\{1 + \exp[-1,7a_i(\theta_j - b_i)]\}$.

Explicitando as expressões P_{ij}, Q_{ij} e P'_{ij}, a informação de Fisher de $U_{1j}, U_{2j}, \ldots, U_{nj}$ no ponto θ_j é dada por

$$\begin{aligned} I(\theta_j) &= \sum_{i=1}^{n} \frac{P'^2_{ij}}{P_{ij}Q_{ij}} \\ &= \sum_{i=1}^{n} \frac{1,7^2 a_i^2 \exp[1,7a_i(\theta_j - b_i)]\{\exp[-1,7a_i(\theta_j - b_i)] + 2 + \exp[1,7a_i(\theta_j - b_i)]\}}{\exp(1,7a_i(\theta_j - b_i))(\exp(-1,7a_i(\theta_j - b_i)) + 2 + \exp(1,7a_i(\theta_j - b_i)))^2} \\ &= \sum_{i=1}^{n} \frac{1,7^2 a_i^2}{\exp[-1,7a_i(\theta_j - b_i)] + 2 + \exp[1,7a_i(\theta_j - b_i)]}. \end{aligned}$$

O processo iterativo para estimação de θ_j (j=1,2,...,N), obtido pelo método do "Scoring" de Fisher, é dado por

$$\begin{aligned} \hat{\theta}_j^{(m)} &= \hat{\theta}_j^{(m-1)} + \sum_{i=1}^{n} \frac{\{\exp[-1,7a_i(\theta_j^{(m-1)} - b_i)] + 2 + \exp[1,7a_i(\theta_j^{(m-1)} - b_i)]\}}{1,7^2 a_i^2} \\ &\quad \times (u_{ij} - P_{ij}) 1,7a_i \\ &= \hat{\theta}_j^{(m-1)} + \sum_{i=1}^{n} \frac{\{\exp[-1,7a_i(\theta_j^{(m-1)} - b_i)] + 2 + \exp[1,7a_i(\theta_j^{(m-1)} - b_i)]\}}{1,7a_i} \\ &\quad \times (u_{ij} - P_{ij}), \end{aligned}$$

onde $\hat{\theta}_j^{(m)}$ é a estimativa de θ na j-ésima iteração.

Apêndice B

PROGRAMA PARA ESTIMAÇÃO DE PARÂMETROS DA FUNÇÃO DE RESPOSTA DE ITEM

```
/* ------------------------------------------------
 * Este programa calcula as estimativas dos parametros
 * das funcoes de resposta de item de  um  teste pelo
 * METODO DE BAKER (Baker,1987)  modificado.
 * O usuario deve fornecer uma matriz R  formada pelas
 * respostas dos individuos aos  itens, sendo  que  as
 * linhas  representam as respostas  de cada individuo
 * aos itens, com virgulas separando as respostas  dos
 * varios individuos. No  exemplo tem-se as  respostas
 * de 23 individuos a 7 itens.
 * As  estimativas  dos  parametros  de  itens  e das
 * habilidades sao apresentadas ao final  do  programa
```

```
 * nos seguintes vetores
 * AE:  vetor de estimativas dos parametros de discri-
 *      minacao dos itens,
 * BE:  vetor  de  estimativas dos parametros de difi-
 *      culdade dos itens,
 * OE:  vetor  de  estimativas  das  habilidades  dos
 *      dos individuos; lembrando as limitacoes  deste
 *      metodo: nao sao estimadas  as  habilidades  de
 *      individuos que acertaram todos ou nenhum item;
 *      valores   99   e  -99   sao   impressos   para
 *      indicar tais eventos, repectivamente.
   ------------------------------------------------*/
OPTIONS LS=72 PS=60;

PROC IML;
R={0 0 0 0 0 0 0,
   0 0 0 0 0 0 1,
   0 0 0 0 1 0 0,
   0 0 0 0 1 0 1,
   0 0 0 0 0 1 0,
   0 0 0 0 0 1 1,
   0 0 0 0 1 1 0,
   0 0 0 0 1 1 1,
   0 1 0 0 1 1 0,
   0 0 0 1 1 1 1,
   1 0 0 0 1 1 1,
```

```
   0 1 0 0 1 1 1,
   0 0 1 0 1 1 1,
   0 1 0 1 1 1 1,
   1 1 0 0 1 1 1,
   0 0 1 1 1 1 1,
   0 1 1 0 1 1 1,
   1 1 0 1 1 1 1,
   1 0 1 1 1 1 1,
   0 1 1 1 1 1 1,
   1 1 1 0 1 1 1,
   1 1 1 1 0 1 1,
   1 1 1 1 1 1 1};

/* M E O NUMERO DE EXAMINANDOS */
M=NROW(R);

/* N E O NUMERO DE ITENS */
N=NCOL(R);
PRINT M N;

RESET STORAGE="DADOST";
STORE R;

OE=R[,+];
RRC=RANK(OE);
RM=R;
```

```
/* RM E A MATRIZ R ORDENADA PELO VETOR RCS */
DO I=1 TO M;
  DO J=1 TO N;
  RM[RRC[I],J]=R[I,J];
  END;
END;
FREE R;

RCS=OE;
/* RCS E O VETOR OE ORDENADO*/
DO I=1 TO M;
   RCS[RRC[I]]=OE[I];
END;

/* OE E O VETOR QUE CONTEM O VALOR INICIAL PARA A
   HABILIDADE DE CADA EXAMINANDO */
DO I=1 TO M;
   IF RCS[I]^=N & RCS[I]^=0 THEN OE[I]=PROBIT(RCS[I]/N);
                                ELSE IF RCS[I]=N THEN OE[I]=3;
                                                 ELSE OE[I]=-3;
  END;
FREE RRC;
FREE RCS;

K=11;
```

```
IF N>M THEN K=5;

F=SHAPE(0,K,1);
A=1;
TI=OE[M]-OE[1];

/* DIVISAO DOS K GRUPOS DE EXAMINANDOS, RK E A MATRIZ
   QUE CONTEM A SOMA DE RIJ PARA CADA GRUPO, F E O
   VETOR QUE CONTEM A FREQUENCIA DE  CADA GRUPO */
DO I=1 TO M;
DO WHILE(OE[I]>OE[1]+A*(TI/K));
                              A=A+1;
                              END;
IF A>K THEN A=K;
F[A]=F[A]+1;
END;

FAUX=SHAPE(0,K,1);
A=1;
DO I=1 TO K;
IF F[I]^=0 THEN DO;
                FAUX[A]=F[I];
                A=A+1;
                END;
END;
```

```
A=A-1;
F=SHAPE(FAUX,A,1);

/* CONSTRUCAO DE F DE MODO QUE A FREQUENCIA SEJA NO
   MINIMO 5 */

AUX=0;
FAUX=SHAPE(0,A,1);
I=0;
DO WHILE(I<A);
AUX=AUX+1;
  DO WHILE(FAUX[AUX]<5 & I<A);
    I=I+1;
    FAUX[AUX]=FAUX[AUX]+F[I];
  END;
END;
IF FAUX[AUX]=0 THEN AUX=AUX-1;
FREE F;
F=SHAPE(FAUX,AUX,1);
FREE FAUX;
IF F[AUX]<5 THEN DO;
   F[AUX-1]=F[AUX-1]+F[AUX];
   AUX=AUX-1;
   FAUX=SHAPE(F,AUX,1);
   FREE F;
   F=FAUX;
```

```
   END;
K=NROW(F);

RK=SHAPE(0,K,N);
A=1;
AUX=F[A];
DO I=1 TO M;
   IF I<=AUX THEN DO;
                  DO J=1 TO N;
                  RK[A,J] =RK[A,J]+RM[I,J];
                  END;
                  END;
             ELSE DO;
                  A=A+1;
                  AUX=AUX+F[A];
                  END;
END;
FREE RM;
FREE TI;

/* OM E O VETOR QUE CONTEM O VALOR INICIAL PARA A
   HABILIDADE DE CADA GRUPO DE EXAMINANDO */
OM=SHAPE(0,K,1);
AUX2=1;
AUX=0;
DO I=1 TO K;
```

```
AUX=AUX+F[I];
   DO J=AUX2 TO AUX;
      OM[I]=OM[I]+OE[J];
   END;
AUX2=AUX+1;
OM[I]=OM[I]/F[I];
END;
FREE OE;
L=SHAPE(0,K,N);
W=SHAPE(0,K,N);

/* CALCULOS PRELIMINARES QUE AJUDARAO NAS
   ESTIMATIVAS */
DO J=1 TO N;
   DO I=1 TO K;
   PKIJ=RK[I,J]/F[I];
   IF PKIJ=1 THEN PKIJ=1-(1/(2*F[I]));
                ELSE IF PKIJ=0 THEN PKIJ=1/(2* F[I]);
   QKIJ=1-PKIJ;
   W[I,J]=PKIJ*QKIJ*F[I];
   L[I,J]=LOG(PKIJ/QKIJ);
   END;
END;
SE=SHAPE(0,1,N);
LAE=SHAPE(0,1,N);
FREE RK;
```

```
/* CALCULO DAS ESTIMATIVAS, KA E O NUMERO DE ITERACOES,
   SE E O VETOR DE PSI ESTIMADO, LAE E O VETOR DE
   LAMBDA ESTIMADO, OM E A HABILIDADE MEDIA ESTIMADA
   DE CADA GRUPO */
EST=SHAPE(0,2,N);
VEST=SHAPE(0,2,2);
DVEST=0.10;
DEST=0.10;
DO KA=1 TO 40
  WHILE(DVEST>.001 & DEST>.001);
  DO J=1 TO N;
  VESTA=VEST;
  ESTA=EST;
  ID=SHAPE(1,K,1);
  X=ID||OM;
  FREE ID;
  V=DIAG(W[,J]);
  VEST=INV(X'*V*X);
  EST[,J]=VEST*X'*V*L[,J];
  DEST=MAX(ABS(EST-ESTA));
  DVEST=MAX(ABS(VEST-VESTA));
  END;
FREE X; FREE V;
  DO I=1 TO K;
  Y= L[I,] - EST[1,];
```

```
  V=DIAG(W[I,]);
  X=EST[2,];
  OM[I]=INV(X*V*X')*X*V*Y';
  END;
FREE X; FREE V; FREE Y;
END;
FREE F;

FREE W; FREE V; FREE VEST; FREE L; FREE Y; FREE X;
FREE DEST; FREE DVEST; FREE VESTA; FREE ESTA; FREE KA;

/* REPARAMETRIZANDO */
AE=EST[2,]*1/1.7;
BE=SHAPE(0,1,N);
OE=SHAPE(0,M,1);
DO I=1 TO N;
BE[I]=-EST[1,I]/EST[2,I];
END;
BE=BE'; AE=AE';

/* IMPRIMINDO O VALOR ESTIMADO DE B E A */
PRINT , "Estimativas dos par. de dificuldade e discriminacao",
         BE AE;

BE=BE'; AE= AE';
STORE EST;
```

```
FREE EST;
LOAD R;

OE=R[,+];
DO I=1 TO M;
   IF OE[I]^=N & OE[I]^=0 THEN OE[I]=PROBIT(OE[I]/N);
                              ELSE IF OE[I]=N THEN OE[I]=99;
                                                  ELSE OE[I]=-99;
  END;

FREE RCS;
S=SHAPE(0,M,1);
F=0.10;
INF=SHAPE(0,M,1);
DO INTER=1 TO 40
   WHILE(F>.001);
OEA=OE;
DO I=1 TO M;
   IF OE[I]^=-99 & OE[I]^=99 THEN DO;
   IF ABS(OE[I])<.000001 THEN OE[I]=0;
   DO J=1 TO N;
   EXPN=EXP(-1.7*AE[J]*(OE[I]-BE[J]));
   EXPP=EXP(1.7*AE[J]*(OE[I]-BE[J]));
   S[I]=S[I]+((R[I,J]-(1/(1+EXPN)))*1.7*AE[J]);
   INF[I]=INF[I]+ ((1.7*1.7*AE[J]*AE[J])/(EXPP+2+EXPN));
   END;
```

```
   OE[I]=OE[I]+S[I]/INF[I];
   END;
END;
S=SHAPE(0,M,1);
INF=SHAPE(0,M,1);
F=MAX(ABS(OE-OEA));
END;

/* IMPRIMINDO O VALOR ESTIMADO DA HABILIDADE*/
PRINT , "Estimativas das habilidades", OE;

FREE OE; FREE EXPP; FREE EXPN; FREE S; FREE INF; FREE F;
FREE INTER;
```

Bibliografia

[1] Baker, F. B. (1987). Item Parameter Estimation via Minimum Logit Chi-Square. Journal of Mathematical and Statistical Psycology, 40, 50–60

[2] Bartholomew, D. J. (1987). Latent Variable Models and Factors Analysis. Charles Griffin e Company Ltd., London

[3] Berkson, J. (1944). Application of Logistic Function to Bio-Assay. Journal of the American Statistical Association, 39, 357–365

[4] Birnbaum, A. (1968). Parte V. Em Lord F. M. e Novick (Eds.) Statistical Theories of Mental Test Scores. Reading, Addison Wesley

[5] Bishop, Y. M. M., Fienberg, S. E. E Holland, P. W. (1975. Multivariate Analysis: Theory and Practice. The MIT Press, Cambridge, Massachusetts

[6] Bock, R. D. e Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: Application an EM Algorithm. Psychometrika, 46, 443–459

[7] Bock, R. D. e Lieberman, M. (1970). Fitting a Response Model for n Dichotomously Scored Items. Psychometrika, 35, 179–197

[8] Brock, R. J. e Arnold, G. C. (1985). Applied Regression Analysis and Experimental Design. Marcel Dekker, New York

[9] Carmines, E. G. e Zeller, R. A. (1981). Reliability and Validity Assessment. Sage, New York

[10] Cramér, H. (1946). Mathematical Methods of Statistics. University Press. Princeton, N. J.

[11] Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. Psychometrika, 16, 297–334

[12] Dempster, A. P., Laird, N.M. e Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society (Série B), 39, 1–38

[13] Draper, N. R. e Smith, H. (1981). Applied Regression Analysis. John Wiley, New York

[14] Efron, B. (1979). Bootstrap Methods: another Look at the Jackknife. Annals of Statistics, 7, 1– 26

[15] Efron, B. e Gong, G.(1983). A Leisurely Look at the Bootstrap, the Jackknife and Cross-Validation. The American Estatistician, 37, 36–48

[16] Gulliksen, H. (1950). Theory of Mental Tests. Wiley, New York

[17] Kalbfleisch, J. G. (1985). Probability and Statistical Inference. Springer, New York

[18] Kale, B. K. (1962). On the Solution of Likelihood Equations by Iteration Processes. The Multiparametric Case. Biometrika, 49, 479–486

[19] Knott, M., Albanese, M. T. e Galbraith, J. (1991). Scoring Attitudes to Abortion. The Statistician, 40, 217–223

[20] Kuder, G. F. e Richardson, M. W. (1937). The Theory of Estimation of Test Reliability. Psychometrika, 2, 151–160

[21] Leite, J. G. e Singer, J. M. (1990). Métodos Assintóticos em Estatística. Apostila do IX SINAPE, IME/USP, S.P.

[22] Lindley, D.V. e Smith, A. F. (1972). Bayesian Estimates for Linear Model. Journal of the Royal Statistical Society, (Série B) 24, 1–41

[23] Lord, F. M. (1953). An Application of Confidence Intervals and Maximum Likelihood to the Estimation of an Examinee's Ability. Psychometrika, 18, 57–76

[24] Lord, F. M. (1980). Applications of Item Response Theory to Pratical Testing Problems. Lawrence, Hillsdale

[25] Lord, F. M. e Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Addison, Reading

[26] Mislevy, R. J. e Stoking, M. L. (1987). A Consumer's Guide to LOGIST and BILOG. Educational Testing Service, Princeton, New Jersey

[27] McGrath, K. e Waterton, J. (1986). Social Attitudes, 1983–1986 Panel Survey. Technical Report. London, SCPR

[28] Novick, M. R. e Lewis, C. (1967) Coefficient Alpha and the Reliability of Composite Measurements. Psychometrika, 31, 1–13

[29] Rulon, P. J. (1939) A Simplified Procedure for Determining the Reliability of a Test by Split-Halves. Harvard Educational Review, 9, 99–103

[30] Rosenberg, M. (1965). Society and the Adolescent Self Image. Princeton University Press, Princeton, N. J.

[31] Samejima F. A. (1973). A Comment on Birnbaum's Three-Parameter Model in Latent Trait Theory. Psychometrika, 38, 221–233

[32] Sanathanan, L. e Blumenthal, S. (1978). The Logistic Model and Estimation of Latent Structure. Journal of American Statistical Association, 73, 794–799

[33] Sijtsma, K. e Molenaar, I. W. (1987). Reliability of Test Scores in Nonparametric Item Response Theory. Psychometrika, 52, 79–97

[34] Stocking, M. L. e Pearlman, M. A. (1989). Item Response Theory for Test Developers. Educational Testing Service. Princeton, New Jersey

[35] Swaminathan, H. e Gifford, J. A. (1986). Bayesian Estimation in Three-Parameter Logistic Model. Psychometrika, 51, 589–601

[36] Tuker, L. R. (1951). Academic Ability Test. Research Memorandum. Educational Testing Service. Princeton, New Jersey

[37] Winsgersky, M. S. e Lord, F. M. (1973). A Computer Program for Estimating Examinee Ability and Item Characteristic Curve Parameters. (RM-76-6) [computer program]. Educational Testing Service, Princeton, New Jersey

[38] Wright, B. D. (1977). Solving Measurement Problems with the Rash Model. Journal of Educational Measurement, 14, 97–116

[39] Vianna, H. M. (1982). Testes em Educação. IBRASA, São Paulo

[40] Yen, W. M. (1987). A Comparation of the Efficiency an Accuracy of BILOG and LOGIST. Psychometrika, 52, 275–291

[41] Yen, W. M. (1984). Obtaining Maximum Likelihood Trait Estimates from Number-Correct Scores for the Three-Parameter Logistic Model. Journal of Educational Measurement, 21, 93–111