MARIA DE LOURDES LIMA BRAGION

ANÁLISE COMBINADA DE EXAMES VESTIBULARES DA UNIVERSIDADE FEDERAL DE LAVRAS USANDO A TEORIA DE RESPOSTA AO ITEM

LAVRAS-MG

2011

MARIA DE LOURDES LIMA BRAGION

ANÁLISE COMBINADA DE EXAMES VESTIBULARES DA UNIVERSIDADE FEDERAL DE LAVRAS USANDO A TEORIA DE RESPOSTA AO ITEM

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador
Dr. Júlio Sílvio de Sousa Bueno Filho

LAVRAS-MG
2010

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Bragion, Maria de Lourdes Lima.
Análise combinada de exames vestibulares da Universidade Federal de Lavras usando a teoria de resposta ao item. – Lavras : UFLA, 2010.
187 p. : il.

Tese (Doutorado) - Universidade Federal de Lavras, 2010.
Orientador: Júlio de Sílvio de Sousa Bueno Filho.
Bibliografia.

1. Inferência bayesiana. 2. Curva característica do item. 3. Função informação do item. 4. Monte Carlo via cadeia de Markov. 5. Simulação. I. Universidade Federal de Lavras. II. Título.

CDD-519.542

MARIA DE LOURDES LIMA BRAGION

ANÁLISE COMBINADA DE EXAMES VESTIBULARES DA UNIVERSIDADE FEDERAL DE LAVRAS USANDO A TEORIA DE RESPOSTA AO ITEM

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 17 de dezembro de 2010.

Dr. Caio Lucidius Naberezny Azevedo	UNICAMP
Dr[a]. Thelma Sáfadi	UFLA
Dr. Daniel Furtado Ferreira	UFLA
Dr. Ulisses Azevedo Leitão	UFLA

Dr. Júlio Sílvio de Sousa Bueno Filho
Orientador

LAVRAS-MG
2010

Dedico este trabalho ao
meu marido,
meus filhos
e minha mãe.

AGRADECIMENTOS

A Deus, por ter me concedido saúde, força e ter me confortado através das muitas promessas em Sua Palavra.

Ao meu marido Nivaldo, por seu amor, sua dedicação e por ter me animado nas horas de angústia. Querido, sem seu apoio seria muito difícil chegar até aqui. Te amo!

Aos meus filhos Daniel e Mariana, pela compreensão quando tive que repartir o tempo devido a vocês para dar conta de meus estudos. Daniel, meu filho, muito obrigada por sempre me ajudar na parte computacional quando precisei.

Agradeço aos meus pais, porque foi por meio deles que um dia comecei a percorrer a jornada da vida e pude chegar onde cheguei.

Ao meu orientador, Júlio Sílvio de Sousa Bueno Filho, por ter dividido comigo um pouco de seu muito conhecimento e me ajudar a chegar onde cheguei.

Aos ensinamentos que todos os professores me proporcionaram. A todos os funcionários do Departamento de Ciências Exatas em especial a Josi por sua eficiência e atenção.

Ao Fábio que muito me ajudou computacionalmente. Você não tem ideia de quanto lhe sou grata por todo o tempo que gastou comigo na elaboração de gráficos e outros aspectos computacionais.

Aos colegas de estudo Patrícia, Luciene, Ademária, Renata e todos os que comigo partilharam conhecimentos.

A CAPES, por me conceder uma bolsa durante todo esse período, o que tornou possível minha dedicação a este trabalho.

A COPESE pelo fornecimento dos dados que precisei para a realização desse trabalho, em especial ao Prof. Marcelo Oliveira que tão prontamente sempre me atendeu.

RESUMO

Os exames vestibulares tornaram-se forma padrão de admissão na Universidade Brasileira a partir dos anos 70. A análise de características das provas e das habilidades por meio da Teoria de Resposta ao Item (TRI) possui a vantagem de poder ser utilizada para comparar habilidades de candidatos e parâmetros referentes aos itens, mesmo em situações em que os candidatos não se submetem às mesmas provas. O emprego da inferência *bayesiana* resulta em estimativas com propriedades interessantes para fins de combinar posteriormente as análises resultantes, mas apesar da existência de *softwares* para a análise de TRI, poucos são diretamente adequados à análise *bayesiana* de ensaios desbalanceados no modelo logístico de três parâmetros (ML3P). O presente trabalho teve por objetivos implementar um método *bayesiano* para analisar conjuntamente todos os cursos, incluindo todos os itens, dos vestibulares 2006-2 a 2009-1 da Universidade Federal de Lavras (UFLA), considerando o ML3P. Foi desenvolvida uma função R escrita em C que agiliza a amostragem Monte Carlo via cadeia de Markov (MCMC) da distribuição a *posteriori* conjunta. O primeiro capítulo traz uma revisão geral sobre a TRI e os procedimentos *bayesianos*. No segundo capítulo, o desempenho da função desenvolvida foi avaliado por meio de um estudo de simulação. Tal estudo indicou que o modelo e o método de estimação produziram resultados bastante satisfatórios, pois foi encontrada alta correlação entre os valores paramétricos e os estimados. Utilizou-se essa função para análises em separado dos vestibulares 2006-2 a 2009-1 da UFLA. Estas análises, apresentadas no capítulo seguinte, revelam que as provas mais difíceis são mais discriminativas e informativas. A análise ao longo do tempo da evolução das propriedades das provas e habilidades dos candidatos foi feita no último capítulo com a análise combinada de todos os vestibulares, baseada nas estimativas (cadeias de Markov) anteriormente obtidas. Pôde-se verificar que, quanto aos parâmetros dos itens, as disciplinas da área de exatas - Física, Matemática e Química, sempre estão classificadas como as mais difíceis e as que mais discriminam. Todas as provas dos vestibulares estudados tiveram, em média, baixa probabilidade de acerto por indivíduos com baixa habilidade. Quanto às habilidades dos candidatos, a média da habilidade para os cursos noturnos foi menor que a dos cursos diurnos. São apresentados exemplos que evidenciam que a metodologia pode revelar importantes propriedades para o planejamento de provas bem como para auxiliar a seleção de candidatos com base em suas habilidades.

Palavras-chave: Inferência *bayesiana*. Curva característica do item. Função informação do item. Monte Carlo via cadeia de Markov. Simulação.

ABSTRACT

The entrance examinations (*vestibular*) have become the standard form for admission to the University of Brazil, from 70 years. The analysis of features of tests and abilities through Item Response Theory (IRT) has the advantage that it can be used to compare the abilities of candidates and parameters referring to the items even in situations where candidates don't submit the same tests. The use of Bayesian inference results in estimates with interesting properties for the purpose of combining the resulting analysis later, but despite the existence of software for the IRT analysis, few are directly suitable for Bayesian analysis of unbalanced trials in three parameters logistic model (3PLM). This study had for objective to implement a Bayesian framework to analyse together all the courses, including all items, of *vestibular* 2006-2 to 2009-1 of Federal University of Lavras (UFLA), considering the 3PLM. A function R was developed writing in C that streamlines the sampling Markov Chain Monte Carlo (MCMC) the joint posterior distribution. The first chapter brings a general revision about IRT and Bayesian methods. In the second chapter, the performance of developed function was evaluated through a simulation study. This study indicated that the model and estimation produced satisfactory results because it was found a high correlation between the parametric values and those estimated. It was used this function to a separate analysis of *vestibular* 2006-2 to 2009-1 of UFLA. These analysis presented in the following chapter, reveal that the most difficult tests are more discriminating and more informative. The analysis over time of changing proprieties of the tests and abilities of candidates was done in the last chapter with the combined analysis of all *vestibular*, based on the estimates (Markov chains) previously obtained. It was verified that, for the parameters of the items, the disciplines of exact area - Physics, Mathematics and Chemistry are always classified as the most difficult and more discriminating. All tests of *vestibular* studied had on average, low probability by guess. Regarding the abilities of the candidates, the average ability for the evening courses was lower than that daytime courses. Examples are presented that indicate that this methodology can reveal important proprieties of evidence for planning of tests as well to help select candidates based on their abilities.

Keywords: Bayesian inference. Item Characteristic curve. Function item information. Markov Chain Monte Carlo. Simulation.

LISTA DE FIGURAS

Capítulo 1 18
Figura 1 Curva característica do item - CCI 28
Figura 2 CCIs com diferentes valores para *b* (parâmetro de dificuldade do item) e *a* = 1 (parâmetro de discriminação do item) 33
Figura 3 CCIs com diferentes valores para *a* (parâmetro de discriminação do item) e *b* = 0 (parâmetro de dificuldade do item) 34
Figura 4 CCIs com diferentes valores para *c* (parâmetro da probabilidade de acerto casual), *a* = 1 (parâmetro de discriminação) e *b* = 0 (parâmetro de dificuldade do item) 35
Figura 5 Comparações de CCIs e função informação do item. 38
Capítulo 2 52
Figura 1 Correlações entre os valores paramétricos (a, b, c, θ) e os valores estimados ($\hat{a}$, $\hat{b}$, $\hat{c}$, $\hat{\theta}$) e respectivos intervalos de confiança a 95%. No eixo das abscissas, o primeiro valor se refere ao número de indivíduos e o segundo, ao número de itens 70
Figura 2 Correlação entre a nota e o valor paramétrico (θ) e respectivo intervalo de confiança a 95%. No eixo das abscissas, o primeiro valor se refere ao número de indivíduos e o segundo, ao número de itens 71
Figura 3 Estimativa do erro quadrático médio (EQM) e respectivo intervalo de confiança a 95 %para cada parâmetro (a, b c , θ). No eixo das abscissas, o primeiro valor se refere ao número de indivíduos e o segundo, ao número de itens 72
Figura 4 Estimativa do viés e respectivo intervalo de confiança a 95% para cada parâmetro (a, b c , θ). No eixo das abscissas, o primeiro valor se refere ao número de indivíduos e o segundo, ao número de itens 72
Capítulo 3 83
Figura 1 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2006-2 da UFLA e respectivos intervalos de credibilidade HPD a 95% 94
Figura 2 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2006-2 da UFLA e respectivos intervalos de credibilidade HPD a 95% 95
Figura 3 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2006-2 da UFLA e respectivos intervalos de credibilidade HPD a 95% 96

Figura 4 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha, vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para exemplos de um item muito difícil (item 8), um item muito ruim (item 11) e um item bom (item 74), do vestibular 2006-2, dispostos nessa ordem 98
Figura 5 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2006-2, dispostas nessa ordem 99
Figura 6 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), para o vestibular 2006-2 102
Figura 7 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2007-1 da UFLA e respectivos intervalos de credibilidade HPD a 95% 103
Figura 8 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2007-1 da UFLA e respectivos intervalos de credibilidade HPD a 95% 104
Figura 9 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2007-1 da UFLA e respectivos intervalos de credibilidade HPD a 95% 105
Figura 10 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para exemplos de um item muito difícil (item 47), um item muito ruim (item 26) e um item bom (item 45), do vestibular 2007-1, dispostas nessa ordem 107
Figura 11 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2007-1, dispostos nesta ordem 109

Figura 12 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita) do vestibular 2007-1 112
Figura 13 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2007-2 da UFLA e respectivos intervalos de credibilidade HPD a 95% 113
Figura 14 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2007-2 da UFLA e respectivos intervalos de credibilidade HPD a 95% 114
Figura 15 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2007-2 da UFLA e respectivos intervalos de credibilidade HPD a 95% 115
Figura 16 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para os itens 27, 2 e 56, do vestibular 2007-2, dispostos nessa ordem . . 117
Figura 17 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2007-2, dispostas nessa ordem 119
Figura 18 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), referentes ao vestibular 2007-2 122
Figura 19 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2008-1 da UFLA e respectivos intervalos de credibilidade HPD a 95% 123
Figura 20 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2008-1 da UFLA e respectivos intervalos de credibilidade HPD a 95% 124
Figura 21 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2008-1 da UFLA e respectivos intervalos de credibilidade HPD a 95% 125

Figura 22 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para exemplos de um item muito difícil (item 4), um item muito ruim (item 2) e um item bom (item 72), do vestibular 2008-1, dispostos nessa ordem 127
Figura 23 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2008-1, dispostas nessa ordem 128
Figura 24 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), do vestibular 2008-1 131
Figura 25 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2008-2 da UFLA e respectivos intervalos de credibilidade HPD a 95% 132
Figura 26 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2008-2 da UFLA e respectivos intervalos de credibilidade HPD a 95% 133
Figura 27 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2008-2 da UFLA e respectivos intervalos de credibilidade HPD a 95% 134
Figura 28 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para exemplos de um item muito difícil (item 19), um item muito ruim (item 3) e um item bom (item 68), do vestibular 2008-2, dispostos nessa ordem 136
Figura 29 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2008-2, dispostas nessa ordem 138

Figura 30 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), referente ao vestibular 2008-2 140
Figura 31 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2009-1 da UFLA e respectivos intervalos de credibilidade HPD a 95% 142
Figura 32 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2009-1 da UFLA e respectivos intervalos de credibilidade HPD a 95% 143
Figura 33 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2009-1 da UFLA e respectivos intervalos de credibilidade HPD a 95% 144
Figura 34 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para os itens 66, 31 e 60, do vestibular 2009-1 146
Figura 35 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2009-1, dispostas nessa ordem 152
Figura 36 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), do vestibular 2009-1 154
Capítulo 4 157
Figura 1 Poder de discriminação das provas dos vestibulares de 2006-2 a 2009-1 da UFLA. Letras iguais na coluna indicam que não há diferenças significativas pelo teste de Tuckey. Para cursos que se encontram sem letras, repetem-se as letras do curso imediatamente superior 165
Figura 2 Grau de dificuldade das provas dos vestibulares de 2006-2 a 2009-1 da UFLA. Letras iguais nas colunas indicam que não há diferenças significativas pelo teste de Tuckey. Para cursos que se encontram sem letras, repetem-se as letras do curso imediatamente superior 168

Figura 3 Probabilidade de acerto por indivíduos com baixa habilidade das provas dos vestibulares de 2006-2 a 2009-1 da UFLA 172
Figura 4 Gráfico do máximo de informação dadas pelas provas ao longo dos Vestibulares de 2006-2 a 2009-1 da UFLA. Letras iguais na coluna indicam que não há diferenças significativas pelo teste de Tuckey. Para cursos que se encontram sem letras, repetem-se as letras do curso imediatamente superior 174
Figura 5 Gráfico da informação ponderada pelas habilidades dadas pelas provas ao longo dos Vestibulares de 2006-2 a 2009-1 da UFLA. Letras iguais na coluna indicam que não há diferenças significativas pelo teste de Tuckey. Para cursos que se encontram sem letras, repetem-se as letras do curso imediatamente superior . . . 176
Figura 6 Gráfico da média das habilidades por curso ao longo dos vestibulares de 2006-2 a 2009-1 da UFLA 178
Figura 7 Gráfico do números de candidados por vaga $\times$ média das habilidades por curso ao longo dos vestibulares de 2006-2 a 2009-1 da UFLA 179

LISTA DE TABELAS

Capítulo 1 18
Capítulo 2 52
Tabela 1 Coeficientes para o modelo quadrático de superfície de resposta nas variáveis de correlação. (*n*: número de indivíduos; *p*: número de itens) 73
Tabela 2 Coeficientes para o modelo quadrático de superfície de resposta do EQM. (*n*: número de indivíduos; *p*: número de itens) 73
Tabela 3 Coeficientes para o modelo quadrático de superfície de resposta do viés. (*n*: número de indivíduos; *p*: número de itens) 74
Tabela 4 Coeficientes para o modelo quadrático de superfície de resposta para o tempo de execução. (*n*: número de indivíduos; *p*: número de itens) 74
Capítulo 3 83
Tabela 1 Disciplinas e itens dos vestibulares 2006-2 a 2009-1 da UFLA . 87
Tabela 2 Relação de candidatos por vagas inscritos aos diversos cursos oferecidos pela UFLA para os vestibulares 2006-2 a 2009-1 . . . 88
Tabela 3 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2006-2 da UFLA e seus respectivos erros de Monte Carlo (EMC) 100
Tabela 4 Médias das habilidades por curso e respectivo desvio padrão (sd) referentes ao vestibular 2006-2 101
Tabela 5 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2007-1 da UFLA e seus respectivos erros de Monte Carlo (EMC) 110
Tabela 6 Médias das habilidades por curso e respectivo desvio padrão (sd) referentes ao vestibular 2007-1 111
Tabela 7 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2007-2 da UFLA e seus respectivos erros de Monte Carlo (EMC) 120
Tabela 8 Médias das habilidades por curso do vestibular 2007-2 e respectivo desvio padrão (sd) 121
Tabela 9 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2008-1 da UFLA e seus respectivos erros de Monte Carlo (EMC) 130
Tabela 10 Médias das habilidades por curso e respectivo desvio padrão (sd), referentes ao vestibular 2008-1 130

Tabela 11 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidades de dois candidatos do vestibular 2008-2 da UFLA e seus respectivos erros de Monte Carlo (EMC) 139
Tabela 12 Médias das habilidades por curso do vestibular 2008-2 e respectivo desvio padrão (sd) 139
Tabela 13 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2009-1 da UFLA e seus respectivos erros de Monte Carlo (EMC) 153
Tabela 14 Médias das habilidades por curso do vestibular 2009-1 e respectivo desvio padrão (sd) 153
Capítulo 4 157
Tabela 1 Relação de candidatos por vagas inscritos aos diversos cursos oferecidos pela UFLA para os vestibulares 2006-2 a 2009-1 . . . 161
Tabela 2 Quadro-resumo da ANAVA das provas e vestibulares da UFLA de 2006-2 a 2009-1 com relação ao parâmetro a 164
Tabela 3 Quadro-resumo da ANAVA das provas e vestibulares da UFLA de 2007-1 a 2009-1 com relação ao parâmetro a 166
Tabela 4 Quadro-resumo da ANAVA das provas e vestibulares da UFLA de 2006-2 a 2009-1 com relação ao parâmetro b 167
Tabela 5 Quadro-resumo da ANAVA das provas e vestibulares da UFLA de 2006-2 a 2009-1 com relação ao parâmetro c 167
Tabela 6 Médias das provas dos vestibulares de 2006-2 a 2009-1 da UFLA com relação aos parâmetros a, b e c 169
Tabela 7 Médias dos parâmetros de discriminação (a), dificuldade (b) e probabilidade de acerto por indivíduos com baixa habilidade (c) das provas dos vestibulares 2006-2 a 2009-1 da UFLA, sendo consideradas por semestre 171
Tabela 8 Quadro-resumo da ANAVA das provas e os vestibulares de 2006-2 a 2009-1 da UFLA com relação ao máximo da informação . . 173
Tabela 9 Quadro-resumo da ANAVA das provas e os vestibulares de 2007-1 a 2009-1 da UFLA com relação ao máximo da informação . . 173
Tabela 10 Quadro-resumo da ANAVA das provas e os vestibulares de 2006-2 a 2009-1 da UFLA com relação à informação ponderada pelas habilidades 175

SUMÁRIO

	CAPÍTULO 1 Introdução geral	18
1	**INTRODUÇÃO**	18
2	**REFERENCIAL TEÓRICO**	21
2.1	**Conceitos gerais**	21
2.2	**Modelos para Teoria de Resposta ao Item**	24
2.2.1	**Curva Característica do Item (CCI)**	28
2.2.2	**Interpretação dos parâmetros do item**	31
2.2.3	**Exemplos de curvas características do item (CCI)**	33
2.2.4	**Curva característica do teste (CCT)**	35
2.2.5	**Função informação do item (FII)**	36
2.2.6	**Função informação do teste (FIT)**	39
2.3	**Métodos de estimação**	39
2.3.1	**Inferência *bayesiana***	40
2.3.2	**Método Monte Carlo via cadeias de Markov (MCMC)**	42
2.3.3	**Algoritmo Metropolis-Hastings (MH)**	44
2.3.4	**Convergência das cadeias**	45
2.3.5	**Intervalo de máxima densidade a *posteriori* (HPD)**	46
2.3.6	**Sumário da tese**	47
	REFERÊNCIAS	48
	CAPÍTULO 2 Avaliação do programa para a análise do modelo de três parâmetros da Teoria de Resposta ao Item	52
1	**INTRODUÇÃO**	54
2	**REFERENCIAL TEÓRICO**	56
2.1	**Inferência bayesiana**	58
3	**METODOLOGIA**	61
3.1	**Processo de simulação**	61
3.1.1	**Escolha das distribuições a *priori***	62
3.1.2	**Distribuição conjunta a *posterior*i e distribuições condicionais completas**	64
3.2	**Implementação do processo amostral**	68
3.3	**Análise do experimento**	68
4	**RESULTADOS**	70
5	**DISCUSSÃO**	76
6	**CONCLUSÃO**	78
	REFERÊNCIAS	79
	CAPÍTULO 3 Análise dos vestibulares 2006-2 à 2009-1 da UFLA	83
1	**INTRODUÇÃO**	85

2 METODOLOGIA 87
2.1 Material 87
2.2 Metodologia 88
2.2.1 Implementação do processo amostral 90
2.2.2 Análise dos resultados 90
3 RESULTADOS 92
3.1 Análise do vestibular 2006-2 92
3.2 Análise do vestibular 2007-1 102
3.3 Análise do vestibular 2007-2 112
3.4 Análise do vestibular 2008-1 122
3.5 Análise do vestibular 2008-2 131
3.6 Análise do vestibular 2009-1 140
4 CONCLUSÃO 155
REFERÊNCIAS 156
CAPÍTULO 4 Análise combinada dos vestibulares 2006-2 a 2009-1 da UFLA 157
1 INTRODUÇÃO 159
2 METODOLOGIA 161
2.1 Material 161
2.2 Metodologia 162
3 RESULTADOS 164
4 CONCLUSÃO 181
REFERÊNCIAS 182
APÊNDICE 183

CAPÍTULO 1

Introdução geral

1 INTRODUÇÃO

A teoria de resposta ao item (TRI) é definida como um conjunto de modelos para a probabilidade de uma pessoa obter um escore a um determinado item, em função de sua habilidade e de características do item (ANDRADE, 2001). Seu surgimento se deu devido a discussões sobre a viabilidade de se compararem as habilidades de indivíduos submetidos a provas diferentes (HAMBLETON; SWAMINATHAN; ROGERS, 1991). A forma de medir tais habilidades até então, denominada teoria clássica dos testes (TCT), só permitia a comparação entre indivíduos ou grupos de indivíduos que tivessem sido submetidos à mesma prova ou então àquelas que produzissem uma mesma média e desvio-padrão, que também são chamadas de provas paralelas (LORD, 1980). No entanto, a obtenção de formas assim paralelas é muito difícil de ser conseguida.

Os modelos mais básicos da TRI são caracterizados por dois tipos de parâmetros: os dos itens e os das habilidades. Os parâmetros dos itens estão relacionados às questões, sendo eles grau de dificuldade (b), poder de discriminação (a) e probabilidade de acerto por candidatos com baixa habilidade (c); e os das habilidades (θ), às características dos candidatos que respondem a estas questões. Uma propriedade muito importante é o fato de que esses parâmetros, tanto dos itens como das habilidades, são invariantes dentro de uma mesma popu-

lação (BAKER, 2001) e é devido a ela que se torna possível a comparação entre grupos de candidatos, tenham eles respondido a provas iguais ou diferentes, assim como a avaliação de todos os itens de forma que os mesmos não dependam do grupo de respondentes.

Para as Universidades em geral é de interesse identificar quais candidatos possuem as melhores habilidades para ocupar suas vagas, assim como identificar quais itens ou tipos de itens são mais eficientes na seleção desses candidatos. No entanto, os exames vestibulares são ainda muito diferenciados no país e, frequentemente, tais candidatos têm a opção de fazer diferentes provas de língua estrangeira. Dessa forma, são submetidos a tipos de provas diferentes, já que nem todos os itens são comuns. Bragion e Bueno Filho (2007) apresentam uma análise *bayesiana* de uma prova isolada (candidatos da Agronomia da UFLA), ignorando essas questões de língua estrangeira.

Outro fator a ser considerado é quanto aos recursos computacionais utilizados para a elaboração dos resultados de interesse na TRI. Duas questões podem ser ressaltadas: a) os *softwares* existentes são, em sua maioria, comerciais, o que, segundo Pinheiro (2006), de certa forma dificulta a expansão do método; b) a elaboração dos resultados de interesse quando se tem um conjunto de dados desbalanceados empregando-se a metodologia *bayesiana* no modelo logístico de três parâmetros (ML3P) exige um maior dispêndio computacional. Isso gera a necessidade do aprimoramento de técnicas computacionais com desenvolvimento de programas para serem executados em *softwares* livres que agilizem a amostragem de Monte Carlo via cadeia de Markov (MCMC) da distribuição a *posteriori* conjunta.

Assim, este trabalho teve como objetivos:

1) estabelecer um algoritmo eficiente e rápido para o modelo de TRI de

três parâmetros com dados desbalanceados usando inferência *bayesiana*, sendo este programado para ser utilizado no *software* R (que é gratuito);

2) testar o algoritmo elaborado através de um estudo de simulação a fim de aplicá-lo na análise de exames vestibulares incluindo todos os cursos e todos os itens;

3) analisar as provas de múltipla escolha dos vestibulares da UFLA de 2006-2 a 2009-1, identificando quais propriedades dos itens o classificam como melhor ou pior a fim de orientar novas provas;

4) comparar as provas ao longo dos anos, procurando estabelecer tendências de progresso na informação por prova;

5) identificar a evolução das habilidades médias dos cursos ao longo dos anos.

Pode-se ressaltar o fato de que um estudo aplicado a dados de vestibular tem uma grande importância, pois virá contribuir para que se tenham vestibulares que apresentem questões bem formuladas quanto ao conteúdo, à distribuição dos níveis de dificuldade e discriminação, alcançando, assim, de forma mais eficiente, os objetivos propostos, que são maximizar as chances de que as pessoas aprovadas e selecionadas sejam de fato as mais capacitadas.

A seguir será apresentado um breve referencial em que os principais conceitos e modelos sobre esta teoria são enunciados. Nas subseções serão abordados comentários sobre os métodos de estimação, idéias básicas sobre inferência *bayesiana* e o método MCMC. No final, será apresentado um sumário do trabalho.

2 REFERENCIAL TEÓRICO

2.1 Conceitos gerais

A TRI surgiu da necessidade de se superarem as limitações apresentadas pela TCT (HAMBLETON; SWAMINATHAN; ROGERS, 1991).

O interesse em se ter um instrumento para medir as aptidões humanas remonta a longa data. No entanto, a inteligência como objeto de estudo da ciência psicológica surgiu no final do século XIX com as pesquisas da Psicometria sendo atribuído ao psicólogo francês Alfred Binet (1857-1911), em colaboração com Victor Henri e Theodore Simon, o desenvolvimento, em 1905, do primeiro teste de inteligência propriamente dito (BINET; SIMON, 1907). Essa abordagem marcou o início do desenvolvimento dos testes de inteligência contemporâneos. Muitos outros testes foram desenvolvidos, a partir de então, podendo ser citados os testes de personalidade, de aptidão verbal, de memória, de raciocínio e o vocacional, entre outros. O termo "QI" (quociente de inteligência = idade mental dividida pela idade cronológica), por exemplo, foi proposto por Wilhelm Stern, em 1912. Em 1916, Lewis Madison Terman sugeriu multiplicar o QI por 100, a fim de eliminar a parte decimal. Em 1917, devido à Primeira Grande Guerra Mundial, o exército americano precisava fazer uma classificação de seus melhores soldados a fim de colocar cada um no posto para o qual melhor se adequava e isso de forma rápida. Houve, pois, a necessidade de que uma prova coletiva para um grupo numeroso fosse elaborada. Surgiram, então, os testes Army Alpha e Beta, desenvolvidos por Robert Yerkes (YERKES, 1921). Esses tipos de teste se estenderam rapidamente ao campo social, industrial e, sobretudo, ao educacional (BAQUERO, 1968).

Em 1950, a TCT já estava bem axiomatizada, contudo apresentava muitas limitações: todas as suas medidas dependiam das características dos indivíduos

que realizavam o teste; tanto a dificuldade do item (definida como a proporção de indivíduos que acertaram ao item) quanto sua discriminação dependia do grupo de indivíduos do qual elas foram obtidas; a habilidade dos indivíduos que realizavam a prova era medida pelo total de acertos obtidos (chamados de escores brutos) e aumentava ou diminuía dependendo da dificuldade do teste, ou seja, testes com dificuldades diferentes produziam diferentes estimativas das habilidades; a comparação entre indivíduos que não foram submetidos à mesma prova não era possível (ANDRADE; TAVARES; VALLE, 2000).

Alguns desses problemas foram levantados por Thurstone (1928), porém, sem conseguir encontrar solução. Ele diz: "um instrumento de medida, na sua função de medir, não pode ser seriamente afetado pelo objeto de medida. Na extensão em que sua função de medir for assim afetada, a validade do instrumento é prejudicada ou limitada" (THURSTONE, 1928, p. 547).

Uma proposta para contornar o problema da comparação entre indivíduos que não realizaram a mesma prova foi a condição de que essas provas deviam ter formas paralelas. Gulliksen (1950) definiu que duas provas podiam ser consideradas formas paralelas quando, após a conversão para a mesma escala, suas médias, desvio-padrão de acertos, bem como demais correlações do número de acertos com todo e qualquer outro critério fossem iguais. Porém, provas paralelas são muito difíceis de serem obtidas.

Essas limitações da TCT fizeram com que se buscassem estruturas de medida alternativa, tanto para o conhecimento da habilidade de um examinado, como das estatísticas dos itens, fazendo com que a apresentação de resultados não fosse feita somente através de percentuais de acertos ou escores dos testes e ainda da dificuldade de comparar resultados de diferentes testes em diferentes situações.

Esses anseios foram resolvidos pela TRI. Ela passa a tratar o problema da

estimação da habilidade e conhecimento de um examinando de forma essencialmente diferente: o enfoque das análises desvincula-se das provas e concentra-se nos itens. Assim, a resposta que o indivíduo dará ao item (e não mais à prova como um todo) dependerá do nível de habilidade que possui, ou, em outras palavras, o item passa a ser um estímulo que o leva a uma resposta, a qual dependerá do nível de sua habilidade. E tanto essa habilidade como as características dos itens são invariantes, isto é, a habilidade de um indivíduo não depende da prova a ele aplicada, assim como as características dos itens também não dependem de quem realiza a prova. É devido a essa invariância dos parâmetros que é possível a comparação entre indivíduos mesmo respondendo a uma prova em que nem todos os itens são comuns.

Na TRI, a característica individual determinante de como responder aos itens de um teste é medida através de uma relação probabilística com cada um dos itens utilizados, sendo a representação gráfica dessa probabilidade chamada curva característica do item (CCI).

Vários autores contribuíram para a elaboração dessa teoria, dentre eles Richardson (1936), que compara os parâmetros dos itens obtidos pela teoria clássica da Psicometria com os moldes que hoje usa a TRI; Lawley (1943) que indica alguns métodos para estimar os parâmetros dos itens, os quais se afastavam da teoria clássica; Tucker (1946), que parece ter sido o primeiro a utilizar a expressão curva característica do item (CCI) e a contribuição de Lazersfeld (1950), que introduziu o conceito de traço latente, ainda que no contexto da medida das atitudes. Entretanto, Lord (1952) é considerado o responsável por dar as bases da moderna TRI. Também de suma importância têm-se os trabalhos de Rasch (1960) e Birnbaum (1968) devido aos modelos por eles elaborados.

Apesar das vantagens da TRI, seu crescimento e divulgação somente co-

meçaram a ocorrer a partir dos anos 80. Isso porque, os algoritmos matemáticos necessários para ajustar os modelos da TRI são de tal complexidade que a tecnologia computacional da época em que surgiu era incapaz de resolvê-los de uma maneira prática. Somente com a revolução na capacidade de processamento (computadores) é que se tornou possível a viabilização dos cálculos que o modelo TRI exige (PASQUALI; PRIMI, 2003). O primeiro *software* desenvolvido para as análises da TRI se deu somente em 1979, com o BICAL de Wright e Mead (PASQUALI; PRIMI, 2003). No Brasil, a primeira aplicação dessa teoria ocorreu em 1995 na análise do Sistema Nacional de Avaliação da Educação Básica (SAEB). A partir de então, seu uso nas avaliações educacionais brasileiras tem sido valorizado e incentivado pelos órgãos governamentais, como o Ministério da Educação (VALLE, 1999).

2.2 Modelos para Teoria de Resposta ao Item

Existem vários modelos para teoria de resposta ao item que diferem em sua forma em função dos diferentes tipos de itens, do número de populações envolvidas e da modelagem da habilidade que está sendo avaliada. Os itens podem ser dicotômicos, isto é, acerta-se ou não; politômicos, em que o modelo atribui uma probabilidade para cada categoria de resposta; ou de caráter contínuo, utilizados em questões abertas. Os modelos mais amplamente utilizados são para itens dicotômicos e serão comentados abaixo. Modelos em que mais de uma habilidade está sendo medida (modelos multidimensionais) podem ser encontrados em Hambleton e Cook (1977); modelos de resposta gradual, em que não são consideradas simplesmente as respostas como corretas ou incorretas, em Samejima (1969); modelos para duas ou mais populações, em Bock e Zimowski (1997); estrutura geral que considera modelos de grupos múltiplos e longitudinais conjuntamente

em Azevedo (2008).

Segue uma apresentação genérica dos modelos de TRI: seja Y_{ij} a variável aleatória associada ao acerto ou erro na resposta do indivíduo *i* ao item *j*, *i* = 1, ..., n, *j* = 1, ..., k. Y_{ij} pode assumir valores 1 (acerto) ou 0 (erro). Um modelo amostral para a observação y_{ij} seria a função distribuição de probabilidade (f.d.p.) de Bernoulli:

$$f_{Y_{ij}}(y_{ij};\pi_{ij}) = \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$$

sendo que π_{ij} está em função dos parâmetros e representa a probabilidade condicional de que o indivíduo *i* responda corretamente ao item *j*. Nesta expressão a forma da probabilidade π_{ij} é dada por:

$$\pi_{ij} = P(Y_{ij} = 1|\theta, a, b, c) = c_j + (1-c_j).F(m_{ij})$$

em que:

a) $m_{ij} = a_j(\theta_i - b_j)$: uma função linear de θ;

b) θ_i : parâmetro da habilidade do indivíduo i;

c) a_j: parâmetro do poder de discriminação do item j;

d) b_j: parâmetro do grau de dificuldade do item j ou índice de locação;

e) c_j: parâmetro da probabilidade de que um indivíduo com habilidade infinitamente baixa acerte o item j;

f) F(.): uma função estritamente não decrescente, dada pela função distribuição acumulada (f.d.a.) de uma determinada distribuição de probabilidade.

O primeiro modelo de TRI foi formalmente introduzido por Lord (1952), o qual considerou $F(.) = \Phi(.)$, sendo $\Phi(.)$ a f.d.a. da distribuição normal padrão. Nesse modelo não estava incluído o parâmetro c. Por utilizar como F(.) a dis-

tribuição acumulada da distribuição normal, é chamado de modelo ogiva normal e é dado pela expressão:

$$\pi_{ij} = P(Y_{ij} = 1|\theta, a, b) = F(m_{ij}) = \Phi[a_j(\theta_i - b_j)]$$

Ao ser determinada f(.), a função densidade de probabilidade (f.d.p.) dessa distribuição, isto é, $\phi[a_j(\theta_i - b_j)]$, pode-se notar que o parâmetro b_j está relacionado à média e o parâmetro a_j, à mudança de inclinação e que, quando b_j é igual ao parâmetro θ_i, $\pi_{ij} = F_\theta(m_{ij}) = \Phi(0) = 0,5$.

$$\frac{\partial[F_\theta(m_{ij})]}{\partial\theta} = \frac{\partial\{\Phi[a_j(\theta_i - b_j)]\}}{\partial\theta}$$

$$= \phi[a_j(\theta_i - b_j)].a_j$$

$$= \frac{1}{\sqrt{2\pi}.1/a_j}.e^{-\frac{1}{2}\left(\frac{\theta_i - b_j}{1/a_j}\right)^2}$$

Mais tarde, Birnbaum (1968) considerou a função de distribuição acumulada da distribuição logística para $F(.)$, isto é, $F(.) = L(m_{ij})$, sendo $L(m_{ij}) = \frac{e^{m_{ij}}}{1+e^{m_{ij}}} = \frac{1}{1+e^{-m_{ij}}}$. Como neste modelo apenas são considerados dois parâmetros para os itens, é chamado modelo logístico de dois parâmetros e é expresso por:

$$\pi_{ij} = P(Y_{ij} = 1|\theta, a, b, c) = \frac{1}{1 + e^{-a_j(\theta_i - b_j)}}$$

No caso em que $a_j = 1$, tem-se o modelo logístico de um parâmetro, conhecido como modelo de Rasch (RASCH, 1960), expresso por:

$$\pi_{ij} = P(Y_{ij} = 1|\theta, a, b, c) = \frac{1}{1 + e^{-(\theta_i - b_j)}}$$

Entretanto, a possibilidade de acerto por indivíduos com baixa habilidade existe. Sendo assim, modelar a probabilidade de resposta correta de pessoas que respondam ao acaso ou tenham baixo nível de conhecimento não é zero, como determinam os modelos de um e dois parâmetros. Questões de múltipla escolha sempre permitem que sejam acertadas, mesmo por quem não domine o conhecimento. Entretanto, Lord (1952) percebeu que, em geral, esse percentual de acerto, em nível muito baixo de habilidade, não era simplesmente igual ao inverso do número de alternativas e sim variado. Como exemplo, considere-se uma questão sobre resolução de equação de segundo grau em que uma de suas alternativas tenha a soma do quadrado de suas raízes como sendo um número negativo. Esta alternativa não será a escolhida por um indivíduo que, mesmo não sabendo resolver uma equação de segundo grau, sabe que a soma de potências pares sempre são positivas. A capacidade de detectar esse fato faz parte da habilidade do indivíduo. Consequentemente, mesmo ele tendo baixa habilidade, a probabilidade de acerto casual dessa questão aumenta. Para representar esse fato mais adequadamente, Birnbaum (1968) propôs a introdução do parâmetro c_j ao modelo que é chamado de modelo logístico de três parâmetros (ML3P). Assim, dependendo do número de parâmetros relacionados ao item, os modelos são classificados como sendo de 1, 2 ou 3 parâmetros.

Considerando-se, pois, 3 parâmetros para os itens a expressão geral de π_{ij} é dada por:

$$\pi_{ij} = P(Y_{ij} = 1|\theta, a, b, c) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(\theta_i - b_j)}}$$

denominado, então, modelo logístico de três parâmetros.

O ML3P pode ser expresso graficamente através do que se chama curva

característica do item (CCI) e será descrita no tópico seguinte.

2.2.1 Curva Característica do Item (CCI)

Curva Característica do Item (CCI) é o nome dado à curva que é obtida ao ser plotado π_{ij} em função de θ, para um determinado item j. Pode ser vista, também, como a proporção de respostas corretas de todos os indivíduos com habilidade θ_i a esse item. (Figura 1).

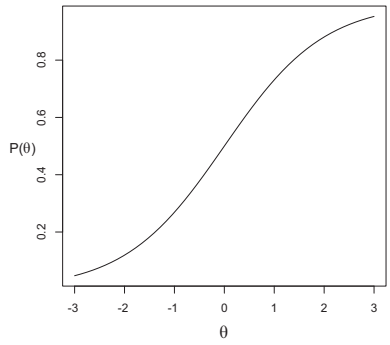


Figura 1 Curva característica do item - CCI

Observa-se que quanto maior for a habilidade de um indivíduo, maior será a probabilidade de que ele responda corretamente ao item, e essa relação não é linear.

Cada item do teste terá sua própria CCI.

Nota-se que o eixo das ordenadas varia de 0 a 1. Isso é coerente, pois o mesmo representa probabilidades de acerto. Para o eixo das abscissas, porém, onde estão representadas as habilidades dos indivíduos, a escala não é tão evidente, no entanto é coerente. O fato é que, diferentemente da avaliação usual, em que a habilidade é diretamente proporcional à nota e varia de 0 a *n*, sendo *n* o número

de itens de um teste, a escala de habilidade pode assumir, teoricamente, qualquer valor real entre $-\infty$ a $+\infty$. Na realidade, o que importa não é sua amplitude e, sim, as relações entre seus pontos. O que se necessita é que seja estabelecida uma origem e uma unidade de medida para a definição desta, sendo que esses valores devem representar, respectivamente, o valor médio (μ) e o desvio padrão (σ) das habilidades da população em estudo. Como a distribuição dessas habilidades pode ser representada por uma distribuição normal, o que usualmente se faz é padronizá-las pela densidade da normal padrão. Assim, é bastante utilizada pela TRI a escala com média 0 e desvio padrão igual a 1, representada por escala (0;1). No entanto, deve ficar claro que, apesar de ser frequente a utilização dessa escala, não necessariamente se necessita de que a mesma seja adotada.

Como exemplo, considere-se um teste com 120 questões, sendo que a média de acertos foi de 60 com um desvio padrão de 10. Esta escala é representada pela escala (60;10). Um indivíduo com habilidade 72, nessa escala, terá habilidade 1,20 na escala (0;1), isto é, em qualquer escala, ele estará a 1,20 desvio-padrão acima da habilidade média [(72-60)/10].

No modelo proposto, o parâmetro *c* não depende da escala, pois trata-se de uma probabilidade e o parâmetro *b* é medido na mesma unidade da habilidade. Assim, a parte do modelo probabilístico proposto que é modificado pela transformação de escala é $m_{ij} = a_j(\theta_i - b_j)$ (ANDRADE; TAVARES; VALLE, 2000). O valor correspondente de uma escala para outra pode ser obtida pela seguinte regra de transformação:

$\theta_i^* = \sigma \times \theta_i + \mu$

$b_j^* = \sigma \times b_j + \mu$

$a_j^* = a_j/\sigma$

Prova:

$$a_j(\theta_i - b_j) = \frac{a_j}{\sigma}.\sigma(\theta_i - b_j) = \frac{a_j}{\sigma}(\sigma.\theta_i - \sigma.b_j) = \frac{a_j}{\sigma}(\sigma.\theta_i + \mu - \mu - \sigma.b_j)$$

$$a_j(\theta_i - b_j) = \underbrace{a_j/\sigma}_{a_j^*}[\underbrace{(\sigma.\theta_i + \mu)}_{\theta_i^*} - \underbrace{(\sigma.b_j + \mu)}_{b_j^*}]$$

Considerando o mesmo exemplo anterior, e sendo $a_j = 0,60$, $b_j = -0,30$, e $c_j = 0,25$, temos:

$\theta_i^* = 10 \times 1,20 + 60 = 72$

$b_j^* = 10 \times (-0,30) + 60 = 57$

$a_j^* = 0,60/10 = 0.06$

A probabilidade de acerto, considerando as duas escalas, será:

Escala (0;1):

$$P(Y = 1|\theta = 1,20) = 0,25 + \frac{(1 - 0,25)}{1 + e^{-0,60(1,20+0,30)}} = 0,78$$

Escala (60;10):

$$P(Y = 1|\theta = 72) = 0,25 + \frac{(1 - 0,25)}{1 + e^{-0,06(72-57)}} = 0,78$$

Isto mostra que a probabilidade de um indivíduo acertar um determinado item não se altera em função da escala adotada para medir sua habilidade.

Esse é um ponto importante do modelo a ser mencionado, ou seja, ele é não identificável quando se tem todos os parâmetros desconhecidos, pois qualquer transformação do tipo acima, não altera a probabilidade representada pelo modelo. A maneira mais comum para torná-lo identificável é fixar uma padronização para

isso. Em geral, adota-se a distribuição normal padronizada como distribuição a *priori* para as habilidades, definindo a escala com média 0 e desvio-padrão 1.

2.2.2 Interpretação dos parâmetros do item

b_j: representa o ponto na escala de habilidade no qual o indivíduo, com habilidade $\theta_i = b_j$ possui $[(c_j+1)/2]$ de probabilidade de responder corretamente a esse item, ou seja,

$$c_j + \frac{1-c_j}{1+e^{-a_j(\theta_i-b_j)}} = \frac{c_j+1}{2} \quad \Rightarrow \quad \frac{1-c_j}{1+e^{-a_j(\theta_i-b_j)}} = \frac{c_j+1}{2} - c_j$$

$$\frac{1-c_j}{1+e^{-a_j(\theta_i-b_j)}} = \frac{c_j+1-2c_j}{2} \quad \Rightarrow \quad \frac{1-c_j}{1+e^{-a_j(\theta_i-b_j)}} = \frac{1-c_j}{2}$$

$$1+e^{-(\theta_i-b_j)} = 2 \quad \Rightarrow \quad e^{-(\theta_i-b_j)} = 1 \quad \Rightarrow \quad e^{-(\theta_i-b_j)} = e^0$$

$$-(\theta_i - b_j) = 0 \quad \Rightarrow \quad \theta_i = b_j$$

Como é medido na mesma escala da habilidade, seu domínio pode variar, teoricamente, entre $-\infty$ a $+\infty$.

a_j: parâmetro do poder de discriminação do item j. Representa a inclinação da curva no ponto b_j (proporcional ao valor da tangente nesse ponto) e faz com que se torne possível diferenciar entre indivíduos que estão abaixo ou acima

do índice de locação.

$$P'(b_j) = \frac{(1-c_j)a_j.e^{-a_j(\theta_i - b_j)}}{[1+e^{-a_j(\theta_i - b_j)}]^2} = \frac{(1-c_j)a_j.e^0}{(1+e^0)^2} = \frac{(1-c_j)a_j}{4} \quad \propto \quad a_j$$

em que $P'(b_j)$ é a derivada primeira de p_{ij} em relação à θ, no ponto b_j.

O domínio deste parâmetro é o conjunto R_+, uma vez que seria incoerente admitir valores negativos para o mesmo, pois, isso indicaria que a probabilidade de responder corretamente a esse item diminuiria com o aumento da habilidade. Um item que possui baixo valor para esse parâmetro indica que o mesmo possui pouco poder de discriminação, isto é, indivíduos com habilidades bem diferentes possuem praticamente a mesma probabilidade de acertá-lo. Segundo Baker (2001), itens cujos parâmetros de discriminação estão entre 0,01 e 0,34 são classificados como itens de discriminação muito baixa; entre 0,35 e 0,64, de baixa discriminação; entre 0,65 e 1,34, de discriminação moderada; entre 1,35 e 1,69, de discriminação alta, e maior que 1,70, de discriminação muito alta.

c_j: parâmetro da probabilidade de que um indivíduo com baixa habilidade responda corretamente o item j. Representa a probabilidade de que o indivíduo, com baixa habilidade, acerte a questão, sem que ele saiba a resposta, isto é,

$$\lim_{\theta \longrightarrow -\infty} \pi_{ij} = c_j$$

Pode ser visto como a assíntota inferior da CCI.

Por se tratar de uma probabilidade, seu domínio está entre 0 e 1.

2.2.3 Exemplos de curvas características do item (CCI)

No que diz respeito à interpretação gráfica de cada um dos parâmetros dos itens, pode ser feita com o auxílio das Figuras 2 a 4 abaixo, em que várias CCIs foram traçadas como exemplo . Para os itens **i** e **ii**, considerou-se $c_j = 0$.

i) Comparações entre curvas características do item (CCI) com a mesma discriminação e diferentes dificuldades

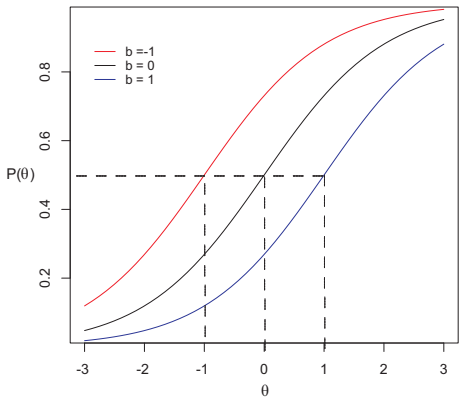


Figura 2 CCIs com diferentes valores para *b* (parâmetro de dificuldade do item) e a = 1 (parâmetro de discriminação do item)

Pode-se observar que a habilidade necessária para que um indivíduo tenha 50% de probabilidade de acerto, que é o que determina o valor do parâmetro b_j é maior quanto mais à direita estiver a CCI e menor quanto mais à esquerda ela estiver. Por exemplo, para ter essa porcentagem de acerto para o item azul, é necessário que o indivíduo possua uma habilidade 1 enquanto que para o item vermelho, com um valor de habilidade mais baixa, -1, já se consegue o mesmo resultado. Logo, o item azul é mais difícil que o vermelho. Assim, o grau de dificuldade do item vai depender de onde ele se encontra na escala de habilidade.

ii) Comparação entre curvas características do item (CCI) com a mesma dificuldade e diferentes discriminações

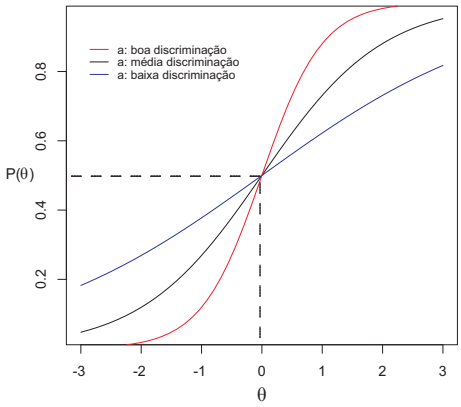


Figura 3 CCIs com diferentes valores para *a* (parâmetro de discriminação do item) e *b* = 0 (parâmetro de dificuldade do item)

Observa-se que quanto maior a inclinação da CCI, isto é, maior o valor do parâmetro a_j, maior é a diferença entre as probabilidades de acerto de indivíduos com diferentes habilidades e vice-versa, ou seja, maior será a capacidade do item diferenciar (discriminar) os indivíduos.

iii) Comparação entre curvas características do item (CCI) com diferentes valores para o parâmetro *c*

Pode-se notar que a diferença na probabilidade de acerto entre indivíduos com habilidades diferentes diminui quando o valor do parâmetro c_j aumenta, ou seja, quanto maior a probabilidade de que um indivíduo com baixa habilidade responda corretamente o item j, se mantivermos fixo o valor do parâmetro a_j, o poder de discriminação diminui.

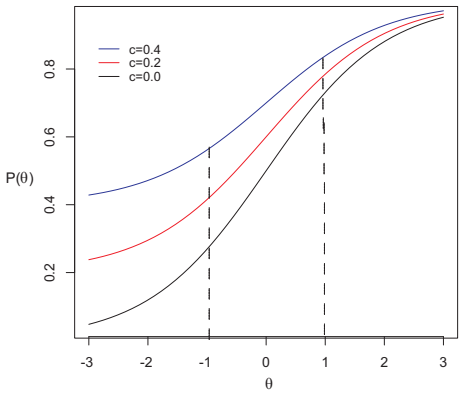


Figura 4 CCIs com diferentes valores para c (parâmetro da probabilidade de acerto casual), a = 1 (parâmetro de discriminação) e b = 0 (parâmetro de dificuldade do item)

2.2.4 Curva característica do teste (CCT)

A curva característica do teste (CCT) é similar à CCI exceto que ela envolve todo o conjunto de questões. Para obtê-la, toma-se a probabilidade de resposta correta para cada θ para todos os itens. Em seguida, essas probabilidades são somadas para cada nível de θ. Após, essas somas são plotadas em função de θ, obtendo-se a CCT. Pode ser indicada por:

$$TS_i = \sum_{j=1}^{k} \pi_{ij}(\theta_i)$$

em que TS_i significa o escore verdadeiro (*true score*) do indivíduo com nível de habilidade θ_i (BAKER, 2001). Pode ser comparado à quantidade de questões corretas dentre um conjunto total de questões. Essa medida pode ser utilizada como sendo a nota de um indivíduo num determinado teste.

O eixo vertical terá seus limites entre 0 e o número de itens no teste. O eixo horizontal continua a ser o nível de habilidade θ. Na CCT, $\theta \rightarrow -\infty$ representa a

soma das probabilidades de acerto casual e reflete o fato de quanto um indivíduo, com baixa habilidade, poderia obter no teste simplesmente pelo "chute".

A forma da CCT vai depender do número de itens e dos valores dos parâmetros de cada um desses itens.

2.2.5 Função informação do item (FII)

A função informação do item (FII) permite analisar o quanto um item traz de informação para a medida de habilidade (ANDRADE; TAVARES; VALLE, 2000). É útil para identificar as questões que são realmente relevantes. Sua equação é obtida através da Informação de Fisher, cuja expressão é representada por:

$$I_j(\theta_i) = -E\left[\frac{\partial^2 lnL(\beta; y)}{\partial \theta_i^2}\right]$$

em que $I_j(\theta_i)$ é a informação fornecida pelo item *j* no nível de habilidade θ_i, $L(\beta; y) = f_{Y_{ij}}(y_{ij}; \pi_{ij})$ e $\beta = (\theta, a, b, c)$.

Para o ML3P, é dada por:

$$I_j(\theta_i) = a_j^2 \frac{(1 - \pi_{ij})}{\pi_{ij}} \left(\frac{\pi_{ij} - c_j}{1 - c_j}\right)^2$$

A prova encontra-se no apêndice A.

Outra forma de obter a função de Informação é através da desigualdade de Cramer Rao (MOOD; GRAYBILL; BOES, 1974), que é dada por:

$$V[Y_{ij}] \geq \frac{(\frac{d}{d\theta_i}\pi_{ij})^2}{-E\left[\frac{\partial^2 lnL(\beta;y)}{\partial \theta_i^2}\right]}$$

sendo $V[Y_{ij}]$ a variância de Y_{ij}.

Portanto,

$$I_j(\theta_i) = \frac{(\frac{d}{d\theta_i}\pi_{ij})^2}{\pi_{ij}(1-\pi_{ij})}$$

Prova:

$$\frac{d}{d\theta_i}\pi_{ij} = \frac{c_j e^{-a_j(\theta_i - b_j)}(-a_j)[1+e^{-a_j(\theta_i-b_j)}] - [1+c_j e^{-a_j(\theta_i-b_j)}]e^{-a_j(\theta_i-b_j)}(-a_j)}{[1+e^{-a_j(\theta_i-b_j)}]^2}$$

$$= \frac{-a_j c_j e^{-a_j(\theta_i-b_j)} - a_j c_j e^{-2a_j(\theta_i-b_j)} + a_j e^{-a_j(\theta_i-b_j)} + a_j c_j e^{-2a_j(\theta_i-b_j)}}{[1+e^{-a_j(\theta_i-b_j)}]^2}$$

$$= \frac{a_j e^{-a_j(\theta_i-b_j)}(1-c_j)}{[1+e^{-a_j(\theta_i-b_j)}][1+e^{-a_j(\theta_i-b_j)}]}$$

$$\frac{d}{d\theta_i}\pi_{ij} = \frac{a_j(1-\pi_{ij})}{1+e^{-a_j(\theta_i-b_j)}}$$

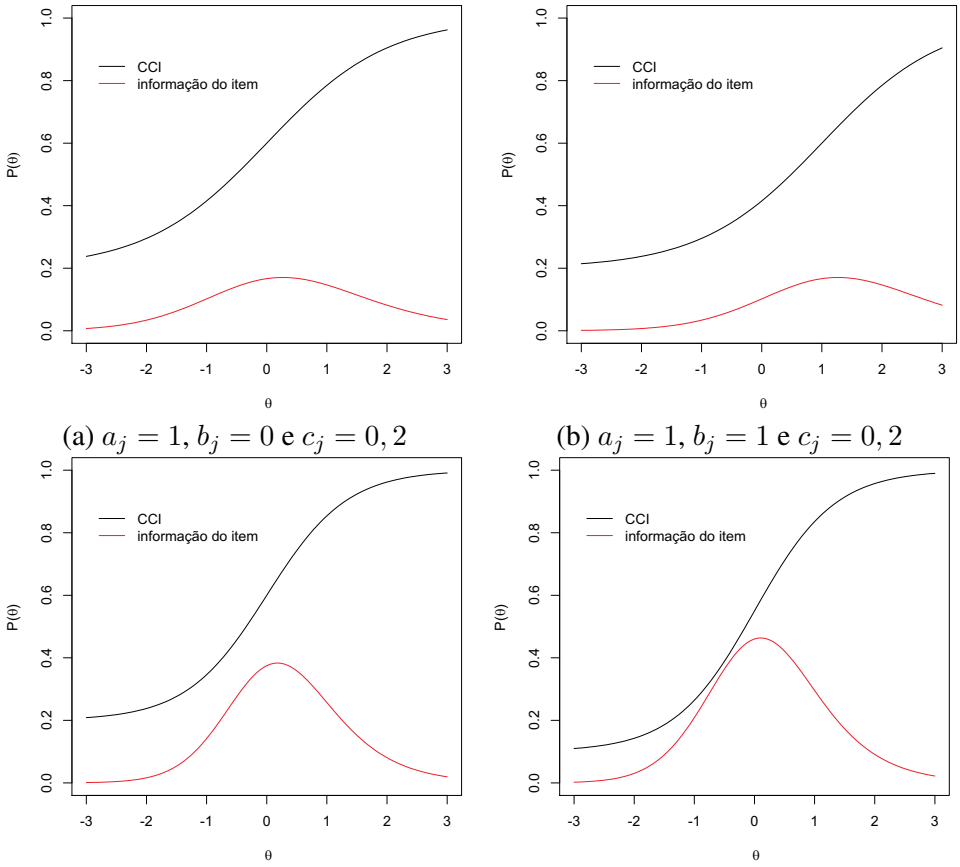
$$I_j(\theta_i) = \frac{\left[\frac{a_j(1-\pi_{ij})}{1+e^{-a_j(\theta_i-b_j)}}\right]^2}{\pi_{ij}(\theta_i)[1-\pi_{ij}(\theta_i)]} = \frac{a_j^2(1-\pi_{ij})^2}{[1+e^{-a_j(\theta_i-b_j)}]^2} \cdot \frac{1}{\pi_{ij}(1-\pi_{ij})}$$

$$= a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \left[\frac{1}{1+e^{-a_j(\theta_i-b_j)}}\right]^2$$

$$\therefore I_j(\theta_i) = a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \left[\frac{\pi_{ij}-c_j}{1-c_j}\right]^2$$

Na Figura 5 estão representados alguns exemplos de CCIs e curvas de

informação com diferentes combinações de valores dos parâmetros do item.



(a) $a_j = 1$, $b_j = 0$ e $c_j = 0,2$

(b) $a_j = 1$, $b_j = 1$ e $c_j = 0,2$

(c) $a_j = 1,5$, $b_j = 0$ e $c_j = 0,2$

(d) $a_j = 1,5$, $b_j = 0$ e $c_j = 0,1$

Figura 5 Comparações de CCIs e função informação do item.

Comparando-se as letras (a) e (b) dessa Figura 5 pode-se notar que a informação é maior quando θ_i se aproxima de b_j; pelas letras (a) e (c) que a informação é maior quanto maior for o valor do parâmetro a_j; e através das letras (c) e (d) que a informação é maior quanto menor for o valor do parâmetro c_j.

2.2.6 Função informação do teste (FIT)

A função informação do teste (FIT) é a informação fornecida pelo teste todo e é obtida pela soma das informações de cada item que compõe o mesmo (BAKER, 2001), isto é,

$$I(\theta_i) = \sum_{j}^{k} I_j(\theta_i)$$

A FIT é um recurso extremamente útil da TRI, pois, por meio dela permite-se saber quão bom está sendo o teste como um todo em fornecer a informação sobre a habilidade de interesse.

É interessante observar que, como se trata da soma das informações do todos os itens, quanto mais itens num teste, maior a quantidade de informação, ou seja, testes mais longos tendem a medir a habilidade dos indivíduos com maior precisão que testes mais curtos.

2.3 Métodos de estimação

Conforme citado no início, os y_{ij} são as respostas de cada indivíduo *i* a cada item *j*, sendo *i* = 1, 2, ..., n, *j* = 1, 2, ..., k. Essas respostas são dependentes da habilidade, pois, conforme a habilidade há uma probabilidade de acertá-las ou errá-las. Porém, dada a habilidade, as respostas que esse indivíduo dará aos diferentes itens da prova são consideradas independentes, isto é, o aluno não estará aprendendo ao longo do teste. Essa propriedade é conhecida como *independência condicional*. Os modelos de TRI satisfazem essa propriedade e também a de *independência entre as respostas de diferentes indivíduos*. Sendo assim, a distribuição conjunta (ou verossimilhança) de $Y = (y_{11}, ..., y_{nk})$, gerada pelas respostas dos *n* indivíduos aos *k* itens, é dada por:

$$f_{Y_{11},...,Y_{nk}}(y_{11},...y_{nk}|\theta,a,b,c) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}})\right]^{y_{ij}} \cdot$$

$$\cdot\left\{1 - \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}}\right]\right\}^{1-y_{ij}}$$

Esse modelo de TRI envolve um total de n + 3k parâmetros desconhecidos, tendo como consequência, um modelo super parametrizado. Os métodos utilizados para estimação destes parâmetros podem ser feitos via inferência frequentista (ou clássica) ou bayesiana. A estimação clássica é feita pelo método da máxima verossimilhança. Usando esta abordagem, várias propostas têm sido formuladas, tais como, a verossimilhança conjunta, verossimilhança marginal e verossimilhança condicional. Azevedo (2003) desenvolveu um trabalho no qual apresenta e discute os principais métodos de estimação.

2.3.1 Inferência *bayesiana*

A Inferência é um ramo da Estatística que visa construir técnicas que permitam fazer afirmações sobre os parâmetros populacionais, com base em dados amostrais. A distribuição dessa amostra é chamada de função de verossimilhança. Ela define uma relação entre os dados e, por exemplo, um parâmetro desconhecido β (MOOD; GRAYBILL; BOES, 1974). Sua notação é dada por $p(y|\beta)$ ou $L(\beta; y)$. A inferência clássica considera esse parâmetro como desconhecido, mas fixo e, portanto, não faz afirmações probabilísticas sobre ele. Na inferência *bayesiana*, o desconhecimento gera incerteza e toda incerteza deve ser quantificada através de probabilidades. Esse desconhecimento ou informação que se tem sobre o parâmetro, antes de ser observada a amostra, deve ser considerado e in-

corporado aos dados, por meio de uma densidade a *priori* $p(\beta)$. Os parâmetros das distribuições a *priori* são denominados de hiperparâmetros e, em geral, são considerados conhecidos. A junção da informação fornecida pelos dados com a densidade a *priori* resulta na densidade a *posteriori* $p(\beta|y)$ e é com base nela que é feita toda inferência *bayesiana* (O'HAGAN, 1994).

O procedimento para construir-se a distribuição a *posteriori* partindo da distribuição a *priori* é o teorema de Bayes, que é dado por:

$$p(\beta|y) = \frac{p(\beta).p(y|\beta)}{p(y)}$$

em que $p(y) = \sum p(\beta).p(y|\beta)$ no caso discreto ou, no caso contínuo, $p(y) = \int p(\beta).p(y|\beta)d\beta$, que não depende de β e, portanto, é uma constante. Logo, $p(\beta|y)$ pode ser escrita como:

$$p(\beta|y) \propto p(\beta).p(y|\beta)$$

Assim, a função de verossimilhança modifica ou atualiza o conhecimento que se tinha sobre o parâmetro antes da obtenção dos dados e a distribuição a *posteriori* reflete o que é conhecido sobre a distribuição a *priori* após a obtenção dos dados. O teorema de Bayes é, pois, a ferramenta que atualiza a informação feita sobre os parâmetros, com base na informação de uma amostra que já ocorreu (BOX; TIAO, 1992).

As distribuições a *priori* podem ser informativas ou não informativas. Como visam representar probabilisticamente o conhecimento que se tem sobre β antes da obtenção dos dados, tendo-se pouca ou nenhuma informação sobre os parâmetros, utilizam-se distribuições *a priori* não informativas. Neste caso, pode-se pensar em todos os valores de β como igualmente prováveis e a distribuição

a *priori* será a distribuição uniforme. Assim, $p(\beta) \propto k$. Outra proposta de distribuição a *priori* não informativa foi feita por Jeffreys (1961), a qual se baseia no uso da informação de Fisher sobre $\beta \in \mathbb{R}$ (PAULINO; TURKMAN; MURTEIRA, 2003). Tal proposta visa minimizar o conteúdo de informação, com o objetivo de que o máximo desta seja extraído da amostra.

O parâmetro β pode ser um escalar ou um vetor, como por exemplo, $\beta = (\beta_1, \beta_2, ..., \beta_p)$ e o que se deseja é inferir sobre os mesmos. No caso desse trabalho, $\beta = (\theta, a, b, c)$. Através da distribuição *a posteriori* conjunta, a distribuição a *posteriori* marginal de cada β_i pode ser obtida e orientar essas inferências. A distribuição a *posteriori* marginal para um determinado β_i é dada por:

$$f(\beta_i|y) = \int \dots \int f(\beta_i, \beta_{-i}) d\beta_{-i}$$

,

em que β_{-i} corresponde ao conjunto complementar de β_i.

A inferência exata somente será feita calculando-se estas integrais analiticamente, o que, com raras exceções, é impossível ou muito trabalhosa. Técnicas de aproximações numéricas são, então, sugeridas sendo bastante utilizados os métodos de Monte Carlo via cadeia de Markov (MCMC). Na próxima seção são fornecidas mais informações sobre esse método.

2.3.2 Método Monte Carlo via cadeias de Markov (MCMC)

O método MCMC consiste num método numérico utilizado para gerar amostras de uma distribuição de interesse. Os valores são gerados de forma iterativa, a partir de distribuições que constituam uma cadeia de Markov. Tais distribuições são as distribuições de transição da cadeia, as quais devem convergir

para uma distribuição estacionária que seja a própria distribuição de interesse. A amostra deve ser grande o suficiente a fim de que a distribuição estacionária possa ter uma boa aproximação da distribuição de interesse. Através de estatísticas calculadas nesta amostra simulada, pode-se estimar os parâmetros correspondentes da distribuição de interesse. É preciso notar que o método MCMC leva a duas propriedades indesejáveis na amostra resultante, quais sejam: a) efeito do valor inicial e b) dependência entre as observações em iterações subsequentes. Uma das formas mais simples de analisar saídas de cadeias de Markov livres destes efeitos é eliminar as primeiras observações (*burn-in*) e fazer saltos de *k* em *k* elementos (*thinning*).

Dos métodos de simulação que utilizam cadeias de Markov, destacam-se o amostrador de Gibbs (AG) e o algoritmo Metropolis e Hastings (MH). O AG foi introduzido no contexto estatístico por Geman e Geman (1984) com grande contribuição de Gelfand e Smith (1990). Nesse algoritmo, para a distribuição de transição da cadeia de Markov são consideradas as distribuições condicionais completas a *posteriori* de cada parâmetro, quando estas possuem uma forma de densidade conhecida e, portanto, fácil de ser amostrada.

A distribuição condicional completa a *posteriori* do parâmetro β_i é obtida quando se considera os demais parâmetros da distribuição a *posteriori* conjunta, os β_{-i}, como conhecidos. É denotada por $f_{\beta_i}(\beta_i|\beta_{-i}, y)$.

Quando distribuições condicionais completas conhecidas não são possíveis de serem obtidas pode-se utilizar o algoritmo MH, o qual consiste em gerar os valores de uma distribuição de transição arbitrária auxiliar da qual se saiba gerar. Esta distribuição será chamada de distribuição geradora de candidatas e será representada por $q(\beta)$. Este algoritmo foi desenvolvido através dos trabalhos de Metropolis et al. (1953) e Hastings (1970). Para esse trabalho, como formas conhecidas das

distribuições condicionais completas a *posteriori* não foram possíveis de serem obtidas, utilizou-se o algoritmo MH. Esse algoritmo encontra-se descrito a seguir. Maiores detalhes sobre os métodos MCMC podem ser encontrados em Gamerman e Lopes (2006).

2.3.3 Algoritmo Metropolis-Hastings (MH)

O procedimento usado para que os valores amostrados pertençam à distribuição de interesse será por meio dos seguintes passos:

1 - atribui-se um valor inicial arbitrário β como primeiro valor da amostra;

2 - calculam-se $P(\beta)$ e $q(\beta)$, sendo $P(\beta) = p(\beta|y)$;

3 - amostra-se β^* da distribuição candidata e calculam-se $P(\beta^*)$ e $q(\beta^*)$;

4 - obtém-se uma razão:

$$r = \frac{\frac{P(\beta^*)}{q(\beta^*)}}{\frac{P(\beta)}{q(\beta)}} = \frac{P(\beta^*)q(\beta)}{P(\beta)q(\beta^*)}$$

5 - e uma probabilidade de aceitação:

$$\alpha(\beta, \beta^*) = min\{1, \frac{P(\beta^*)q(\beta)}{P(\beta)q(\beta^*)}\}$$

6 - gera-se $u \sim U(0, 1)$;

7 - aceita-se β^* como um valor da densidade de interesse se $u \leq \alpha$. Caso contrário, rejeita-se β^*, repete-se β na amostra e retorna-se ao passo 3.

Esse processo é repetido até que o número de iterações desejado seja

atingido.

Segundo Chib e Greenberg (1995), várias são as famílias que podem ser escolhidas para a densidade da candidata. No entanto, conforme afirma Patz e Junker (1999), devido a grande liberdade na escolha da densidade para a mesma, essa escolha pode ser feita de modo conveniente a fim de simplificar a implementação do algoritmo.

Como dito acima, para anular a influência do "chute" inicial, desprezam-se os valores das primeiras iterações ("*burn-in*") e toma-se valores de certa em certa distância ("*thinning*") a fim de que se obtenha uma amostra aleatória.

2.3.4 Convergência das cadeias

A fim de evitar que um número excessivo ou insuficiente de iterações no processo de amostragem seja feito e verificar se a cadeia gerada convergiu para a distribuição de interesse, faz-se necessário o monitoramento da mesma. Há vários métodos utilizados para esse monitoramento disponíveis na literatura, como o critério de Gelman e Rubin (1992), o diagnóstico de Raftery e Lewis (1992), entre outros, que são citados por Nogueira, Sáfadi e Ferreira (2004) os quais analisaram a aplicabilidade e desempenho de cada um deles.

Neste trabalho, foram adotados os critérios de Gelman e Rubin (1992) e o diagnóstico de Raftery e Lewis (1992). No método de Gelman e Rubin (1992) são consideradas mais de uma cadeia, nas quais cada uma delas parte de valores iniciais distintos. Para cada parâmetro de interesse é comparada a variabilidade entre e dentro das cadeias amostradas. A convergência é avaliada por meio de um fator de redução $\hat{R}$. Considera-se que ela foi alcançada quando esse fator for aproximadamente 1. O diagnóstico de Raftery e Lewis (1992) estima o número de iterações necessárias para se alcançar a convergência, o tamanho mínimo do

burn-in e a distância mínima do *thinning*. No pacote *coda* ("*Output Analysis and Diagnostics for MCMC*"), que pode ser instalado no *software* R, encontram-se implementados estes métodos.

Após a obtenção das amostras, inferências de interesse sobre os parâmetros podem ser realizadas, como obtenção de estimativas pontuais (média, mediana ou moda) ou por região (intervalos de credibilidade ou intervalo de máxima densidade a *posteriori* - HPD). No próximo tópico, será apresentado um breve comentário sobre o intervalo HPD.

2.3.5 Intervalo de máxima densidade a *posteriori* (HPD)

As características de uma distribuição podem ser resumidas através de estimativas pontuais, como a média, mediana ou moda. No entanto, um resumo da distribuição a *posteriori* mais informativo que qualquer estimativa pontual é obtido por uma região do espaço paramétrico de β que contenha uma parte substancial dessa distribuição (PAULINO; TURKMAN; MURTEIRA, 2003). Essa região é delimitada por um intervalo que possui um nível de credibilidade $100(1-\alpha)\%$. Porém, esse intervalo de credibilidade não é único. É possível obter uma infinidade deles com o mesmo nível $100(1-\alpha)\%$ de credibilidade. Entretanto, o que mais interessa é aquele que possui menor amplitude. Este intervalo de interesse pode ser obtido tomando-se os valores de θ com maior densidade a *posteriori*. Essa região obtida por esse intervalo de amplitude mínima é chamada de intervalo de máxima densidade a *posteriori* ou HPD (*Highest Probability Density Interval*).

2.3.6 Sumário da tese

Neste capítulo foi feita uma revisão sobre os principais conceitos e modelos mais usuais de TRI, suas ferramentas, método de estimação, idéias básicas sobre inferência bayesiana, assim como o objetivo desta tese. No capítulo 2, será apresentado um estudo de simulação para o ML3P o qual teve como objetivo testar o algoritmo (programa R) empregado que tornou possível realizar uma análise de todas as provas em conjunto (para cada vestibular), de forma mais rápida e eficiente. No capítulo 3, serão feitos estudos dos vestibulares 2006-2 a 2009-1 da UFLA, aplicando-se o algoritmo desenvolvido no capítulo 2. No capítulo 4, serão analisados os 6 vestibulares conjuntamente, comparando-se as provas e os cursos.

REFERÊNCIAS

ANDRADE, D. F. Comparando desempenhos de grupos de alunos por intermédio da teoria de resposta ao item. **Estudos em Avaliação Educacional**, São Paulo, v. 23, n. 1, p. 31-69, jan./jun. 2001.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. **Teoria da resposta ao item**: conceitos e aplicações. São Paulo: Associação Brasileira de Estatística - SINAPE, 2000.

AZEVEDO, C. L. N. **Métodos de estimação na teoria de resposta ao item**. 2003. 121 p. Dissertação (Mestrado em Estatística)-Universidade de São Paulo, São Paulo, 2003.

_____. **Modelos longitudinais de grupos múltiplos multiníveis na teoria de resposta ao item**: métodos de estimação e seleção estrutural sob uma perspectiva bayesiana. 2008. 265 p. Tese (Doutorado em Estatística)-Universidade de São Paulo, São Paulo, 2008.

BAKER, F. B. **The basics of item response theory**. 2^{nd}. ed. Wisconsin: University of Wisconsin, 2001.

BAQUERO. G. **Testes psicométricos e projetivos**: esquemas para construção, análise e avaliação. São Paulo: Edições Loyola, 1968.

BINET, A.; SIMON, T. Le développement de l'intelligence chez les enfants. **Année Psychologique**, France, v. 14, n. 14, p. 1-94, 1908.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. In: LORD, F. M.; NOVICK, M. R. (Ed.). **Statistical theories of mental test scores**. Reading, MA: Addison-Wesley, 1968. p. 397-549.

BOCK, R. D.; ZIMOWSKI, M. F. Multiple group IRT. In: LINDEN, W. J. van der; HAMBLETON, R. K. (Ed.). **Handbook of modern item response theory**. New York: Spring-Verlag, 1997. p. 433-448.

BOX, G. E. P.; TIAO, G. C. **Bayesian inference in statistical analysis**. New York: J. Wiley, 1992.

BRAGION, M. L. L.; BUENO FILHO, J. S. S. Análise dos candidatos e do vestibular 2006-2, do curso de agronomia da UFLA, usando um modelo de teoria de resposta ao item (TRI). **Revista Matemática e Estatística**, São Paulo, v. 25, n. 3, p. 39-55, 2007.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hasting algorithm. **The American Statistician**, Alexandria, v. 49, n. 4, p. 327-335, Nov. 1995.

GAMERMAN, D.; LOPES, H. **Markov chain Monte Carlo**: stochastic simulation for bayesian inference. 2^{nd}. ed. London: Chapman e Hall/CRC Texts in Statistical Science, 2006.

GELFAND, A. E.; SMITH, A. F. M. Sampling based approaches to calculating marginal densities. **Journal of the American Statistical Association**, Alexandria, v. 85, n. 410, p. 398-409, June 1990.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, Hayward, v. 7, n. 4, p. 457-511, 1992.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibss distributions and the bayesian restoration of images. **IEEE Transactions on Pattern Analysis and Machine Intellence**, Washington, v. 6, n. 6, p. 721-741, Nov. 1984.

GULLIKSEN, H. **Theory of mental tests**. New York: J. Wiley, 1950.

HAMBLETON, R. K.; COOK, L. L. Latent trait models and their use in the analysis of educational test data. **Journal of Educational Measurement**, Washington, v. 14, n. 2, p. 75-96, 1977.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. **Fundamentals of item response theory**. Newbury Park: Sage Publications, 1991.

HASTINGS, W. K. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika**, London, v. 57, n. 1, p. 97-109, Mar. 1970.

JEFREYS, H. **Theory of probability**. Oxford: Clarendon, 1961.

LAWLEY, D. N. On problems connected with item selection and test construction. **The Royal Society of Edinburgh**, Edinburgh, v. 61-A, n. 2, p. 273-287, 1943.

LAZARSFELD, P. F. The logical and mathematical foundation of latent structure analysis. In: STAUFFER, S. A. et al. **Measurement and prediction**. Princeton, NJ: Princeton Univerty, 1950. v. 4, p. 362-412.

LORD, F. M. **Applications of item response theory to practical testing problems**. Hillsdale, NJ: Erlbaum, 1980.

_____. **A theory of test scores**. Iowa City, IA: Psychometric Society, 1952.

METROPOLIS, N. et al. Equations of state calculations by fast computing machines. **Journal of Chemical Phisics**, New York, v. 21, n. 6, p. 1087-1092, June 1953.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. 3^{nd}. ed. Tokio: McGraw-Hill, 1974.

NOGUEIRA, D. A.; SÁFADI, T.; FERREIRA, D. F. Avaliação de critérios de convergência para o método de Monte Carlo via cadeias de Markov. **Revista Brasileira de Estatística**, Rio de Janeiro, v. 65, n. 224, p. 59-88, 2004.

O'HAGAN, A. **Kendall's advanced theory of statistics**. London: Ed. Arnold, 1994. (Bayesian Inference, v. 2B).

PASQUALI, L.; PRIME, R. Fundamentos da teoria de resposta ao item - TRI. **Avaliação Psicológica**, Porto Alegre, v. 2, n. 2, p. 99-110, 2003.

PATZ, R. J.; JUNKER, B. W. Applications and extensions of MCMC in IRT: multiple item types, missing data and rated responses. **Journal of Educational and Behavioral Statistics**, New York, v. 24, n. 4, p. 342-366, Winter 1999.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003.

PINHEIRO, C. E. **Implementação de métodos estatísticos para avaliação educacional no software R**. 2006. 151 p. Dissertação (Mestrado em Estatística)-Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, 2006.

RAFTERY, A. L.; LEWIS, S. Comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. **Statistical Science**, Hayward, v. 7, n. 4, p. 493-497, 1992.

RASCH, G. **Probabilistic models for some intelligence and attainment tests**. Copenhagen: Danish Institute for Educational Research, 1960.

RICHARDSON, M. W. Notes on the rationale of item analysis. **Psychometrika**, Princeton, v. 1, n. 1, p. 69-76, 1936.

SAMEJIMA, F. A. **Estimation of latent ability using a response pattern of graded scores**. 1969. Disponível em: <http://www.psychometrika.org/journal/online/MN17.pdf>. Acesso em: 09 jul. 2009.

THURSTONE, L. L. Attitude can be measured. **American Journal of Sociology**, Chicago, v. 33, n. 4, p. 529-554, Jan. 1928.

TUCKER, L. R. Maximum validity of a test with equivalent items. **Psychometrika**, Princeton, v. 11, p. 1-13, Mar. 1946.

VALLE, R. C. **Teoria de resposta ao item**. 1999. 99 p. Dissertação (Mestrado em Estatística)-Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, 1999.

YERKES, R. M. **Psychological examining in the United States Army**. Washington: Government Printing Office, 1921. (Memoirs of the National Academy of Sciences, v. 15).

CAPÍTULO 2

Avaliação do programa para a análise do modelo de três parâmetros da Teoria de Resposta ao Item

RESUMO

Realizou-se um estudo de simulação para os parâmetros dos itens e das habilidades em um modelo logístico de três parâmetros, da teoria de resposta ao item. Nove tipos de testes com todas as combinações de 50, 100 e 200 itens com 50, 100 e 200 indivíduos foram simulados em oito repetições de cada configuração. Foram simulados dois grupos de provas diferentes em 10% dos itens. Procedeu-se à análise *bayesiana* usando o algoritmo Metropolis-Hastings para obter amostras da distribuição a *posteriori* dos parâmetros. Calcularam-se medidas de correlação, viés, erro quadrático médio (EQM) e tempo de ajuste. Os resultados obtidos indicam altas correlações e mais baixo viés e EQM, melhorando consistentemente com o aumento do tamanho amostral. Para estimar restrições no tempo de execução, foram ajustadas superfícies de resposta em função do número de itens e estudantes. A função R implementada mostrou-se adequada para a análise de exames do tipo vestibular de tamanho moderado, com centenas de itens e poucos milhares de candidatos e em computadores pessoais.

Palavras-chave: Correlação. Simulação. Viés.

ABSTRACT

It was conducted a simulation study for the parameters of the items and abilities in three parameters logistic model of item response theory. Nine types of tests with all combinations of 50, 100 and 200 items with 50, 100 and 200 individuals were simulated in eight repetitions of each configuration. It was simulated two groups of different tests in 10% of items. It proceeded to the Bayesian analysis using the Metropolis-Hastings algorithm to obtain sample of the joint posterior distribution of all parameters. It was calculated measures of correlation, bias and mean square error (MSE) and time of tuning. The results indicate high correlation and lower bias and MSE, consistently improving with increasing samples size. For estimate the execution time constrains, response surfaces were adjusted in function of number of items and students. The R function implemented proved adequate for analysis of tests of entrance examinations (*vestibular*) type of moderate size, with hundreds of items and few thousand of candidates and personal computer.

Keywords: Bias. Correlation. Simulation.

1 INTRODUÇÃO

O modelo logístico de três parâmetros (ML3P) é o mais comumente utilizado em TRI.

Em geral, os *softwares* disponíveis não manejam bem situações com dados desbalanceados, isto é, provas em que nem todos os itens são comuns ao grupo de respondentes, e que se queira obter os resultados por meio da metodologia *bayesiana*, além do dispêndio computacional exigido que será muito maior. Assim sendo, se faz necessário o aprimoramento de algoritmos e programas que facilitem tais cálculos, tornando-os rotineiros em análise de dados de vestibular e congêneres.

Além da necessidade de que melhores e mais eficazes algoritmos com facilidade de implementação e redução de custos, deve-se pensar também em sua implementação (os *softwares* que serão empregados). A elaboração de programas para a utilização em *softwares* livres implica em uma contribuição muito mais abrangente (e em geral o R vem sendo uma solução para a comunicação científica e a difusão de técnicas).

Assim sendo, neste capítulo foi feito um estudo de simulação para os parâmetros dos itens e das habilidades em um modelo logístico de três parâmetros da teoria de resposta ao item (TRI), combinando no mesmo modelo dois tipos de provas diferentes. Empregando-se a metodologia *bayesiana*, elaborou-se uma rotina baseada no método de simulação Monte Carlo via cadeia de Markov (MCMC) de tal forma que todos os itens (que compõe as muitas provas) e indivíduos pudessem ser analisados conjuntamente. Considerando a implicação que um *software* livre possa ter no sentido de facilitar a expansão desses estudos, essa rotina foi escrita em linguagem C para ser utilizada no *software* R. Maiores detal-

hes sobre o algoritmo podem ser solicitados via *e-mail*: <lourdinha.bragion@gmail.com> ou <fmcron@gmail.com>.

O objetivo do estudo relatado neste capítulo é, pois, verificar se a função R, implementada para esse modelo, pode ser considerada adequada para a análise de exames do tipo vestibular em que tal situação ocorre. Demais questões relacionadas a ele, como facilidade de implementação e rapidez de execução computacional também fizeram parte desse objetivo.

Na próxima seção será apresentado o referencial teórico e em seguida, a metodologia utilizada. Na sequência, têm-se os resultados obtidos, as conclusões e, por último, a bibliografia utilizada.

2 REFERENCIAL TEÓRICO

O crescimento e a divulgação da TRI encontram-se diretamente ligados ao progresso das máquinas e ao desenvolvimento de *softwares* apropriados que possibilitem e viabilizem os cálculos que seus modelos exigem. Isso começou a ocorrer a partir dos anos 80. O primeiro *software* para esse fim foi o BICAL de Wright, Mead e Bell (1979), seguido depois pelo LOGIST (WINGERSKY; BARTON; LORD, 1982) e pelo BILOG (MISLEVY; BOCK, 1984), tendo este último tornado-se uma referência na área.

Os programas anteriormente citados oferecem opções de cálculo para estimação dos parâmetros por meio da máxima verossimilhança conjunta (MVC) e máxima verossimilhança marginal (MVM). Nessa metodologia, por meio do uso do algoritmo EM (BAKER, 1992), primeiro se estimam os parâmetros dos itens e, em seguida, considerando-os como conhecidos e fixos, estimam-se os das habilidades. Entretanto, como afirmam Patz e Junker (1999b), apesar dessa metodologia ter sido fundamental para o sucesso das implementações em TRI, à medida que se aumenta a complexidade dos modelos de TRI, as aplicações do EM se tornam mais complexas, fazendo-se necessários métodos mais rápidos e de fácil implementação. Patz e Junker (1999a, b) discutem as vantagens da utilização do método de simulação Monte Carlo via cadeia de Markov (MCMC) para resolver problemas complexos em psicometria, tanto em termos de facilidade de implementação - e consequentemente em custos computacionais - como em relação à flexibilidade para incorporar modelos mais complexos.

O método MCMC é um método iterativo para gerar amostras de uma distribuição de interesse. A principal vantagem do MCMC sobre o BILOG é que ele oferece uma boa aproximação da distribuição à *posteriori* conjunta para os

parâmetros, enquanto que o BILOG produz apenas estimativas pontuais desses parâmetros (JONES; NEDIAK, 2000). A distribuição a *posteriori* é o elemento fundamental que serve de base ao desenvolvimento de toda a inferência *bayesiana* (O'HAGAN, 1994). No entanto, amostrar diretamente dessa distribuição quase nunca é possível, pois, dificilmente ela possui uma forma conhecida. Amostras dessa distribuição são, então, tiradas indiretamente pelo uso do algoritmo MCMC através de uma distribuição de transição da cadeia que deve convergir para uma distribuição estacionária que seja a própria distribuição de interesse (CHIB; GREENBERG, 1995; TIERNEY, 1994). Com essa amostra é possível calcular resumos estatísticos (como a média, desvio padrão) e outros diagnósticos (JOHNSON; JUNKER, 2003; PATZ; JUNKER, 1999a). Como exemplo de um *software* que permite a implementação de modelos para se fazer inferência *bayesiana* via MCMC pode-se citar o WinBugs (SPIEGELHLTER; THOMAS; BEST, 1999).

Gelfand e Smith (1990) popularizaram uma versão do Monte Carlo via cadeias de Markov (MCMC) chamada amostrador de Gibbs, AG (GEMAN; GEMAN, 1984). O AG foi aplicado pela primeira vez em problemas de TRI por Albert (1992). No entanto, somente após Patz e Junker (1999a, b) terem discutido as vantagens do MCMC para resolver problemas complexos em psicometria é que sua utilização em TRI tem recebido crescente atenção (SINHARAY, 2004). Estudos de simulação têm sido realizados para averiguar a eficiência do MCMC em diversos tipos de modelos. Como alguns exemplos podem ser citados Bolt, Cohen e Wollack (2001) para uma mistura de dois tipos de amostragem do modelo de resposta nominal, usando o *software* Winbugs; Segall (2002), para testes adaptativos, também usando o *software* Winbugs; Johnson e Junker (2003), para modelos de respostas *unfolding*, usando o R (R Development Core Team); Torre e Patz (2005), para estudo da multidimensionalidade da habilidade, entre outros.

Uma característica muito importante da TRI é tornar possível a comparação entre indivíduos que realizaram provas diferentes. Para isso é desejável, no entanto, que essas provas possuam itens comuns. A combinação é feita por um processo conhecido como equalização (ANDRADE, 2001). Em Senno (2006) podem ser encontrados os principais métodos de equalização via inferência frequentista.

Pinheiro (2006) apresenta implementações computacionais utilizando a linguagem R para análise de itens e equalização de um teste. No entanto, o procedimento de estimação por ele adotado é o da máxima verossimilhança marginal. É importante, pois, que se tenham rotinas implementadas em *softwares* livres utilizando metodologias *bayesianas*.

2.1 Inferência bayesiana

Na inferência *bayesiana*, toda informação que se tem sobre o parâmetro deve ser considerada e incorporada aos dados, por meio de uma distribuição a *priori*. A distribuição resultante é chamada distribuição a *posteriori* e é com base nela, que é feita toda inferência *bayesiana* (O'HAGAN, 1994).

A distribuição conjunta a *priori* de β será representada por $p(\beta)$, sendo β um vetor de parâmetros correspondente aos parâmetros dos itens e das habilidades. Assumindo-se que os parâmetros são independentes é dada por:

$$p(\beta) = [\prod_{j=1}^{k} p(a_j)p(b_j)p(c_j)][\prod_{i=1}^{n} p(\theta_i)]$$

A distribuição dos dados, também chamada de função de verossimilhança, se refere à distribuição conjunta de $Y = (Y_{11}, ..., Y_{ij}, ..., Y_{nk})$, sendo Y_{ij} a variável aleatória associada ao acerto ou erro na resposta do indivíduo *i* ao item *j*, *i* =

1, ..., n, *j* = 1, ..., k, ($Y_{ij} = 1$, se o indivíduo *i* acerta o item *j* e $Y_{ij} = 0$ se ele erra). Como os modelos de TRI satisfazem a propriedade de *independência condicional*, isto é, para uma dada habilidade, as respostas aos diferentes itens da prova são independentes e, também, de *independência entre as respostas de diferentes indivíduos*, essa distribuição conjunta será representada por:

$$L(\beta; y) = p(y|\beta) = \prod_{i=1}^{n}\prod_{j=1}^{k} \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$$

sendo π_{ij} dado por:

$$\pi_{ij} = P(Y_{ij} = 1|\theta, a, b, c) = c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}}$$

A notação da distribuição a *posteriori* conjunta é dada por $p(\beta|y)$ e é obtida pela aplicação do teorema de Bayes, sendo expressa por:

$$p(\beta|y) \propto L(\beta; y)[\prod_{j=1}^{k} p(a_j)p(b_j)p(c_j)][\prod_{i=1}^{n} p(\theta_i)]$$

Em geral, é difícil obter uma forma analítica fechada para a distribuição a *posteriori*. Na maioria dos casos, as distribuições marginais a *posteriori* para os parâmetros envolvem integrações trabalhosas ou mesmo impossíveis. Assim sendo, utilizaram-se técnicas de MCMC para simular valores para os parâmetros de uma cadeia de Markov cuja distribuição estacionária seja a distribuição a *posteriori* dos parâmetros do modelo.

Dos métodos de simulação que utilizam cadeias de Markov, destacam-se o amostrador de Gibbs - AG - e o algoritmo Metropolis e Hastings - MH (HASTINGS, 1970; METROPOLIS et al., 1953). O AG considera as distribuições condicionais completas a *posteriori* de cada parâmetro para a distribuição de transição

quando essas possuem forma conhecida. Caso contrário, faz-se uso do algoritmo MH, o qual consiste em gerar os valores de uma distribuição de transição arbitrária auxiliar da qual se saiba gerar (distribuição geradora de candidatas). A fim de simplificar a implementação do algoritmo MH, a distribuição geradora de candidatas pode ser escolhida como tendo a mesma forma da distribuição a *priori* dos parâmetros de interesse (CHIB; GREENBERG, 1995; PATZ; JUNKER, 1999a). Maiores detalhes sobre esses dois algoritmos podem ser encontrados em Chib e Greenberg (1995), Gamerman (1997) e Paulino, Turkman e Murteira (2003).

Na próxima seção está descrita a metodologia utilizada no processo de simulação e de estimação dos parâmetros, assim como a forma de implementação do processo amostral e da análise do experimento. Seguem-se a apresentação dos resultados e a discussão.

3 METODOLOGIA

3.1 Processo de simulação

A fim de avaliar o desempenho do algoritmo implementado, foram simulados nove testes referentes a todas as combinações de 50, 100 e 200 itens com 50, 100 e 200 indivíduos num total de oito repetições para cada simulação.

Para gerar o vetor aleatório $Y = [Y_{11}, ..., Y_{ij}]^t$, seguiu-se os seguintes passos: gerou-se um vetor aleatório $\beta = [a_1, ..., a_k, b_1, ..., b_k, c = c_1..., c_k, \theta_1, ..., \theta_n]^t$. Os parâmetros de discriminação (a_j) foram gerados a partir de uma distribuição gamma (5,3); os parâmetros de dificuldade (b_j), a partir de uma distribuição beta usando a expressão: $6 \times (beta(5, 5) - 0, 5)$; os parâmetros para a probabilidade de que um indivíduo com baixa habilidade acerte o item (c_j), a partir de uma distribuição beta(2,5); e os parâmetros de habilidade (θ_j), a partir de uma distribuição normal padronizada. A seguir, Y_{ij} foi gerado a partir de uma distribuição Binomial$(1, \pi_{ij})$.

Em cada teste, deixou-se que apenas 90% dos itens fossem respondidos por todos os indivíduos. Para os outros 10%, foram elaborados dois grupos de questões diferentes de forma que 50% dos indivíduos respondessem a um grupo e os outros 50% ao outro grupo de questões. Dessa forma, foram originados dois tipos de provas diferentes. Para diferenciar quais indivíduos responderam a cada um deles, considerou-se uma matriz de delineamento $X = \{x_{ij}\}$ para as respostas como provinda de um ensaio desbalanceado, sendo $x_{ij} = 1$ se o indivíduo *i* respondeu ao item *j* e $x_{ij} = 0$ caso contrário. Assim sendo, a função de verossimilhança

passou a ser:

$$p(y|\theta, a, b, c, x) = \prod_{i=1}^{n}\prod_{j=1}^{k} \pi_{ij}^{y_{ij}x_{ij}}(1-\pi_{ij})^{(1-y_{ij})x_{ij}}$$

e a distribuição conjunta a *posteriori*:

$$p(\theta, a, b, c|y, x) \propto p(y|\theta, a, b, c, x)[\prod_{j=1}^{k} p(a_j)p(b_j)p(c_j)][\prod_{i=1}^{n} p(\theta_i)]p(x)$$

sendo $p(x) = 1$, pois é um delineamento conhecido. Devido a isso, essa expressão continuará sendo escrita da mesma forma como anteriormente, ou seja:

$$p(\beta|y) \propto L(\beta; y)[\prod_{j=1}^{k} p(a_j)p(b_j)p(c_j)][\prod_{i=1}^{n} p(\theta_i)]$$

A estimação desses parâmetros foi feita via inferência *bayesiana* usando os procedimentos MCMC.

3.1.1 Escolha das distribuições a *priori*

A distribuição a *priori* para cada elemento do parâmetro θ foi uma distribuição normal (0,1), pois admite-se que as características dessa distribuição estão de acordo com as da população que realiza uma prova.

Para cada elemento do parâmetro a, adotou-se uma distribuição log-normal (0;0,5), pois, admitindo-se valores negativos para eles, a probabilidade de responder corretamente ao item diminuiria com o aumento da habilidade, o que se tornaria incoerente. Os valores 0 e 0,5 refletem uma média de 1 e um desvio padrão também de 1. Segundo Baker (2001), valores acima de 1,70 para esse parâmetro são considerados como tendo discriminação muito alta. Uma log-normal (0;0,5)

possui 91% de seus valores até 2 e 98% até 3. Dessa forma, com a escolha dessa distribuição a *priori* não se espera obter muitos itens com valores de $a_j > 3$. Segundo Harwell e Janosky (1991), variâncias das *prioris* para os parâmetros dos itens têm pouco efeito nas estimativas dos mesmos quando o número de indivíduos é maior que 250.

Para cada elemento do parâmetro b, atribuiu-se uma distribuição uniforme (-3,3). O objetivo foi que pouca informação *a priori* fosse refletida sobre esse parâmetro. Isso porque, em uma prova, não se conhecem os padrões adotados para elaborar o número de itens quanto a seu grau de dificuldade. Assim, apesar do domínio de b estar relacionado ao de θ, não é razoável supor que tenha uma distribuição normal. A escolha do intervalo [-3,3], no entanto, reflete uma cobertura de 99,87% da distribuição das habilidades e o fato de se ter escolhido uma distribuição a *priori* normal (0,1) para cada elemento de θ minimiza a expectativa de que $\theta < -3$ e $\theta > 3$ ocorram.

Para cada elemento do parâmetro c foi escolhida a distribuição beta (2,4), pois, tratando-se de uma probabilidade, a informação de que se dispõe, é que seu valor encontra-se no intervalo (0,1). A escolha dos valores para seus hiperparâmetros se deu devido ao fato de que eles refletem numa moda igual à $0,25$, que é consistente com o que se espera num estimador de máxima verossimilhança quando se tem questões com 4 alternativas.

Nas simulações, os valores de cada parâmetro foram gerados a partir de distribuições diferentes das escolhidas para as distribuições a *priori*. O objetivo foi testar a eficiência do algoritmo em recuperar o valor paramétrico.

3.1.2 Distribuição conjunta a *posteriori* e distribuições condicionais completas

A distribuição conjunta a *posteriori* será:

$$p(\beta|y) \propto L(\beta;y)[\prod_{j=1}^{k} p(a_j)p(b_j)p(c_j)][\prod_{i=1}^{n} p(\theta_i)]$$

$$p(\beta|y) \propto \prod_{i}^{n}\prod_{j}^{k} \frac{1}{a_j\sqrt{\pi/2}} e^{-\frac{1}{2}[2ln(a_j)]^2} . \frac{1}{6} . \frac{1}{B(2,4)} c_j(1-c_j)^3 . \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta_i^2}$$

$$\left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}}\right]^{y_{ij}} \left\{1 - \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}}\right]\right\}^{(1-y_{ij})}$$

$$p(\beta|y) \propto \prod_{i}^{n}\prod_{j}^{k} \frac{c_j(1-c_j)^3}{a_j} . \frac{e^{-\frac{1}{2}[4ln^2(a_j)+\theta^2]}}{B(2,4)} .$$

$$\left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}}\right]^{y_{ij}} \left\{1 - \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}}\right]\right\}^{(1-y_{ij})}$$

As distribuições condicionais completas de cada parâmetro estão dadas a seguir.

i) Para o parâmetro a_j:

$$p(a|\theta,b,c,y) = \frac{p(\theta,a,b,c|y)}{\int p(\theta,a,b,c|y)da} =$$

$$= \frac{\prod_i^n \prod_j^k p(\theta_i)p(a_j)p(b_j)p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}}{\int \prod_i^n \prod_j^k p(\theta_i)p(a_j)p(b_j)p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}da_j} =$$

$$= \frac{\prod_i^n p(a_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}}{\int \prod_i^n p(a_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}da_j} \propto \prod_i^n p(a_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}$$

$$p(a|\theta,b,c,y) \propto \prod_i^n \frac{1}{a_j\sqrt{\pi/2}} e^{-\frac{1}{2}[2ln(a_j)]^2}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}$$

A distribuição condicional resultante é dada explicitamente por:

$$p(a|\theta,b,c,y) \propto \prod_i^n \frac{e^{-2ln^2(a_j)}}{a_j}.$$

$$\cdot \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i-b_j)}}\right]^{y_{ij}} \left\{1 - \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i-b_j)}}\right]\right\}^{(1-y_{ij})}$$

ii) Para o parâmetro b_j:

$$p(b|\theta,a,c,y) = \frac{p(\theta,a,b,c|y)}{\int p(\theta,a,b,c|y)db} =$$

$$= \frac{\prod_i^n \prod_j^k p(\theta_i)p(a_j)p(b_j)p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}}{\int \prod_i^n \prod_j^k p(\theta_i)p(a_j)p(b_j)p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}db_j} =$$

$$= \frac{\prod_i^n p(b_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}}{\int \prod_i^n p(b_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}db_j} \propto \prod_i^n p(b_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}$$

$$p(b|\theta,a,c,y) \propto \prod_i^n \frac{1}{6}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}$$

Que pode ser reescrita como:

$$p(b|\theta,a,c,y) \propto \prod_i^n \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i-b_j)}}\right]^{y_{ij}} .$$

$$. \left\{1 - \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i-b_j)}}\right]\right\}^{(1-y_{ij})}$$

iii) Para o parâmetro c_j:

$$p(c|\theta,a,b,y) = \frac{p(\theta,a,b,c|y)}{\int p(\theta,a,b,c|y)dc} =$$

$$= \frac{\prod_i^n \prod_j^k p(\theta_i)p(a_j)p(b_j)p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}}{\int \prod_i^n \prod_j^k p(\theta_i)p(a_j)p(b_j)p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}dc_j} =$$

$$= \frac{\prod_i^n p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}}{\int \prod_i^n p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}dc_j} \propto \prod_i^n p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}$$

$$p(c|\theta,a,b,y) \propto \prod_i^n \frac{1}{B(2,4)}c_j(1-c_j)^3\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}$$

Cuja forma final é dada por:

$$p(c|\theta,a,b,y) \propto \prod_{i}^{n} \frac{c_j(1-c_j)^3}{B(2,4)}.$$

$$\cdot \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i-b_j)}}\right]^{y_{ij}} \left\{1 - \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i-b_j)}}\right]\right\}^{(1-y_{ij})}$$

iv) Para o parâmetro θ_i:

$$p(\theta|a,b,c,y) = \frac{p(\theta,a,b,c|y)}{\int p(\theta,a,b,c|y)d\theta} =$$

$$= \frac{\prod_i^n \prod_j^k p(\theta_i)p(a_j)p(b_j)p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}}{\int \prod_i^n \prod_j^k p(\theta_i)p(a_j)p(b_j)p(c_j)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}d\theta_i} =$$

$$= \frac{\prod_j^k p(\theta_i)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}}{\int \prod_j^k p(\theta_i)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}d\theta_i} \propto \prod_j^k p(\theta_i)\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}$$

$$p(\theta|a,b,c,y) \propto \prod_j^k \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta_i^2}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{(1-y_{ij})}$$

Que tem forma final:

$$p(\theta|a,b,c,y) \propto \prod_j^k e^{-\frac{1}{2}\theta_i^2}.$$

$$\cdot \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}}\right]^{y_{ij}} \left\{1 - \left[c_j + (1-c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}}\right]\right\}^{(1-y_{ij})}$$

Como nenhuma dessas condicionais têm forma conhecida, para obter amostras da distribuição a *posteriori* de interesse foi utilizado o algoritmo MH.

3.2 Implementação do processo amostral

Foi gerada uma cadeia inicial com 550.000 iterações para todos os parâmetros, das quais descartaram-se os primeiros 50.000 valores a fim de que a influência do "chute" inicial pudesse ser anulada (*burn-in*). A seguir, os pontos amostrais foram tomados de 100 em 100 iterações (*thinning*), de forma que pudesse ser obtida uma amostra aleatória. Assim, a amostra final constou de 5.000 observações.

Para implementar o algoritmo MH, foi escolhida como distribuição geradora de candidatas a mesma da distribuição a *priori* dos parâmetros de interesse. Isso foi feito de forma a simplificar a implementação do algoritmo.

3.3 Análise do experimento

As medidas utilizadas para avaliação dos resultados obtidos pelo processo de simulação foram as medidas de correlação (ρ), erro quadrático médio (EQM) e viés em 8 blocos completos casualizados. O comportamento de cada uma dessas medidas foi expresso em função do número de itens e indivíduos por meio do seguinte modelo estatístico:

$$\widehat{Z}_{ij} = \alpha_0 + \alpha_1 n + \alpha_{11} n^2 + \alpha_2 p + \alpha_{22} p^2 + \alpha_{12}(n \times p) + \varepsilon_{ij}$$

em que:

a) $\widehat{Z}_{ij}$: representa as variáveis dependentes $\rho(a, \widehat{a})$ ou $\rho(b, \widehat{b})$ ou $\rho(c, \widehat{c})$ ou $\rho(\theta, \widehat{\theta})$ ou $\widehat{EQM_{ij}}$ ou $\widehat{vis}_{ij}$, conforme esteja-se avaliando as variáveis correlação entre os valores paramétricos e os valores estimados dos parâmetros dos itens e dos indivíduos, EQM ou viés;

b) α_0: representa uma constante;

c) *n*: número de indivíduos;

d) *p*: número de itens;

e) α_1: coeficiente de regressão da variável independente n;

f) α_{11}: coeficiente de regressão da variável independente n^2;

g) α_2: coeficiente de regressão da variável independente p;

h) α_{22}: coeficiente de regressão da variável independente p^2;

i) α_{12}: coeficiente de regressão para a interação das variáveis independentes n e p;

j) ε_{ij}: erros aleatórios associados ao ajuste do modelo.

O tempo de execução das simulações foi analisado em função das combinações itens $\times$ indivíduos através de superficie de resposta, seguindo o mesmo modelo descrito acima.

Os programas utilizados para a simulação, implementação da metodologia *bayesiana* e análise dos resultados foram elaborados com o código escrito em linguagem C, com chamada no R 2.9.0 (R DEVELOPMENT CORE TEAM, 2009). Foi utilizado um processador Core i7 - 965 - 3.20 ghz com 12 gb de memória RAM.

4 RESULTADOS

Nas Figuras 1 e 2 encontram-se representadas, respectivamente, as correlações entre o valor paramétrico e o valor estimado dos parâmetros dos itens e das habilidades e entre a nota e o valor paramétrico para as combinações dos 9 tipos de testes e seus respectivos intervalos de confiança a 95%.

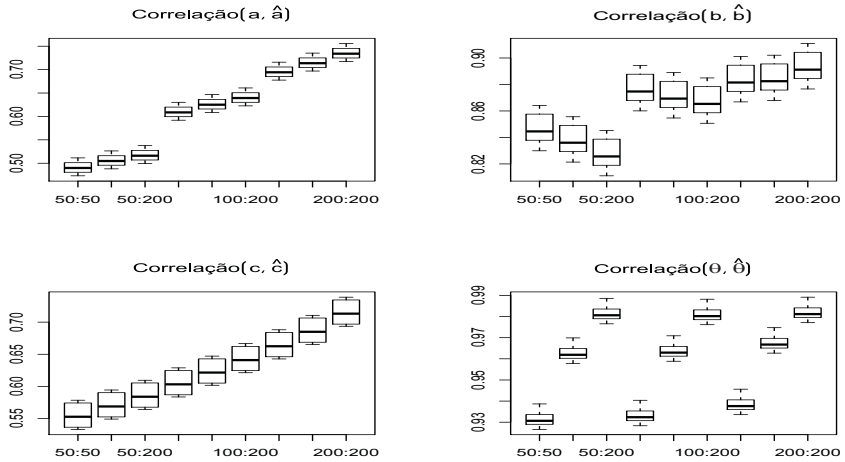


Figura 1 Correlações entre os valores paramétricos (a, b, c, θ) e os valores estimados ($\hat{a}$, $\hat{b}$, $\hat{c}$, $\hat{\theta}$) e respectivos intervalos de confiança a 95%. No eixo das abscissas, o primeiro valor se refere ao número de indivíduos e o segundo, ao número de itens

Observa-se que foram obtidos valores satisfatórios para as correlações entre o valor paramétrico e o valor estimado para os parâmetros dos itens. Como esperado, a correlação cresce com o aumento do número de indivíduos. Quanto às habilidades, houve alta correlação com o valor paramétrico. Pode-se observar, também, que para os parâmetros dos itens, a correlação aumenta conforme aumenta o número de indivíduos. E para os parâmetros da habilidade, a correlação aumenta conforme aumenta o número de itens. Isto mostra que os parâmetros dos itens são mais bem estimados quando se tem um número maior de indivíduos rea-

lizando o teste, e os parâmetros das habilidades são mais bem estimados quando se aplica um número maior de itens no teste.

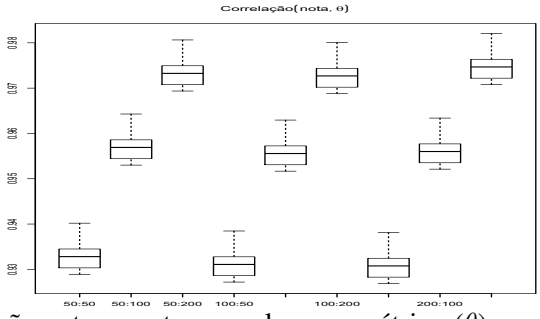


Figura 2 Correlação entre a nota e o valor paramétrico (θ) e respectivo intervalo de confiança a 95%. No eixo das abscissas, o primeiro valor se refere ao número de indivíduos e o segundo, ao número de itens

A partir da Figura 2, pode-se observar que a correlação entre a nota tradicional e a habilidade foi também bastante alta, semelhante à correlação entre a habilidade paramétrica e a estimada, sendo essa correlação crescente em função do número de itens.

As estimativas do erro quadrático médio (EQM) e do viés para cada parâmetro e seus respectivos intervalos de confiança estão representadas nas Figuras 3 e 4.

Por meio desses resultados, pode-se observar que houve boas aproximações dos valores estimados em relação ao valor paramétrico, pois foram obtidos baixos valores para o EQM e o viés. Pode-se observar também que esses valores tendem a zero à medida que o número de itens e indivíduos cresce. Para a simulação realizada, considerou-se o número máximo de 200 indivíduos realizando a prova. No entanto, nas avaliações em geral, esse número é bem mais elevado, o que implicará em uma consistente melhora nas estimativas.

Nas Tabelas 1 a 3 têm-se os resultados da análise de regressão para as variáveis de correlação, EQM e viés.

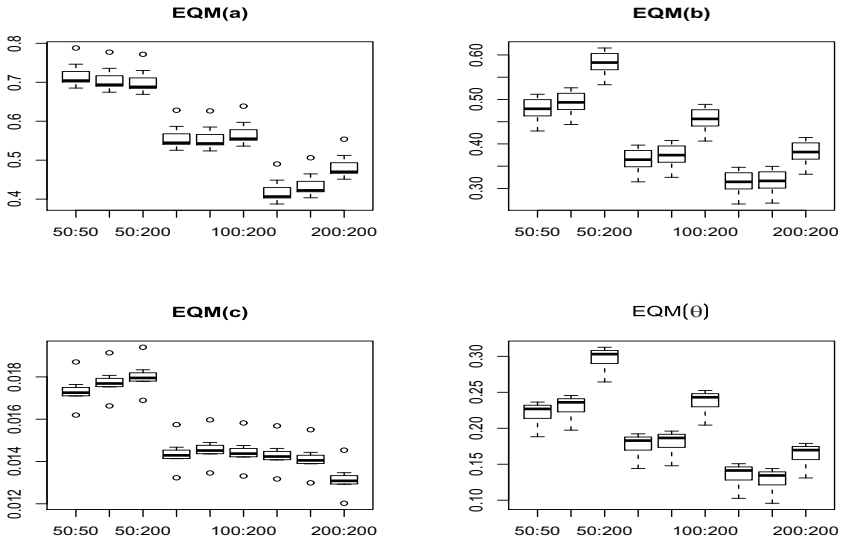


Figura 3 Estimativa do erro quadrático médio (EQM) e respectivo intervalo de confiança a 95 %para cada parâmetro (a, b c , θ). No eixo das abscissas, o primeiro valor se refere ao número de indivíduos e o segundo, ao número de itens

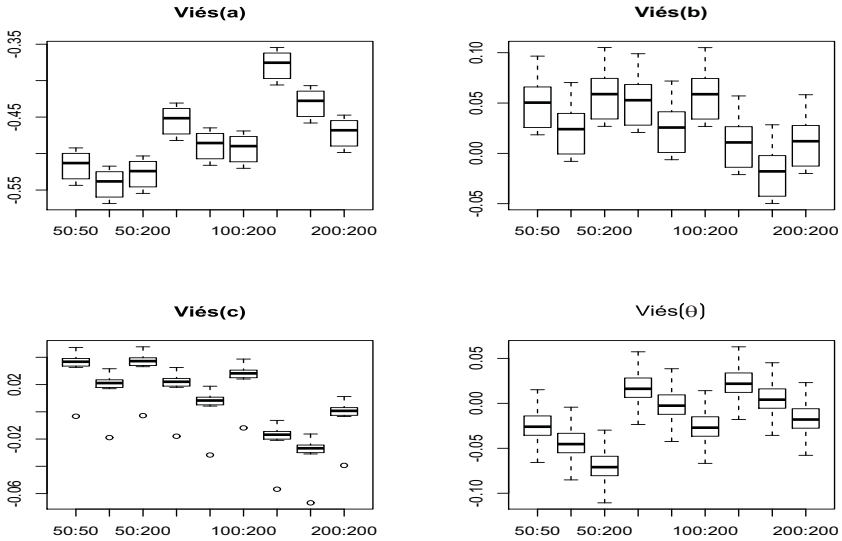


Figura 4 Estimativa do viés e respectivo intervalo de confiança a 95% para cada parâmetro (a, b c , θ). No eixo das abscissas, o primeiro valor se refere ao número de indivíduos e o segundo, ao número de itens

Tabela 1 Coeficientes para o modelo quadrático de superfície de resposta nas variáveis de correlação. (*n*: número de indivíduos; *p*: número de itens)

Coef	$\rho(a, \hat{a})$	$\rho(b, \hat{b})$	$\rho(c, \hat{c})$	$\rho(\theta, \hat{\theta})$	$\rho(nota, \theta)$
μ	**3,22e-01*****	**8,19e-01*****	**4,95e-01*****	**8,91e-01*****	**9,09e-1*****
n	**3,86e-03*****	**1,07e-03***	1,38e-03	3,03e-05	-7,33e-05
n^2	**-1,01e-05****	-3,57e-06	-2,78e-06	1,22e-07	2,08e-07
p	4,54e-04	-3,03e-04	4,47e-04	**1,08e-03*****	**7,93e-04*****
p^2	-1,24e-06	4,51e-07	-1,13e-06	**-2,92e-06*****	**-2,12e-06*****
$n * p$	5,97e-07	1,27e-06	8,71e-06	-2,86e-07	1,54e-07
QME	4,97e-03	1,31e-03	6,70e-03	1,19e-03	1,14e-04

***** significativo a 0,05%, ****** significativo a 0,01%, ******* significativo a 0,001%, pelo teste F

Pode-se observar que a correlação para os parâmetros a e b está em função do número de indivíduos e a correlação para o parâmetro da habilidade, em função do número de itens. Esse fato já foi constatado quando da observação das Figuras 1 e 2 acima e trata-se de um resultado coerente, pois para estimar parâmetros de itens necessita-se que haja indivíduos para respondê-los. Consequentemente depende do número de indivíduos que realizam a prova. Da mesma forma, a estimação das habilidades desses indivíduos vai depender do número de questões que compõem a avaliação.

Tabela 2 Coeficientes para o modelo quadrático de superfície de resposta do EQM. (*n*: número de indivíduos; *p*: número de itens)

Coef	EQM(a)	EQM(b)	EQM(c)	EQM(θ)
μ	**9,53e-01*****	**6,34e-01*****	**2,03e-02*****	**2,71e-01*****
n	**-5,19e-03****	**-3,99e-03*****	**-1,14e-04*****	-1,24e-03
n^2	1,21e-05	**1,19e-05****	**3,91e-07****	3,11e-06
p	-5,47e-04	-2,42e-04	1,87e-05	-2,01e-04
p^2	1,05e-06	4,06e-06	-4,00e-08	3,26e-06
$n * p$	3,55e-06	-1,64e-06	-8,19e-08	-2,13e-06
QME	1,62e-02	4,82e-03	8,25e-06	2,59e-03

***** significativo a 0,05%, ****** significativo a 0,01%, ******* significativo a 0,001%, pelo teste F

Para o EQM, observa-se que houve diferenças significativas apenas para os

parâmetros dos itens, tendo eles a tendência de se aproximarem de zero conforme aumenta-se o número de indivíduos que realizam o teste. Esse fato é relevante, pois, considerando que na maioria das avaliações o número de candidatos é bem maior, a tendência é que o valor estimado se aproximará muito do verdadeiro valor paramétrico.

Tabela 3 Coeficientes para o modelo quadrático de superfície de resposta do viés. (*n*: número de indivíduos; *p*: número de itens)

Coeficientes	viés(a)	viés(b)	viés(c)	viés(θ)
μ	**-5,47e-01*****	1,08e-01	7,93e-01	-6,43e-02
n	1,88e-03	5,30e-04	-2,36e-04	1,62e-03
n^2	-3,14e-06	-3,11e-06	-6,35e-07	-5,26e-06
p	-9,59e-04	-1,38e-03	-8,25e-04	-5,31e-04
p^2	4,27e-06	5,83e-06	3,16e-06	8,79e-07
$n * p$	-3,63e-06	-3,24e-07	7,57e-07	2,28e-07
QME	5,80e-03	1,43e-02	2,18e-03	7,47e-03

***** significativo a 0,05%, ****** significativo a 0,01%, ******* significativo a 0,001%, pelo teste F

O viés não apresentou nenhum coeficiente significativo e, como foi observado na Figura 4, seus valores foram muito baixos em todas as situações simuladas.

Na Tabela 4 tem-se os resultados da análise de regressão para a variável tempo de execução do programa.

Tabela 4 Coeficientes para o modelo quadrático de superfície de resposta para o tempo de execução. (*n*: número de indivíduos; *p*: número de itens)

Coeficientes	estimativa	erro padrão
μ	502,78	1211,94
n	-15,81	15,22
n^2	0,05	0,06
p	9,54	5,22
p^2	-0,02	0,06
$n * p$	**0,71*****	0,03
QME	1235673	

******* significativo a 0,001% pelo teste F

Pode-se observar que a superfície de resposta para o modelo quadrático da variável tempo de execução do programa, apenas o coeficiente da interação entre indivíduos e itens foi significativo, o qual obteve um valor de 0,71, com um nível de significância de 0,001% pelo teste F. Assim, essa relação será expressa por:

$$tempo = 0,71(\pm 0,03)n \times p$$

sendo *n* o número de indivíduos e *p*: número de itens, com o $tempo$ medido em segundos. Essa relação mostra que o tempo de execução do programa aumenta de forma proporcional com o aumento do número de itens e indivíduos.

5 DISCUSSÃO

Ao se observar os resultados das correlações entre a média estimada e o valor paramétrico, verifica-se que os mesmos foram muito satisfatórios. Valores muito baixos também foram obtidos para o EQM e o viés. Isso indica que o algoritmo recuperou os valores paramétricos com grande sucesso.

Com relação ao tempo de execução do programa, também pode-se verificar que ele cresce de forma proporcional ao número de itens e indivíduos. Como exemplo, considere-se um vestibular com 70 itens em que nem todos são comuns ao grupo de candidatos e, tendo sido realizado por 5000 indivíduos de diferentes cursos. O tempo computacional gasto para gerar uma amostra de 5000 valores para cada parâmetro será de:

$$tempo = 0,71(\pm 0,03'') \times 5000 \times 70 = 248,68'' \pm 11,80''$$

ou

$$tempo \in [65h48'; 72h21']$$

Se, ao invés de 5000 tivéssemos 50.000 indivíduos, o tempo de execução seria 10 vezes maior. Percebe-se que o tempo aumenta de acordo com o tamanho do problema, o que é óbvio, mas a observação a ser feita é que esse aumento ocorre de forma multiplicativa. Isto também indica que a análise combinada de um conjunto grande de provas pode se valer de alguns atalhos, como a aproximação normal para a análise conjunta após a análise individual de cada subgrupo em separado. De forma geral, os vestibulares possuem tamanhos moderados em

que o número de itens não chega a ultrapassar a casa da centena e o número de candidatos, a alguns milhares. Assim, fazendo-se uso apenas de um equipamento de mesa, o programa fornece resultados em tempo moderado.

Estes resultados são considerados fortes fatores a indicar que a metodologia adotada pode ser aplicada a dados reais e que o algoritmo utilizado pode ser considerado adequado e viável, pois trata-se de um programa que utiliza *software* livre e que é relativamente rápido quanto ao tempo gasto para a obtenção dos resultados "corretos".

6 CONCLUSÃO

Pode-se concluir que a função R utilizada fornece estimativas muito próximas dos verdadeiros valores, sendo de fácil execução e relativa rapidez para geração de cadeias de Markov, podendo, portanto, ser considerada adequada para a análise de exames do tipo vestibular de tamanho moderado, com centenas de itens e poucos milhares de candidatos, assim como para outros conjuntos de dados nos quais indivíduos respondam a provas em que nem todos os itens são comuns.

REFERÊNCIAS

ALBERT, J. H. Bayesian estimation of normal ogive item response curves using Gibbs sampling. **Journal of Educational Statistics**, New Jersey, v. 17, n. 3, p. 251-269, 1992.

ANDRADE, D. F. Comparando desempenhos de grupos de alunos por intermédio da teoria de resposta ao item. **Estudos em Avaliação Educacional**, São Paulo, v. 23, n. 1, p. 31-69, jan./jun. 2001.

BAKER, F. B. **Item response theory**: parameter estimation techniques. New York: M. Dekker, 1992.

_____. **The basics of item response theory**. 2^{nd}. ed. Wisconsin: University of Wisconsin, 2001.

BOLT, D. M.; COHEN, A. S.; WOLLACK, J. A. A mixture item response model for multiple-choice data. **Journal of Educational and Behavioral Statistics**, New York, v. 26, n. 4, p. 381-409, Winter 2001.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hasting algorithm. **The American Statistician**, Alexandria, v. 49, n. 4, p. 327-335, Nov. 1995.

GAMERMAN, D. **Markov chain Monte Carlo**: stochastic simulation for bayesian inference. London: Chapman & Hall, 1997.

GELFAND, A. E.; SMITH, A. F. M. Sampling based approaches to calculating marginal densities. **Journal of the American Statistical Association**, Alexandria, v. 85, n. 410, p. 398-409, June 1990.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibss distributions and the bayesian restoration of images. **IEEE Transaction on Pattern Analysis and Machine Intellence**, Washington, v. 6, n. 6, p. 721-741, Nov. 1984.

HARWELL, M. R.; JANOSKY, J. E. An empirical study of the effects of small datasets and varying prior variances on item parameters estimation in bilog. **Applied Psychological Measurement**, Pennsylvania, v. 15, n. 3, p. 279-291, 1991.

HASTINGS, W. K. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika**, London, v. 57, n. 1, p. 97-109, Mar. 1970.

JOHNSON, M. S.; JUNKER, B. W. Using data augmentation and Markov chain Monte Carlo for the estimation of unfolding response models. **Journal of Educational and Behavioral Statistics**, New York, v. 28, n. 3, p. 195-230, 2003.

JONES, D. H.; NEDIAK, M. **Item parameter calibration of LSAT items using MCMC approximation of Bayes posterior distributions**. 2000. Disponível em: <http://rutcor.rutgers.edu/pub/rrr/reports2000/07.pdf>. Acesso em: 09 jan. 2009.

METROPOLIS, N. et al. Equations of state calculations by fast computing machines. **Journal of Chemical Phisics**, New York, v. 21, n. 6, p. 1087-1092, June 1953.

MISLEVY, R. J.; BOCK, R. D. **BILOG**: maximum likelihood item analysis and test scoring logistic models. Chicago: Scientific Software, 1984.

O'HAGAN, A. **Kendall's advanced theory of statistics**. London: Ed. Arnold, 1994. (Bayesian Inference, v. 2B).

PATZ, R. J.; JUNKER, B. W. Applications and extensions of MCMC in IRT: multiple item types, missing data and rated responses. **Journal of Educational and Behavioral Statistics**, New York, v. 24, n. 4, p. 342-366, Winter 1999a.

_____. A straightforward approach to Markov chain Monte Carlo methods for item response models. **Journal of Educational and Behavioral Statistics**, New York, v. 24, n. 2, p. 146-178, Summer 1999b.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003.

PINHEIRO, C. E. **Implementação de métodos estatísticos para avaliação educacional no software R**. 2006. 151 p. Dissertação (Mestrado em Estatística)-Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, 2006.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing, reference index: version 2.9.0. Vienna: R Foundation for Statistical Computing, 2009. Disponível em: <http://www.R0project.org>. Acesso em: 05 jul. 2009.

SEGALL, D. O. An item response model for characterizing test compromise. **Journal of Educational and Behavioral Statistics**, New York, v. 27, n. 2, p. 163-179, Summer 2002.

SENNO, R. M. **Métodos de equalização na teoria clássica e na teoria de resposta ao item**. 2006. 121 p. Dissertação (Mestrado em Estatística)-Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, 2006.

SHINHARAY, S. Experiences with Markov chain Monte Carlo convergence assessment in two psychometrics examples. **Journal of Educational and Behavioral Statistics**, New York, v. 29, n. 4, p. 461-488, Winter 2004.

SPIEGELHALTER, D.; THOMAS, A.; BEST, N. **WinBugs version 1.2**: computer program. Cambridge: Institute of Public Health, 1999.

TIERNEY, L. Markov chains for exploring posterior. **The Annais of Statistics**, Hayward, v. 22, n. 4, p. 1701-1762, Dec. 1994.

TORRE, J.; PATZ, R. J. Making the most of what we have: a practical application of multidimensional item response theory in tes scoring. **Journal of Educational and Behavioral Statistics**, New York, v. 30, n. 3, p. 295-311, 2005.

WINGERSKY, M. S.; BARTON, M. A.; LORD, F. M. **LOGIST user's guide**. Princeton: Educational Testing Service, 1982.

WRIGHT, B. D.; MEAD, R. J.; BELL, S. R. **BICAL**: calibrating items with the Rasch model. Chicago: University of Chicago - School of Education, 1979. (Statistical Laboratory Research Memorandum, n. 23B).

CAPÍTULO 3

Análise dos vestibulares 2006-2 à 2009-1 da UFLA

RESUMO

Seis exames vestibulares da Universidade Federal de Lavras (UFLA), entre 2006 a 2009, foram analisados quanto aos parâmetros dos itens e das habilidades do modelo logístico de 3 parâmetros. As distribuições a *posteriori* de interesse foram obtidas via MCMC (Monte Carlo via cadeia de Markov) desenvolvido no capítulo anterior, o qual foi rodado no programa R, aplicando-se a metodologia *bayesiana*. Também foi utilizada a função de informação de cada item e do teste para detectar quais itens e provas informavam mais sobre os candidatos. Os vestibulares não apresentaram alta probabilidade de acerto por indivíduos com baixa habilidade. Verificou-se que as provas mais difíceis, com grau de dificuldade próximo da média de habilidade do grupo dos melhores candidatos, são mais informativas. Tais resultados podem ser usados tanto na seleção de candidatos quanto no estudo das propriedades dos itens com vistas à futura melhoria da qualidade dos exames.

Palavras-chave: Análise *bayesiana*. Monte Carlo via cadeia de Markov. Teoria de resposta ao item.

ABSTRACT

Six entrance examinations (*vestibular*) at the Federal University of Lavras (UFLA) between 2006 to 2009 were analysed for parameters items and abilities of three parameters logistic model. The posterior distributions of interest were obtained by Markov Chain Monte Carlo (MCMC) algorithm developed in the previous chapter, which was rotate en R program, applying the Bayesian methodology. Is was also used information function of each item and the test to detect which items and tests informed more about candidates. The *vestibular* didn't show a high probability by guess. It was found that the most difficult tests, with degree of difficulty near the average of ability group of the best candidates are more informative. Such results can be used both in the selection of candidates as in the study proprierties of items with a view to improving the quality of future examinations.

Keywords: Bayesian analysis. Item response theory. Monte Carlo Markov chain.

1 INTRODUÇÃO

A Universidade Federal de Lavras (UFLA) está situada no sul de Minas Gerais e foi fundada em 1908 com o nome de Escola de Agricultura de Lavras. Neste ano contava apenas com três alunos. Foi somente em 1994 que foi transformada em Universidade, após ser federalizada em 1963. Atualmente está organizada em 16 departamentos didático-científicos que preparam cerca de 4.000 estudantes de graduação em 22 cursos. Também oferece cursos de mestrado e doutorado. Procuram a UFLA não somente alunos da região como também de outros Estados brasileiros. A Comissão Permanente de Processo Seletivo (COPESE) é o órgão responsável pela seleção dos candidatos.

Selecionar indivíduos que possuam maiores habilidades é o principal objetivo dos exames vestibulares, os quais se tornaram padrão nas Universidades Brasileiras a partir dos anos 70. No entanto, para que esse objetivo possa ser atingido, um importante fator a ser considerado é a qualidade dos itens. Apesar de, nas últimas décadas, a teoria de resposta ao item (TRI) ter sido aplicada com sucesso para construção e análise de diferentes tipos de testes, poucos estudos ainda são encontrados com relação a sua aplicação para questões de vestibular utilizando a metodologia *bayesiana*. Um exemplo de TRI aplicado a dados de vestibular, utilizando essa metodologia, pode ser encontrado em Bragion e Bueno Filho (2007). Porém, o estudo realizado por eles é feito apenas para uma prova isolada (candidatos da Agronomia do vestibular 2006-2 da UFLA), em que foram utilizadas apenas as questões comuns a todos os estudantes a fim de garantir um delineamento balanceado.

Neste capítulo, teve-se como objetivo analisar todas as provas dos vestibulares da UFLA, no período de 2006-2 a 2009-1, por meio da TRI, fazendo uso

da metodologia *bayesiana* e do algoritmo de Monte Carlo via cadeias de Markov (MCMC) (HASTINGS, 1970; METROPOLIS et al., 1953), sendo considerado o modelo logístico de três parâmetros (ML3P). A estimação dos parâmetros do modelo foi feita através do algoritmo desenvolvido no capítulo anterior, o qual foi rodado no programa R. A função de informação do item e do teste também foi usada como uma importante ferramenta para identificação de quais itens e provas informavam mais sobre os candidatos. Espera-se, com esse estudo, estar contribuindo para a melhoria do seu processo seletivo.

Na seção seguinte descreve-se o material e a metodologia utilizada. Segue-se o estudo de cada um dos vestibulares separadamente com seus respectivos resultados e a discussão sobre eles.

2 METODOLOGIA

2.1 Material

Os dados analisados no presente estudos foram gentilmente cedidos pela COPESE-UFLA e referem-se aos vestibulares de 2006-2 a 2009-1. Todos esses seis vestibulares constaram de 66 itens, com 4 alternativas, sendo que os itens de número 27 a 34 se referem à disciplina de Língua Estrangeira com opção entre Inglês ou Espanhol. Foi elaborada uma matriz para as respostas de cada indivíduo aos 66 itens. No entanto, como um grupo de candidatos respondeu às questões de Inglês e outro grupo às de Espanhol, duplicou-se as colunas que se referiam à disciplina de Língua Estrangeira, alterando para 74 o número de itens. A distribuição de itens por disciplina encontra-se na Tabela 1.

Tabela 1 Disciplinas e itens dos vestibulares 2006-2 a 2009-1 da UFLA

Disciplina	**Itens**
Português	1 - 10
Geografia	11 - 18
Histtória	19 - 24
Filosofia	25 - 26
Espanhol	27 - 34
Inglês	35 - 42
Biologia	43 - 50
Física	51 - 58
Matemática	59 - 66
Química	67 - 74

O número de cursos oferecidos e candidatos inscritos foram variados e estão relacionados na Tabela 2.

Para implementação da metodologia e análise dos resultados foi utilizado

Tabela 2 Relação de candidatos por vagas inscritos aos diversos cursos oferecidos pela UFLA para os vestibulares 2006-2 a 2009-1

Cursos	Candidatos(Candidatos/vagas)					
	2006-2	**2007-1**	**2007-2**	**2008-1**	**2008-2**	**2009-1**
Administração (AD)	284(11,1)	380(23,0)	2959(11,4)	388(25,6)	299(8,9)	387(18,4)
Agornomia (AG)	651(9,2)	826(18,1)	580(7,9)	751(17,3)	602(6,7)	807(14,3)
Eng. Alimentos (AL)	229(10,2)	288(19,9)	223(10,1)	288(21,1)	227(10,5)	289(20,6)
Ciênc. Biológicas (CB)	326(15,2)	424(30,1)	316(14,1)	421(32,3)	288(12,9)	371(13,5)
Cinc. Computação (CC)	235(17,9)	252(19,3)	214(9,6)	263(19,6)	176(5,1)	263(11,5)
Ed. Física (ED)	-	284(10,2)	220(5,1)	293(22,6)	52(2,0)	221(7,2)
Eng. Agrícola (EA)	154(6,5)	121(8,2)	113(5,0)	135(9,9)	118(5,3)	138(9,9)
Eng. Florestal (EF)	238(10,8)	285(19,5)	273(12,1)	339(26,7)	271(12,5)	317(23,0)
Física (FS)	-	-	-	-	58(2,2)	44(2,3)
Matemática (MA)	-	153(6,1)	107(3,1)	97(7,4)	70(2,5)	63(3,9)
Med. Veterinária (MV)	604(26,3)	639(47,4)	529(23,5)	663(51,8)	543(23,9)	686(24,4)
Química (QI)	104(4,7)	135(8,7)	118(5,6)	113(8,1)	90(4,0)	108(7,9)
Sist. Informação (SI)	-	280(11,8)	190(5,6)	244(12,0)	207(5,7)	196(8,7)
Zootecnia (ZO)	228(10,0)	266(18,5)	164(7,2)	210(16,0)	155(7,0)	197(7,1)
Ed. Fís. e Esportes (EB)	-	-	-	-	97(3,9)	
Total	3053	4333	3342	4205	3253	4087

um computador equipado com processador Core i7 - 965 - 3.20 ghz com 12 gb de memória RAM.

O algoritmo utilizado foi o implementado no capítulo anterior, programado para rodar na linguagem R (R DEVELOPMENT CORE TEAM, 2009), sendo consumidos 20 dias de processamento para gerar todas as cadeias dos 6 vestibulares.

2.2 Metodologia

O modelo matemático que foi utilizado para a análise da TRI foi o modelo de Birnbaum (1968), chamado de modelo logístico de três parâmetros (ML3P) expresso por:

$$\pi_{ij} = P(Y_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(\theta_i - b_j)}}$$

A expressão gráfica desse modelo é chamada Curva Característica do Item

(CCI). Foram somadas todas as probabilidades computadas com a CCI de todos os itens, obtendo-se a Curva Característica do Teste - CCT (BAKER, 2001), dada pela seguinte expressão:

$$TS_i = \sum_{j=1}^{k} \pi_{ij}(\theta_i)$$

Para quantificar o grau de precisão da estimativa da habilidade de um determinado item e identificar as questões que são realmente relevantes para selecionar os candidatos, foram construídos gráficos da Função de Informação do Item (FII) (ANDRADE; TAVARES; VALLE, 2000). Sua expressão, para o ML3P, é dada por:

$$I_j(\theta_i) = a_j^{\,2} \frac{(1-\pi_{ij}}{\pi_{ij}} [\frac{\pi_{ij} - c_j}{1-c_j}]^2$$

Da mesma forma que para a CCT, foram somadas todas as probabilidades computadas com a FII de todos os itens, obtendo-se a Função Informação do Teste (FIT), cuja finalidade foi de verificar o quanto o teste, como um todo, foi informativo (BAKER, 2001). É dada pela expressão:

$$I(\theta_i) = \sum_{j}^{k} I_j(\theta_i)$$

Uma descrição mais detalhada sobre cada uma dessas funções referidas encontra-se na seção 2 do capítulo 1.

A variável observada é o acerto ou erro em cada resposta de cada candidato a cada item. Cada uma dessas respostas foi modelada seguindo uma densidade de *Bernoulli*. O modelo de TRI adotado foi o ML3P. A descrição completa do modelo estatístico adotado, da verossimilhança, das prioris e da implementação da amostragem para a análise bayesiana se encontra na seção 5.2 do capítulo 2.

2.2.1 Implementação do processo amostral

Foram geradas cadeias de Markov para obter uma amostra da distribuição conjunta a *posteriori* a partir das distribuições completas de cada parâmetro (*Gibbs Sampling*). Como não foi possível obter formas algébricas de amostragem direta para as condicionais completas, foi utilizado o algoritmo Metropolis e Hastings (MH), sendo geradas duas cadeias iniciais com 610.000 iterações para cada um dos parâmetros dos itens e das habilidades. Considerou-se um *burn-in* de 10.000 observações e um *thinning* de 100, ficando a amostra final com 6.000 observações.

Para tornar mais simples a forma algébrica da probabilidade de aceitação foi escolhida a distribuição geradora de candidatas como sendo a mesma da distribuição a *priori* dos parâmetros de interesse.

Para análise da convergência foram adotados o critério de Gelman e Rubin (1992) e o diagnóstico de Raftery e Lewis (1992), além de uma análise visual do traço das cadeias.

2.2.2 Análise dos resultados

Para identificação de quais itens eram bons ou não, assim como os que mais informavam sobre os candidatos, foram construídos gráficos da CCI e FII para cada um deles. Esses gráficos foram obtidos calculando-se as probabilidades de acerto em função de θ em cada uma das iterações, ou seja, considerando-se todos os 6.000 valores gerados na amostra para cada parâmetro. A seguir, tomou-se a moda dessas probabilidades para cada θ. Esse procedimento foi realizado para cada item.

O mesmo procedimento foi adotado para gerar a FII.

Em um mesmo gráfico foram plotados a CCI juntamente com seu HPD, a FII e o histograma das habilidades estimadas a fim de melhor visualizar, não so-

mente as características dos itens quanto ao seu grau de dificuldade, poder de discriminação e probabilidade de acerto por indivíduos com baixa habilidade como, também, identificar quais itens seriam os mais interessantes para a população em questão. A fim de proporcionar melhor leitura da escala no ponto máximo da FII e ficar mais harmonioso com o histograma, escolheu-se uma escala diferente para a mesma. Isto foi feito de forma a que a altura da informação máxima estivesse sempre no meio do gráfico, preservando seus valores.

Essas mesmas ferramentas foram utilizadas para a análise das provas. No entanto, como cada prova é composta de vários itens, foi utilizada a CCT e a FIT para cada uma delas. Como o número de questões de cada uma dessas provas é desigual (o que pode ser verificado na subseção 4.1), foram divididos os resultados das fórmulas da CCT e FIT pelo número de itens de cada prova, a fim de torná-las com iguais condições de comparação.

Para cada vestibular foram selecionados três itens com características distintas a fim de exemplificar situações que auxiliem na elaboração de novas provas. O conteúdo de cada um deles está disponível no seguinte endereço:

<http://www.copese.ufla.br/copese/provasAnteriores.asp?tipo=V>.

Para o vestibular como um todo, semelhante ao que foi feito para cada item, foi plotado, em um mesmo gráfico, o histograma das habilidades estimadas, a CCT com seu respectivo HPD e a FIT.

Foram também construídas tabelas com as médias das habilidades estimadas por curso. Para a estimativa dessas habilidades foi considerada a moda da cadeia gerada para cada candidato e a seguir calculada a média dessas estimativas por curso. Essa metodologia foi aplicada a cada vestibular separadamente.

3 RESULTADOS

Nas subseções que se seguem estão representados e comentados os resultados dos vestibulares 2006-2 a 2009-1. Os métodos para avaliação da convergência referidos nas mesmas são o critério de Gelman e Rubin e o diagnóstico de Raftery e Lewis.

Em cada vestibular analisado, diversos itens se destacaram em relação aos demais pelas características apresentadas, tais como: os de maior poder de discriminação, ou os de maior grau de dificuldade, ou os de maior (ou menor) probabilidade de acerto por indivíduos com baixa habilidade. No entanto, como se tornaria longa e exaustiva a descrição de cada um deles, isso foi feito apenas para três itens em cada vestibular, sendo esses itens escolhidos de forma a representar as diversas combinações de tais características. Um pouco mais de detalhes será feito apenas para o último vestibular. Como o objetivo da análise desses vestibulares foi identificar questões com propriedades importantes para o planejamento de novas provas e como para itens com valores de parâmetros semelhantes tais propriedades são as mesmas, considerou-se que a representação de apenas três deles é suficiente para exemplificar como a metodologia pode revelar importantes propriedades para o planejamento de provas, bem como auxiliar a seleção de candidatos com base em suas habilidades.

3.1 Análise do vestibular 2006-2

Nas Figuras 1 a 3 estão representadas as estimativas dos parâmetros dos 74 itens do vestibular 2006-2 com seus respectivos intervalos de credibilidade HPDs. No que diz respeito à convergência, apenas os itens 43, 50 e 65 tiveram suas estatísticas de Gelman-Rubin maiores que o valor desejado para o parâmetro a e o

item 65 para o parâmetro *c*.

Observando-se a Figura 1, claramente pode-se perceber uma divisão entre o poder de discriminação dos itens 32 para baixo e 33 para cima, apresentando-se estes últimos como os que apresentaram mais altos valores para o parâmetro *a*. Comparando-se com a Figura 2 verifica-se que essa divisão se repete quanto ao grau de dificuldade. É interessante observar que o grupo de itens que apresentaram maior poder de discriminação no vestibular 2006-2 foram, também, os que apresentaram maior grau de dificuldade e os que apresentaram menor poder de discriminação, dificuldade menor. No grupo de itens de 32 para baixo, onde se encontram os de menos poder de discriminação, excetua-se os itens 6, 8, 16, 21 e 22 por apresentaram um valor maior para o parâmetro *a*. Observando-se o grau de dificuldade desses itens pode-se observar que eles também apresentaram um grau de dificuldade maior. De acordo com esses resultados, pode-se dizer que nesse vestibular de 2006-2, os itens mais difíceis foram os que apresentaram maior poder de discriminação. Esses itens se referem às provas de Inglês, Biologia, Física, Matemática e Química. De forma geral e desconsiderando-se os itens que não obtiveram convergência, pode-se observar que, no vestibular 2006-2, o item que apresentou maior estimativa pontual para o parâmetro *a* foi o item 44 (Biologia) e a menor o item 69 (Química). O item 44 apresenta maior poder de discriminação e também está entre o grupo dos itens mais difíceis, apresentando probabilidade de acerto casual dentro da média esperada para questões com 4 alternativas. O item 69 tem menor poder de discriminação e está entre o grupo dos mais fáceis, com probabilidade de acerto casual um pouco abaixo da média esperada. Observa-se mais uma vez que os itens mais difíceis foram também os que apresentaram valores mais altos para o parâmetro *a* e vice-versa. Considerando a classificação apresentada na seção 2.1.2 do capítulo 1, a respeito desse parâmetro,

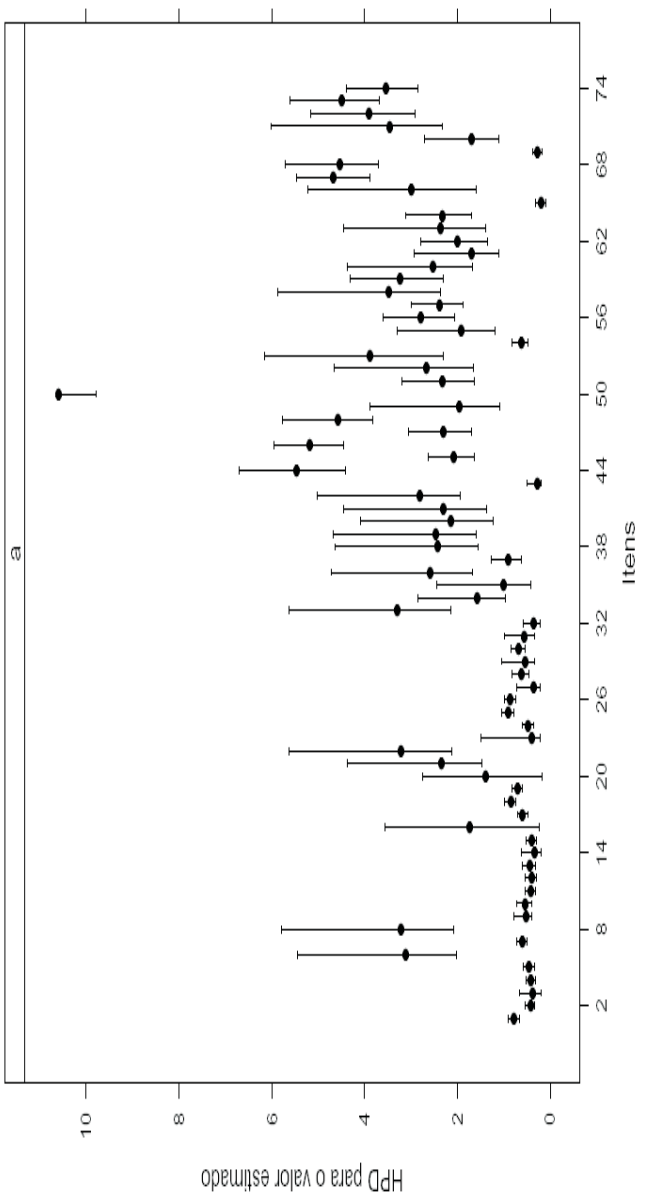


Figura 1 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2006-2 da UFLA e respectivos intervalos de credibilidade HPD a 95%

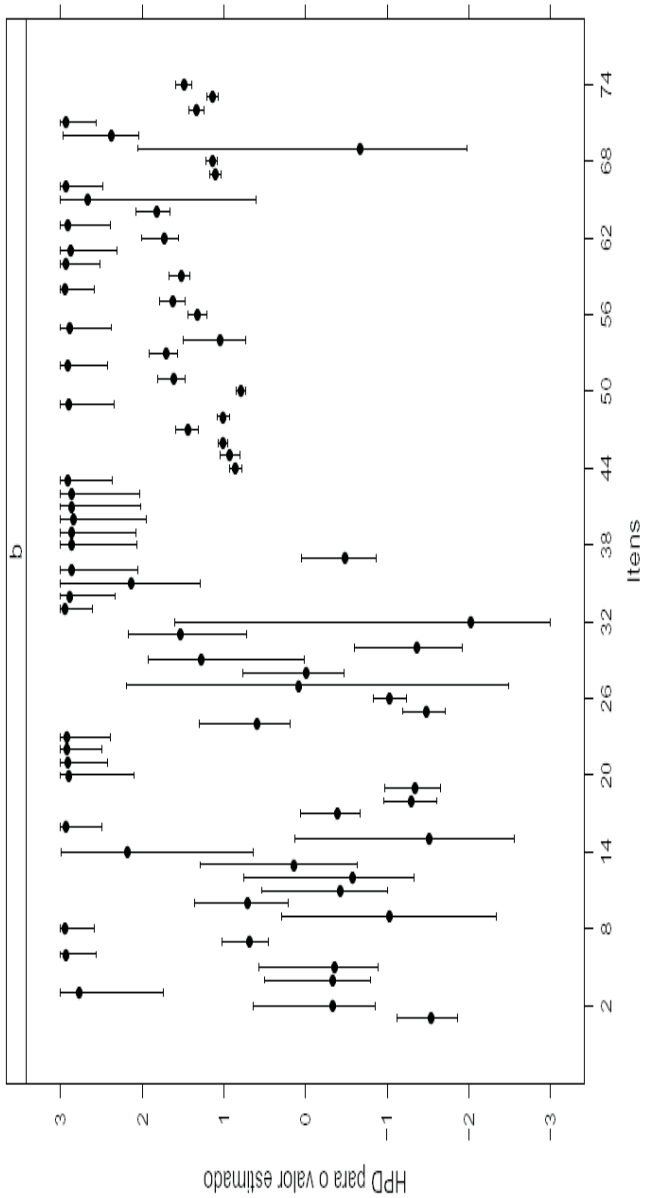


Figura 2 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2006-2 da UFLA e respectivos intervalos de credibilidade HPD a 95%

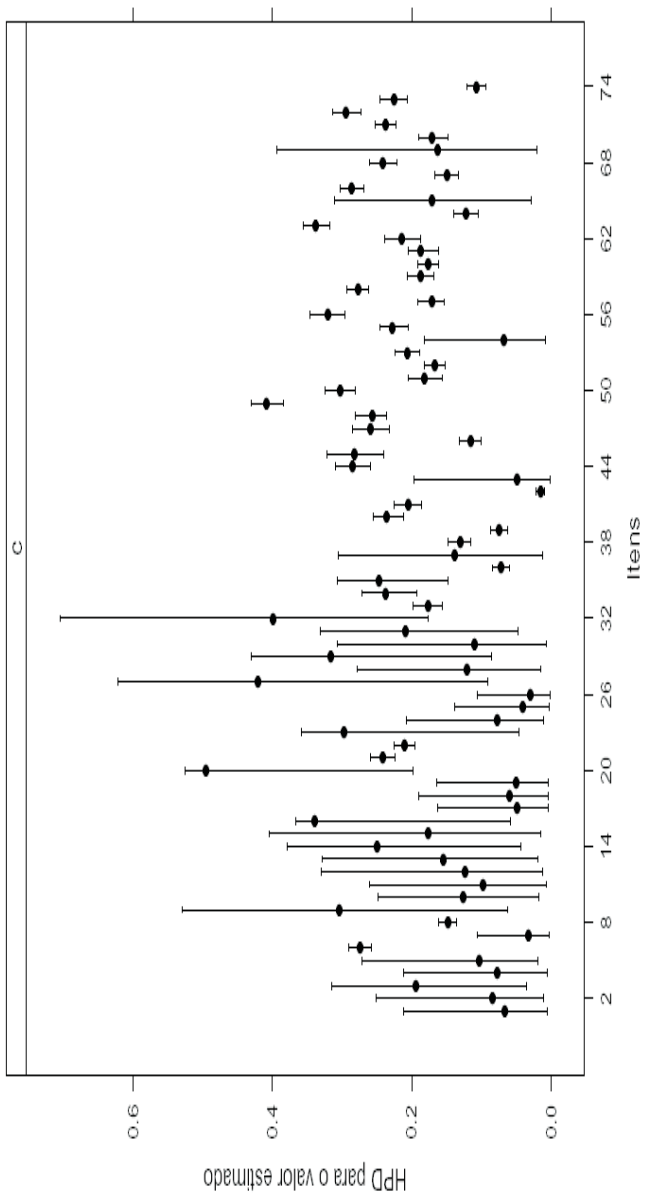


Figura 3 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2006-2 da UFLA e respectivos intervalos de credibilidade HPD a 95%

tem-se que 2 itens tiveram muito baixa discriminação, 20 baixa discriminação, 8 discriminação moderada, 3 alta e 38 muito alta.

O item mais difícil (maior valor do parâmetro *b*) foi o item 33 (Espanhol) e, o mais fácil, o item 32 (Espanhol). Novamente, verificando-se o poder de discriminação desses itens, observa-se que o item 33, mais difícil também possui bom poder de discriminação e o item 32, mais fácil, baixo poder de discriminação.

Quanto ao parâmetro *c*, 44 itens tiveram intervalos que não abrangeram o valor 0,25, que é o valor esperado para esse vestibular, já que o mesmo possui quatro alternativas. Isso significa que 60,3% dos itens possuem valores de *c* estatisticamente diferentes de 0,25, sendo que 35 deles estão abaixo desse valor. Dessa forma, pode-se dizer que o vestibular 2006-2 da UFLA teve um considerável número de itens com baixa probabilidade de acerto por candidatos com baixa habilidade.

A Figura 4 apresenta o histograma das habilidades estimadas dos candidatos ao vestibular 2006-2, juntamente com o gráfico da FII e da CCI com seu respectivo intervalo de credibilidade HPD para três itens, sendo um do grupo dos mais fáceis (item 11) e com baixa discriminação, e dois do grupo dos mais difíceis (8 e 74), com boa discriminação, sendo, no entanto, um muito mais difícil que o outro e verificar a implicação dessas características na escolha de candidatos com habilidades de interesse. Gráficos desse tipo foram desenvolvidos para todos os itens, sendo, porém esses três escolhidos por representarem algumas situações típicas, importantes de serem observadas para futuro planejamento de novas provas.

De acordo com essa Figura, pode-se fazer as seguintes observações:

1)item 8 (Português: a = 3,69; b = 2,85; c = 0,15) - é um item bom apenas para candidatos com habilidades muito elevadas, pois, apesar de possuir bom

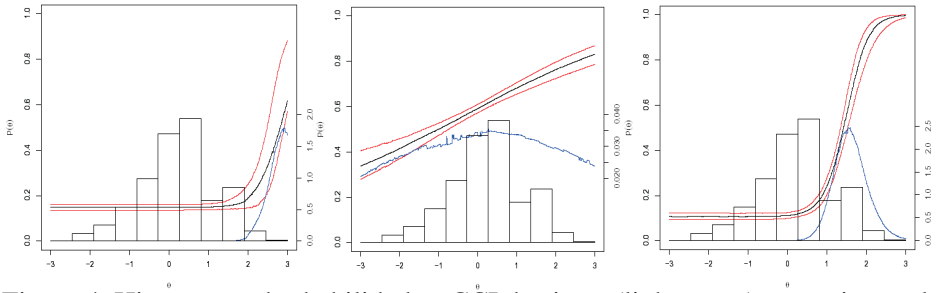


Figura 4 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha, vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para exemplos de um item muito difícil (item 8), um item muito ruim (item 11) e um item bom (item 74), do vestibular 2006-2, dispostos nessa ordem

poder de discriminação, é muito difícil. Está fora da habilidade do grupo em geral;

2) item 11 (Geografia: a = 0,43; b = -0,28; c = 0,33) - é um item muito ruim. É fácil para o grupo, mas de baixo poder de discriminação;

3) item 74 (Química: a = 3,58; b = 1,50; c = 0,11) - item bom. Possui bom poder de discriminação e dificuldade compatível com um grupo razoável dos melhores candidatos.

Pode-se observar que, apesar de ser de interesse que se tenham itens com elevado poder de discriminação, é necessário que seja atentado para que o grau de dificuldade não seja tão elevado a ponto de não haver indivíduos com nível de habilidade suficiente para respondê-lo corretamente.

A média da FIT e da CCT com seus respectivos intervalos de credibilidade HPD para cada disciplina do vestibular 2006-2, junto com o histograma das estimativas das habilidades estão representadas na Figura 5.

A maior quantidade de informação é obtida em torno do valor do parâmetro *b* e aumenta quanto maior for o valor do parâmetro *a*. Por meio das Figuras 1 a

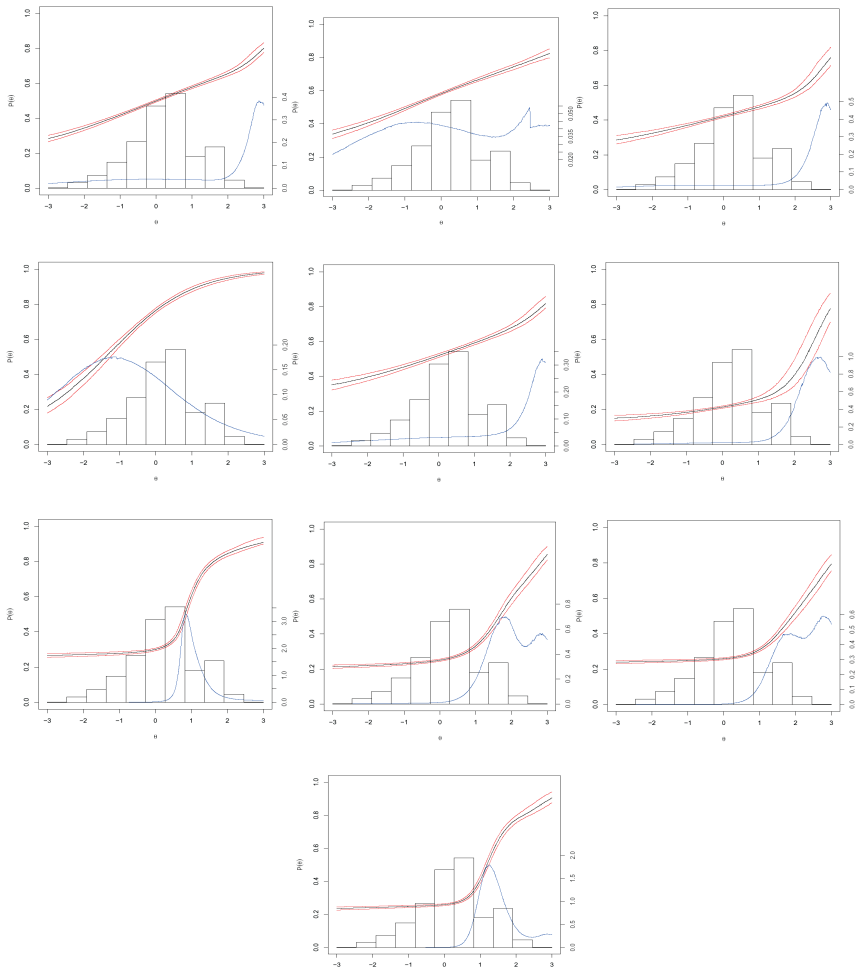


Figura 5 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2006-2, dispostas nessa ordem

3, observou-se que as provas mais difíceis e que mais discriminaram no vestibular 2006-2, foram as provas de Inglês, Biologia, Física, Matemática e Química.

Na Figura 5, pode-se verificar que, essas mesmas provas foram, também, as mais informativas. O fato é que curvas de informação situadas à direita são mais informativas para o grupo dos melhores candidatos, e elas são resultantes de itens mais difíceis. Provas muito fáceis, como foi o caso da prova de Filosofia, são pouco informativas para esse grupo (a curva de informação se encontra à esquerda), dentro do qual se tem o interesse que seja selecionado os melhores. A prova menos informativa, neste vestibular de 2006-2, foi a de Geografia.

Comparando-se as duas disciplinas de Língua Estrangeira, pode-se observar que, no vestibular 2006-2, apesar de ambas apresentarem um elevado grau de dificuldade, a disciplina de Inglês apresentou um grau de dificuldade mais condizente com o grupo em questão. Foi, portanto, mais informativa que aquela para um grupo maior dos melhores candidatos.

Foi possível estimar as habilidades de todos os candidatos ao vestibular 2006-2, sendo obtida uma média de -0,03 e desvio padrão de 0,84. Todos os valores convergiram de acordo com os métodos adotados. A correlação com as notas foi de 85%. Os valores para a maior e a menor habilidade estimada encontram-se na Tabela 3.

Tabela 3 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2006-2 da UFLA e seus respectivos erros de Monte Carlo (EMC)

Candidato	Habilidade estimada	I.C. - HPD		EMC.
		inferior	superior	
2140	-2,85	-4,51	-1,78	0.046
998	2,18	1,722	2,63	0.004

Houve 1283 inscritos que optaram pela disciplina de Espanhol no vestibular 2006-2, correspondendo a 42% dos candidatos. A média das habilidades estimadas desse grupo foi de 0,16 com um desvio padrão de 1,06. Para os que optaram

pelo Inglês, a média foi de -0,16 e desvio padrão de 0,61, com um total de 1770 inscritos, representando 58% do total.

As médias das habilidades estimadas por curso do vestibular 2006-2 e seus respectivos desvios padrão estão representados na Tabela 4.

Tabela 4 Médias das habilidades por curso e respectivo desvio padrão (sd) referentes ao vestibular 2006-2

Curso	Média	sd	Curso	Média	sd
AD	-0,05	0,91	EA	-0,08	0,86
AG	0,05	0,88	EF	0,20	0,82
AL	0,08	0,81	MV	0,08	0,85
CB	0,23	0,87	QI	0,10	0,98
CC	0,01	0,79	ZO	0,11	0,96

Observa-se que, no vestibular 2006-2, o curso que obteve menor habilidade média foi o de Engenharia Agrícola e maior o de Ciências Biológicas. Pode-se observar também que os desvios padrão não diferem muito um do outro o que indica que pode ser usada uma variância comum para comparar estas médias.

A Figura 6 representa os resultados do vestibular 2006-2 como um todo, isto é, possui o histograma das habilidades estimadas dos candidatos juntamente com o gráfico da FIT e da CCT, com seu respectivo intervalo de credibilidade HPD.

Pode-se observar que esse vestibular de 2006-2 apresentou um grau de dificuldade superior à média da população. No entanto, essa dificuldade se encontra em um nível condizente com um grupo razoável dos melhores candidatos, tanto de forma geral como para cada curso. Considerando-se que acima de um desvio padrão da média de uma distribuição normal está aproximadamente 16% da população, para esse vestibular isso corresponde a buscar quantos indivíduos possuem habilidade acima de 0,81, que é o grupo onde se encontram os indivíduos de interesse que sejam os selecionados. Houve 560 candidatos nessas condições.

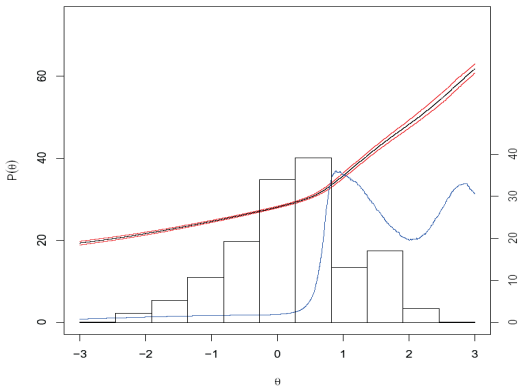


Figura 6 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), para o vestibular 2006-2

Observando-se a curva de informação do teste vê-se que o maior conteúdo de informação abrange de forma satisfatória tais indivíduos. Quanto ao poder de discriminação verifica-se que a CCT possui uma inclinação que torna possível a diferenciação entre eles. Por meio dessa curva pode-se verificar também, que indivíduos com baixa habilidade não conseguem acertar 60% das questões, que é a condição para passar no vestibular. Acertam, aproximadamente, apenas 20 questões da prova toda.

3.2 Análise do vestibular 2007-1

Na Figura 7 a 9 estão representadas as estimativas dos parâmetros dos 74 itens do vestibular 2007-1 com seus respectivos intervalos de credibilidade HPDs. Houve convergência para os parâmetros de todos os itens.

Por meio dessa Figura, pode-se observar que, no vestibular 2007-1, dois itens se destacaram quanto ao maior valor da estimativa pontual para o parâmetro

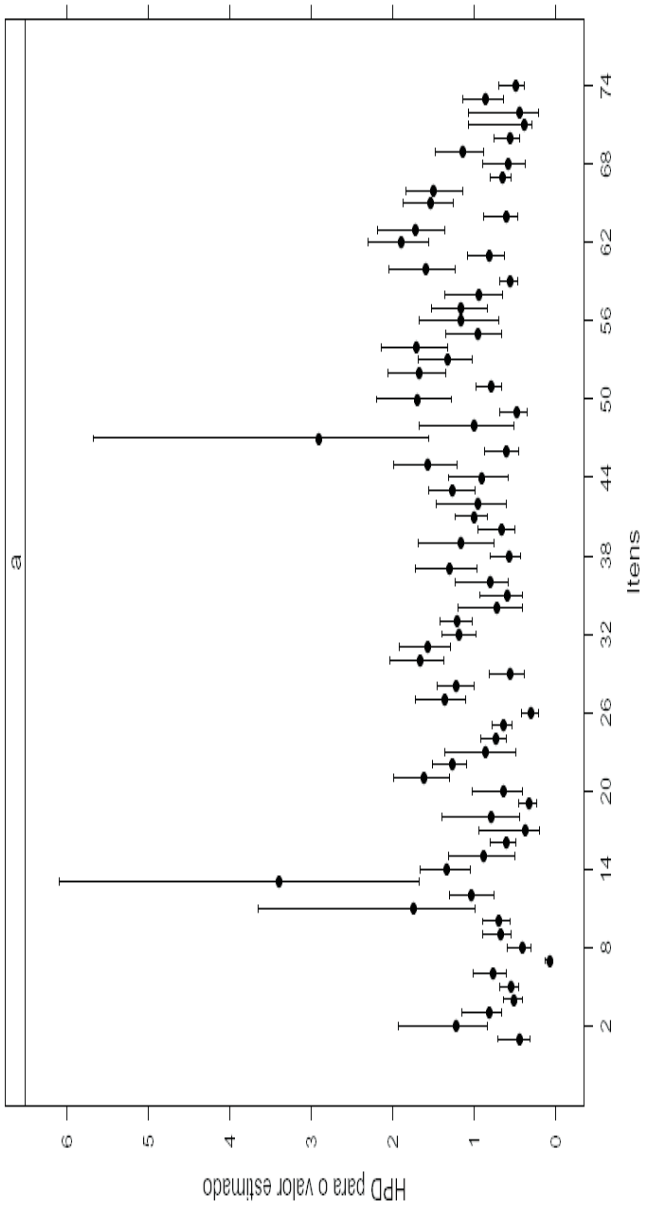


Figura 7 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2007-1 da UFLA e respectivos intervalos de credibilidade HPD a 95%

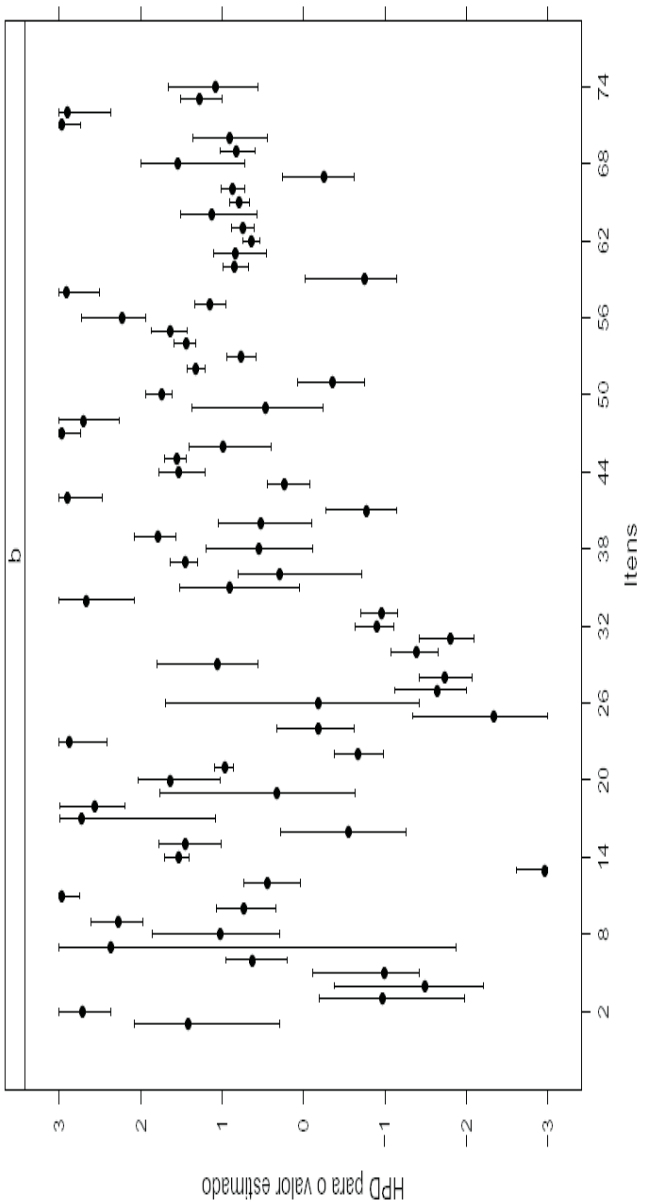


Figura 8 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2007-1 da UFLA e respectivos intervalos de credibilidade HPD a 95%

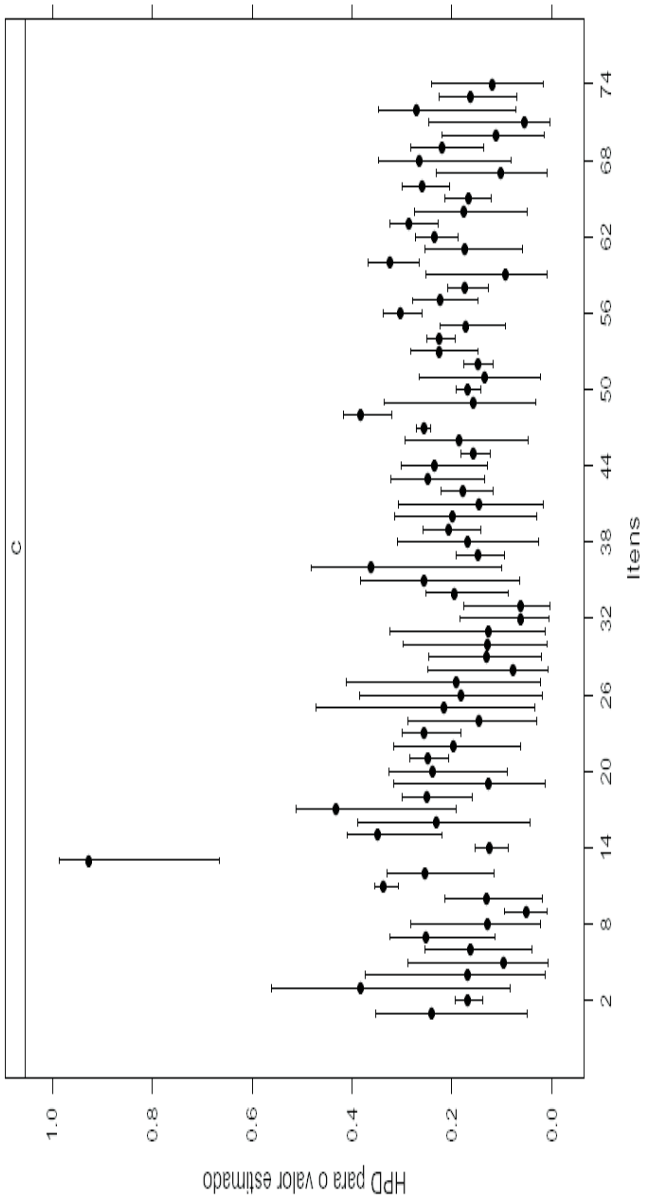


Figura 9 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2007-1 da UFLA e respectivos intervalos de credibilidade HPD a 95%

a: o item 13 (Geografia) e o item 47 (Biologia). Pode-se destacar também o item 11 (Geografia) como um item com alto poder de discriminação. Com excessão do item 13, os outros dois também apresentaram grau de dificuldade bem elevado. O menor valor para o parâmetro *a* foi o do item 7 (Português) seguido do item 26 (Filosofia), ambos com baixo poder de discriminação. De acordo com a classificação apresentada na seção 2.1.2 do capítulo 1 para esse parâmetro, 3 itens tiveram muito baixa discriminação, 18 baixa discriminação, 37 discriminação moderada, 9 alta e 7 muito alta.

Com relação ao parâmetro *b* verifica-se que a maioria das questões teve grau de dificuldade mais elevado sendo, entretanto, bem heterogêneos os seus valores. Os itens mais difíceis foram os de número 11 (Geografia), 47 (Biologia) e 71 (Química) e os mais fáceis os de número 13 (Geografia) e 25 (Filosofia). De forma geral, neste vestibular de 2007-1, as 5 últimas provas (Inglês, Biologia, Física, Matemática e Química), da mesma forma como ocorreu no vestibular anterior, também foram as provas mais difíceis.

Quanto ao parâmetro *c*, houve 26 itens com intervalos não abrangendo o valor 0,25, isto é, 35% dos itens possuem valores para o parâmetro *c* estatisticamente diferentes de 0,25, sendo que apenas 5 deles estão acima desse valor.

Um item a ser destacado nesse vestibular de 2007-1 é o item 13, pois foi um item muito fácil e com elevada probabilidade de acerto por indivíduos com baixa habilidade. Não era esperado, portanto, que apresentasse bom poder de discriminação. Procurou-se averiguar mais acuradamente a razão e foi constatado que se tratava de uma questão anulada. Consequentemente, todos os candidatos receberam 1 ponto para essa questão, independente da alternativa escolhida. Assim, o "acerto" a ela é garantido, "chutando-se" ou não. Isso explica o fato da estimativa do parâmetro *c* ter sido tão elevada e a do grau de dificuldade tão baixa.

A Figura 10 representa o histograma das habilidades estimadas dos candidatos ao vestibular 2007-1 juntamente com o gráfico da FII e da CCI com seu respectivo intervalo de credibilidade HPD para três itens representativos de características importantes a serem observadas quando da elaboração de novas provas: um item muito difícil, um item muito ruim e um item bom. Dos itens comentados acima, entre os que apresentaram maior poder de discriminação (11 e 47) escolheu-se o item 47 (Biologia) e entre os de menor (7 e 26), o item 26 (Filosofia). Entretanto, como um deles é muito difícil e o outro ruim, para representar a característica de um item bom, o terceiro item escolhido foi o de número 45 (Biologia).

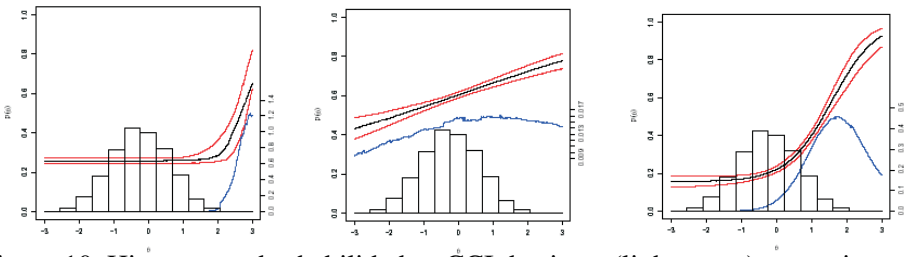


Figura 10 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para exemplos de um item muito difícil (item 47), um item muito ruim (item 26) e um item bom (item 45), do vestibular 2007-1, dispostas nessa ordem

Pode-se observar, pela Figura, que:

1) item 47 (Biologia: a = 3,46; b = 2,90; c = 0,26) - é considerado um item muito difícil, pois está fora da habilidade do grupo, apesar de possuir alto poder de discriminação;

2) item 26 (Filosofia: a = 0,32; b = 0.03; c = 0,21) - é um item muito ruim, pois, apesar de fácil para o grupo, tem baixo poder de discriminação;

3) item 45 (Biologia: a = 1,60; b = 1,57; c = 0,16) - item bom. Possui bom

poder de discriminação e um grau de dificuldade compatível com as habilidades dos melhores candidatos.

Observação: A explicação do elevado valor obtido para a estimativa do parâmetro *b* do item 47 desse vestibular 2007-1 e, consequentemente, de não ter havido nenhum candidato com habilidade suficiente para acertá-la, foi devido ao fato de que não houve alternativa correta para ela, conforme averiguado com o coordenador do curso de Biologia. No entanto, essa questão não foi anulada e considerou-se como resposta correta a alternativa **B**. Assim, como não houve resposta correta, realmente "trata-se de um item muito difícil" e fora da habilidade dos candidatos. Resta a opção de "chute" pelos mesmos. A estimativa obtida para esse parâmetro ($c = 0{,}26$) foi, portanto, coerente para uma questão de 4 alternativas em que não é possível saber qual a alternativa é correta. Essa ocorrência somada ao ocorrido com o item 13 (comentado acima) pode ser visto como mais um fator que confirma a confiabilidade da metodologia utilizada e do algoritmo empregado.

A média da FIT e da CCT com seus respectivos intervalos de credibilidade HPD para cada disciplina do vestibular 2007-1, junto ao histograma das estimativas das habilidades, estão representadas na Figura 11.

Pode-se observar que, no vestibular 2007-1, a disciplina mais informativa foi a de Matemática e a menos informativa a de Filosofia. Comparando-se a FIT com o grau de dificuldade das provas como um todo, pode-se perceber que as provas mais difíceis possuem o gráfico da FIT à direita da média da turma (histograma) e, portanto, trazem mais informação para quem tem habilidade maior. Provas muito fáceis, como a de Filosofia, tem a FIT à esquerda. A prova de Espanhol também teve o gráfico da FIT à esquerda da habilidade média dos candidatos. No entanto, foi mais informativa que a de Filosofia. Isto se deve ao fato de que esta foi mais discriminativa que àquela, o que pode ser visto por meio da

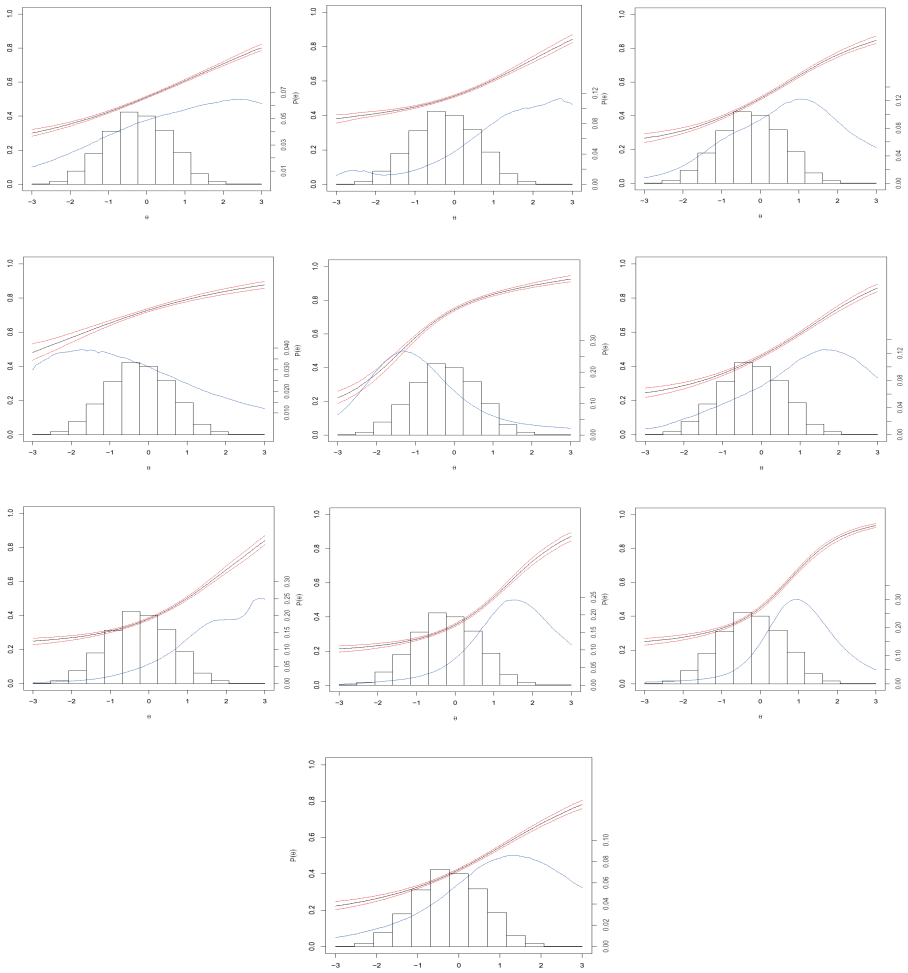


Figura 11 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2007-1, dispostos nesta ordem

inclinação de sua CCT (quanto maior a discriminação, maior a informação). A informação que esta prova forneceu (Espanhol) é boa para candidatos abaixo da

média das habilidades, o que também é útil, pois se houvesse somente questões com grau de dificuldade mais elevado tornar-se-ia difícil quantificar o quanto sabe indivíduos com habilidades mais baixa.

Com relação à prova de Língua Estrangeira, para este vestibular de 2007-1, apesar da prova de Inglês ter sido mais difícil que a de Espanhol, traz mais conteúdo de informação para o grupo de interesse.

O grau de dificuldade das provas como um todo foram condizentes com o nível dos candidatos. Pode-se dizer que, com exceção das provas de Português e Filosofia, todas as demais provas forneceram bom poder de discriminação, o que pode ser visto através da inclinação da CCT.

Estimou-se as habilidades de todos os candidatos do vestibular 2007-1, obtendo-se convergência para todos os valores. A média geral foi de -0,01 com um desvio padrão de 0,89. Houve alta correlação entre as habilidades estimadas e as notas com um valor de 0,94. Na Tabela 5 encontram-se relacionadas a maior e a menor estimativa da habilidade.

Tabela 5 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2007-1 da UFLA e seus respectivos erros de Monte Carlo (EMC)

Candidato	Habilidade estimada	I.C. - HPD		EMC.
		inferior	superior	
38	-2,96	-3,95	-1,91	0.025
2098	3,08	2,28	3,82	0.017

Dos candidatos inscritos no vestibular 2007-1, 1843 optaram pela disciplina de Espanhol, representando 42,53% do total. A habilidade média desses candidatos foi de -0,21 com um desvio padrão de 0,80. Para a opção Inglês, houve 2490 candidatos, ou seja, 57,47%, com uma habilidade média de 0,14 e um desvio padrão de 0,92.

Na Tabela 6 encontram-se as médias das habilidades estimadas para cada curso do vestibular 2007-1 com seus respectivos desvios padrão.

Tabela 6 Médias das habilidades por curso e respectivo desvio padrão (sd) referentes ao vestibular 2007-1

Curso	Média	sd	Curso	Média	sd
AD	0,02	0,90	EF	0,16	0,88
AG	0,03	0,83	MA	-0,33	0,77
AL	0,35	0,88	MV	0,22	0,91
CB	0,24	0,87	QI	-0,15	0,86
CC	0,24	0,90	SI	-0,23	0,79
EA	0,01	0,76	ZO	0,01	0,76
ED	-0,59	0,69			

Pode-se observar que, no vestibular 2007-1, o curso que obteve menor habilidade média foi o de Educação Física e maior o de Engenharia de Alimentos. Observa-se, também, que os desvios padrão possuem seus valores semelhantes indicando que poderia ser considerado um único valor como variância comum.

Na Figura 12 estão representados o histograma das habilidades estimadas dos candidatos, a curva da FIT e a CCT com seu respectivo intervalo de credibilidade HPD do vestibular 2007-1 todo.

Por meio dessa Figura pode-se observar que o vestibular de 2007-1 teve um grau de dificuldade acima da média da população, porém, condizente com um grupo de candidatos com maiores habilidades. Como a informação é maior em torno do valor do grau de dificuldade, pode-se dizer, também, que esse vestibular foi informativo para esses candidatos. Isso é confirmado pela CCT, que se encontra à direita da média das habilidades. Houve 716 candidatos com habilidade estimada acima de um desvio padrão da média e, através da inclinação da CCT pode-se dizer que é possível uma boa discriminação entre eles. Nesse vestibular de 2007-1, um indivíduo com baixa habilidade tem a chance de acertar 20 questões por mero

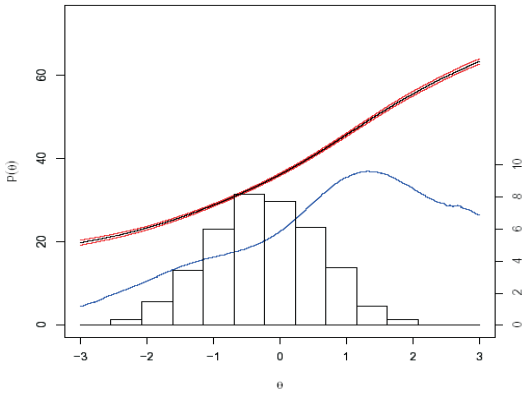


Figura 12 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita) do vestibular 2007-1

acaso.

3.3 Análise do vestibular 2007-2

No vestibular 2007-2, todos os parâmetros dos 74 itens tiveram sua convergência atingida. Suas estimativas juntamente com seus respectivos intervalos de credibilidade HPDs encontram-se na Figura 13 a 15 abaixo.

Por meio dessas Figuras verifica-se que, no vestibular 2007-2, as cinco últimas provas (Inglês, Biologia, Física, Matemática e Química - a partir do item 35) foram as mais difíceis e tenderam a ser, também, as que apresentaram maior poder de discriminação. Os itens 5 (Português), 7 (Português), 27 (Espanhol) e 57 (Física) foram os itens que apresentaram maiores valores para o parâmetro de discriminação, sendo o de número 57 o que obteve maior estimativa pontual. Comparando-se com o grau de dificuldade dos mesmos, nota-se que os itens 5, 7 e 57 foram itens muito difíceis e o item 27 muito fácil. Os itens com menores

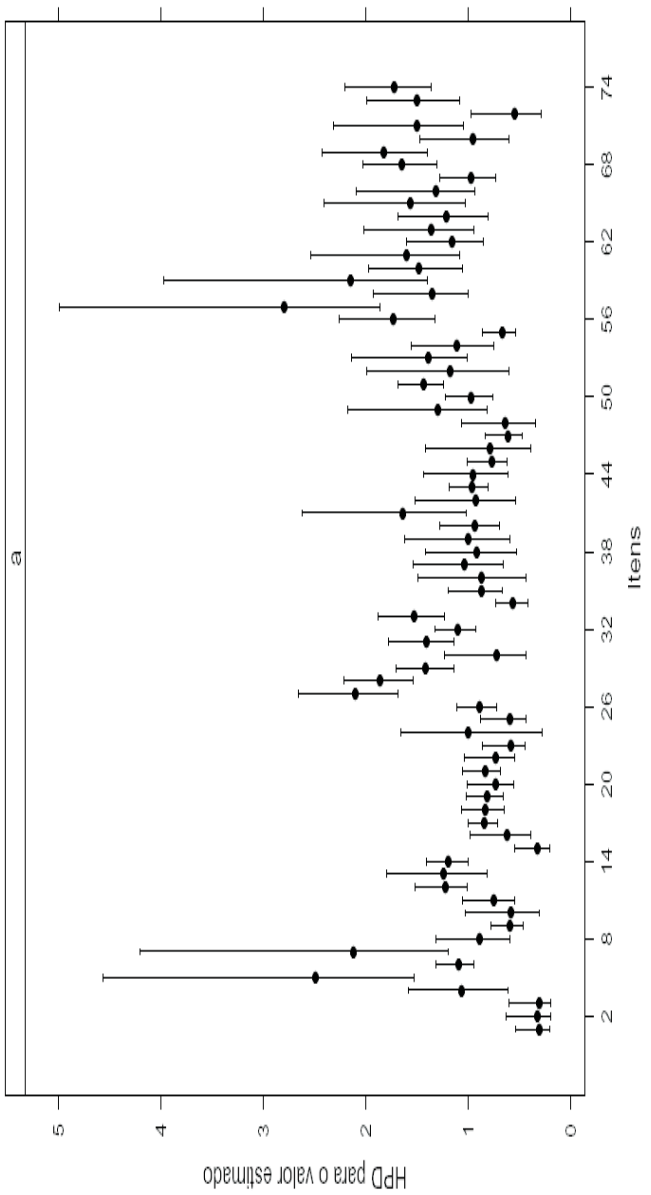


Figura 13 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2007-2 da UFLA e respectivos intervalos de credibilidade HPD a 95%

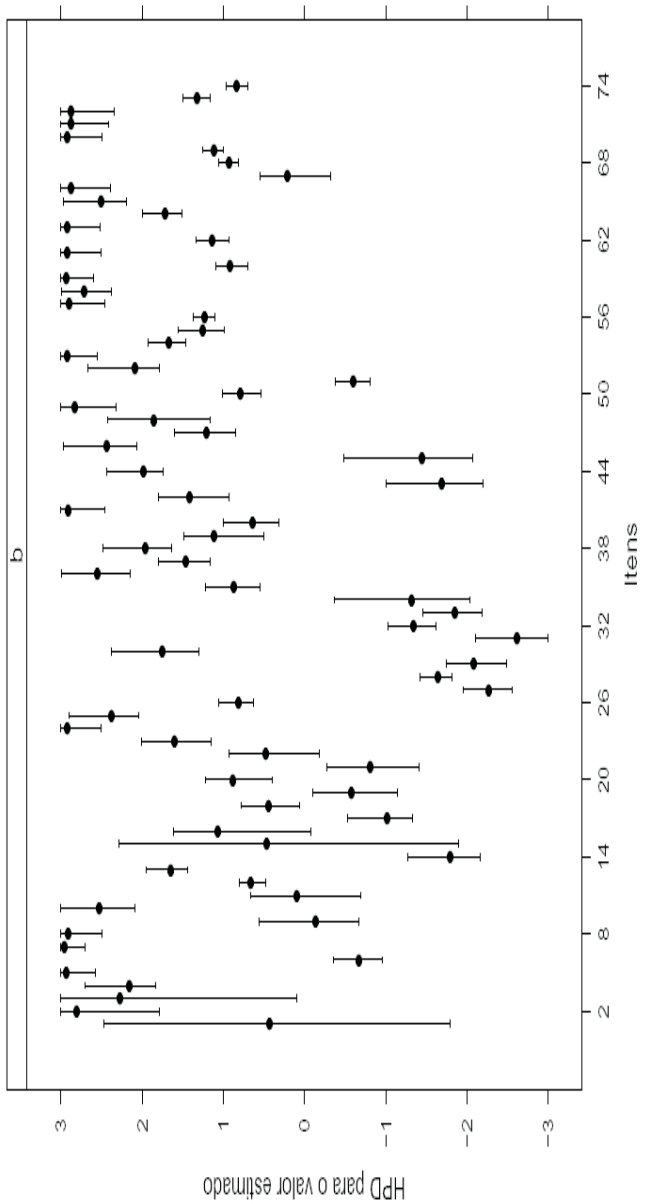


Figura 14 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2007-2 da UFLA e respectivos intervalos de credibilidade HPD a 95%

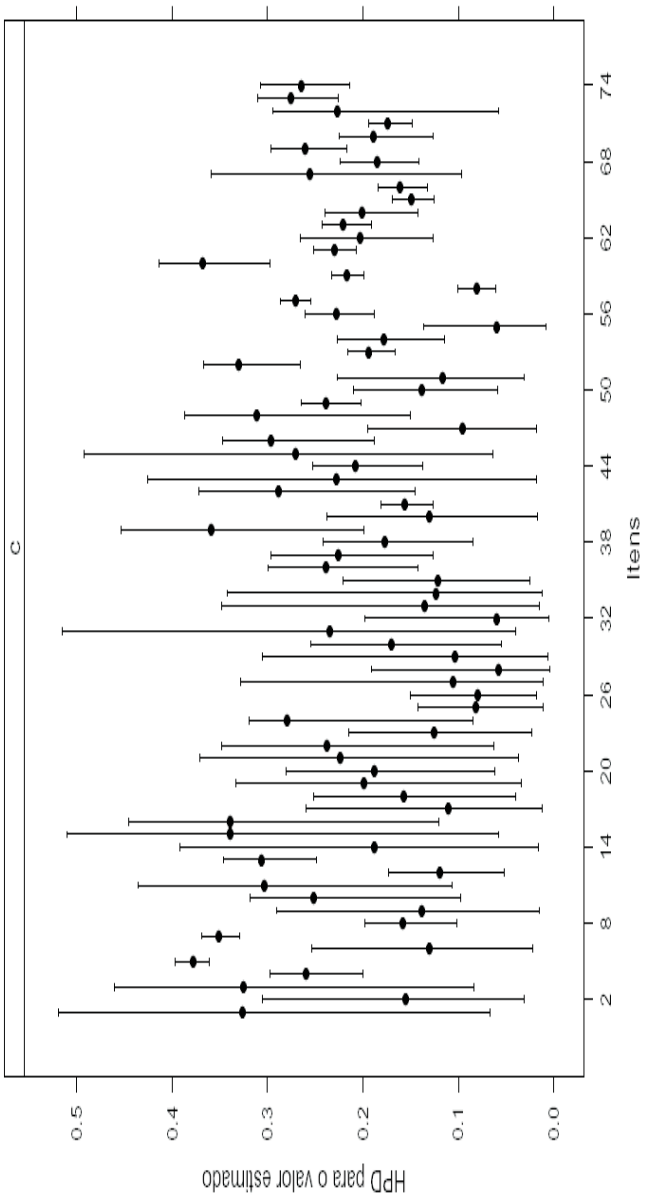


Figura 15 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2007-2 da UFLA e respectivos intervalos de credibilidade HPD a 95%

valores para esse parâmetro foram os de número 1, 2 e 3, todos de Português, tendo o item 3 obtido o menor valor dentre todos os itens desse vestibular. Os itens 2 e 3 tiveram baixo valor para o parâmetro de discriminação e apresentaram elevado grau de dificuldade. De forma geral, conforme a classificação dada para o parâmetro *a*, apresentada na seção 2.1.2 do capítulo 1, tem-se que 4 itens tiveram muito baixa discriminação, 9 baixa discriminação, 38 discriminação moderada, 14 alta e 9 muito alta, podendo-se, portanto, ser considerado como um vestibular que teve boa discriminação como um todo.

Para o parâmetro *b*, o item que apresentou maior grau de dificuldade foi o de número 7 (Português) e o menor o de número 31 (Espanhol).

Quanto ao parâmetro *c*, 31 itens tiveram seus valores estatisticamente diferentes de 0,25, o que corresponde a 42% do total. Desses 31, apenas 5 tiveram seus valores estatisticamente maiores que 0,25. Logo, esse vestibular de 2007-2 não apresentou muita probabilidade de que um candidato com baixa habilidade viesse a acertar suas questões de forma aleatória.

Os itens 5, 7 e 57 foram itens que apresentaram valores pontuais semelhantes para seus parâmetros, ou seja, alto valor para o parâmetro *a*, elevado grau de dificuldade e alta probabilidade de acerto casual. Itens com valores paramétricos semelhantes possuem características semelhantes. Portanto, como já foram discutidas questões dessa natureza em outros vestibulares (itens com alto poder de discriminação, porém muito difíceis), é interessante analisar outros tipos de itens, ou seja, aqueles que possuam outros tipos de CCIs, visando extrair dos mesmos características que auxiliarão no planejamento de novas provas. Pelo observado, percebe-se uma tendência a que os itens mais difíceis sejam também os que apresentam maior poder de discriminação. No entanto, como comentados acima, o item 27 foi um item fácil, mas que apresentou bom poder de discriminação, e o

item 2 foi um item difícil porém que apresentou baixo poder de discriminação. Escolheu-se, portanto, esses dois itens, juntamente com um item bom (o item de número 56 - Física), para analisar o gráfico que os representam. Assim, na Figura 16 estão juntamente representados, para esses três itens, o histograma das habilidades estimadas dos candidatos ao vestibular 2007-2, o gráfico da FII e a CCI com seu respectivo intervalo de credibilidade HPD.

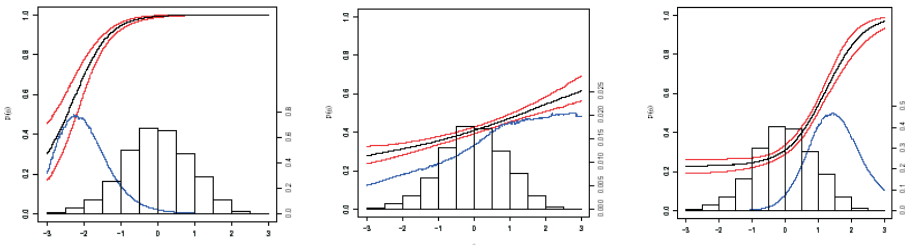


Figura 16 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para os itens 27, 2 e 56, do vestibular 2007-2, dispostos nessa ordem

Pode-se observar que:

1)item 27 (Espanhol: a = 2,13; b = -2,26; c = 0,15) - é um item com baixa probabilidade de acerto por indivíduos com baixa habilidade, boa discriminação, porém, muito fácil. É um tipo de item útil para fornecer informação sobre candidatos com baixa habilidade;

2) item 2 (Português: a = 0,38; b = 2,52; c = 0,17) - é um item muito ruim, pois, além de ser muito difícil para o grupo, tem baixo poder de discriminação e traz pouca informação. Pode-se verificar aqui a propriedade de invariância da TRI, isto é, este foi um item difícil independente dos candidatos que realizaram a prova;

3) item 56 (Física: a = 1,76; b = 1,23; c = 0,23) - é um item bom. Tem um

grau de dificuldade que se adéqua ao nível de habilidade dos melhores candidatos e proporciona boa discriminação entre eles. No entanto, atentando-se para o seu conteúdo de informação e comparando-o com o do item 27, pode-se verificar que ele possui um valor máximo menor, isto é, a informação que ele fornece para o grupo dos candidatos com níveis de habilidades maiores é menor que a informação fornecida pelo item 27 para candidatos com níveis de habilidades menores. Isso mostra que itens fáceis podem ser muito úteis e informativos e itens difíceis podem ser bons ou ruins.

Na Figura 17 estão contidos o gráfico da média da FIT, da CCT com seu respectivo intervalo de credibilidade HPD para cada disciplina e o histograma das estimativas das habilidades para o vestibular 2007-2.

A maioria das provas do vestibular de 2007-2 apresentou grau de dificuldade superior à média do grupo de candidatos. Isso pode ser verificado por meio da posição em que se encontra a CCT, pois a mesma é mais informativa em torno do valor do parâmetro b. Essas provas são: Português, Geografia, Filosofia, Inglês, Biologia, Física, Matemática e Química. Destas, pode-se verificar que a prova de Português foi muito difícil para o nível de habilidade dos candidatos, sendo muito pouco informativa para aqueles que possuem habilidade abaixo de 2.

A prova mais informativa desse vestibular de 2007-2 foi a de Matemática. No entanto, observa-se que a prova de Química foi mais informativa para o grupo de candidatos de maior interesse. A prova de História, apesar de possuir uma dificuldade condizente com o nível de habilidade do grupo, foi pouco informativa. A prova de Espanhol foi muito fácil para os candidatos a esse vestibular. Foi uma prova muito informativa e com boa discriminação, mas, para candidatos com baixa habilidade.

É interessante comentar aqui a relação que tem a informação com o grau de

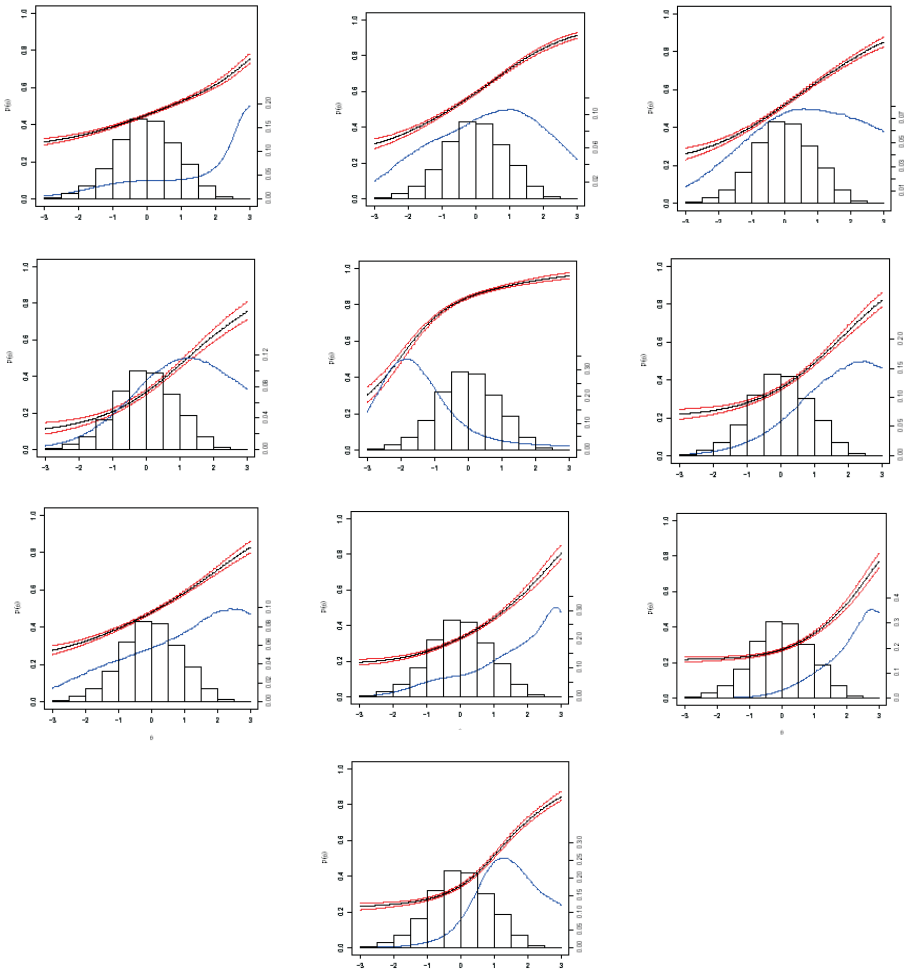


Figura 17 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2007-2, dispostas nessa ordem

dificuldade e com o poder de discriminação. Quanto mais difícil, a prova se torna mais informativa para candidatos com habilidade maior, ou seja, quanto mais à

direita a CCT, é mais informativa nessa região. No entanto, se for muito difícil a informação será para habilidades muito altas e não informará praticamente nada para o grupo das habilidades em questão. Foi o que ocorreu com a prova de Português. A informação também está relacionada ao parâmetro da discriminação, sendo que quanto maior o valor deste parâmetro, mais informativa será a prova. No entanto, se a questão é muito fácil (estiver mais a esquerda da média) a informação será maior para um grupo de baixa habilidade e também informará muito pouco para as habilidades de interesse, que foi o que ocorreu com a prova de Espanhol. Portanto, uma prova que possua alto valor para o poder de discriminação e tenha grau de dificuldade um pouco mais difícil que a média da habilidade dos candidatos é a mais indicada para selecionar os melhores. No entanto, mesmo as provas fáceis podem ser muito úteis e informativas.

A média das habilidades dos candidatos no vestibular 2007-2 foi de -0,02 com um desvio padrão de 0,89. Foram obtidas as estimativas para a habilidade de todos os candidatos, atingindo-se a convergência para todas elas. A correlação com as notas foi de 0,92. Um desvio padrão acima da média da habilidade dos candidatos corresponde ao valor de 0,87. Obteve-se 553 candidatos nessas condições, representando 16,55% dos inscritos. A estimativa para os valores da maior e da menor habilidade encontram-se na Tabela 7.

Tabela 7 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2007-2 da UFLA e seus respectivos erros de Monte Carlo (EMC)

Candidato	Habilidade estimada	I.C. - HPD		EMC.
		inferior	superior	
2498	-2,89	-4,04	-2,02	0.019
1293	2,62	2,03	3,21	0.006

Houve 1565 inscritos no vestibular 2007-2 que optaram pela disciplina de

Espanhol, correspondendo a 46,83% dos candidatos. A média das habilidades estimadas desse grupo foi de -0,21 com um desvio padrão de 0,87. Para os que optaram pelo Inglês, a média foi de 0,15 e desvio padrão de 0,87, com um total de 1777 inscritos, representando 53,17% do total.

As médias das habilidades estimadas de cada curso do vestibular 2007-2 e seus respectivos desvios padrão, estão na Tabela 8.

Tabela 8 Médias das habilidades por curso do vestibular 2007-2 e respectivo desvio padrão (sd)

Curso	Média	sd	Curso	Média	sd
AD	0,08	0,86	EF	0,09	0,81
AG	-0,08	0,81	MA	-0,33	0,80
AL	0,30	0,89	MV	0,26	0,89
CB	0,15	0,94	QI	-0,16	0,87
CC	0,49	0,84	SI	-0,31	0,76
EA	0,10	0,78	ZO	-0,12	0,75
ED	-0,72	0,77			

Por meio dessa Tabela, observa-se que, no vestibular 2007-2, a menor habilidade média foi para os candidatos do curso de Educação Física e a maior para os de Ciência da Computação. Observa-se, também, que os desvios padrão não diferiram muito um dos outros, podendo ser considerado um único valor para todos (uma variância comum).

O histograma das habilidades estimadas dos candidatos ao vestibular 2007-2 juntamente com o gráfico da FIT e da CCT com seu respectivo intervalo de credibilidade HPD, do vestibular todo, encontra-se representado na Figura 18

Por meio dessa Figura, pode-se observar que, no vestibular 2007-2, ao se considerar o vestibular como um todo, houve um elevado grau de dificuldade e um moderado poder de discriminação. Um candidato que nada sabe deve acertar aproximadamente 20 questões em todo vestibular.

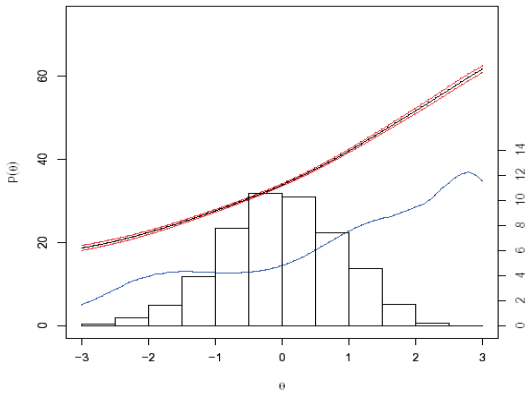


Figura 18 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), referentes ao vestibular 2007-2

3.4 Análise do vestibular 2008-1

Na Figura 19 a 21 estão representadas as estimativas dos parâmetros dos 74 itens do vestibular 2008-1 com seus respectivos intervalos de credibilidade HPDs. Obteve-se convergência para todos os parâmetros de todos esses itens.

Observa-se que, no vestibular 2008-1, em relação ao parâmetro a, os itens que apresentaram maiores valores para este parâmetro foram os de número 4 (Português), 53 (Física), 56 (Física) e 60 (Matemática), sendo que os itens 4, 56 e 60 também apresentaram elevado grau de dificuldade (o item 56 apresentando o maior grau de dificuldade do vestibular todo). No entanto, o item 53, apesar de ter apresentado elevado valor para a estimativa do parâmetro de discriminação, obteve elevada probabilidade de acerto por indivíduos com baixa habilidade. Foi também um item muito fácil. Investigando o motivo de tal anomalia, constatou-se que esse item foi anulado. Este fato é interessante, pois indica que o modo como foi estimado os parâmetros dos itens apresenta resultados coerentes. Como a questão foi

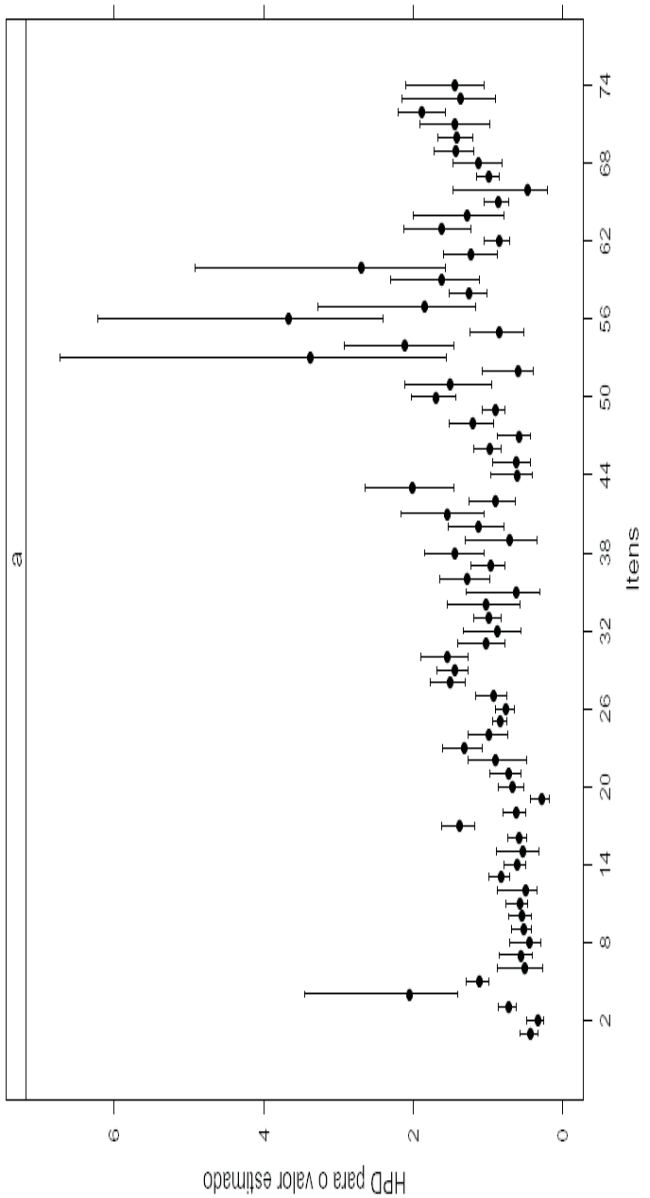


Figura 19 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2008-1 da UFLA e respectivos intervalos de credibilidade HPD a 95%

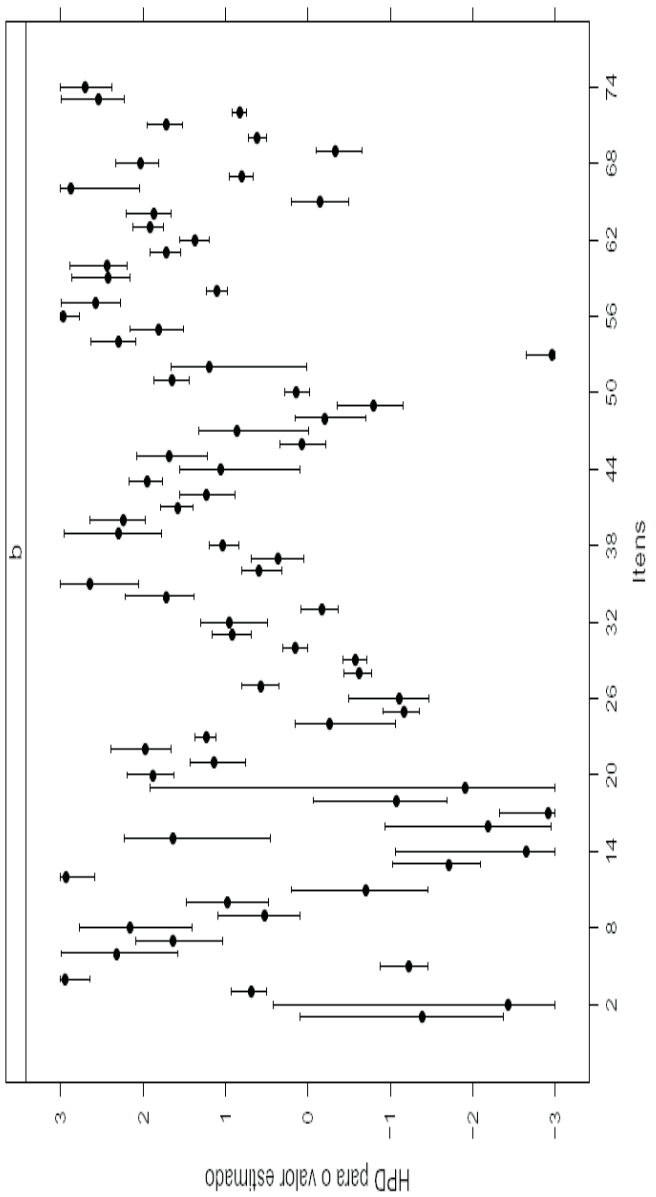


Figura 20 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2008-1 da UFLA e respectivos intervalos de credibilidade HPD a 95%

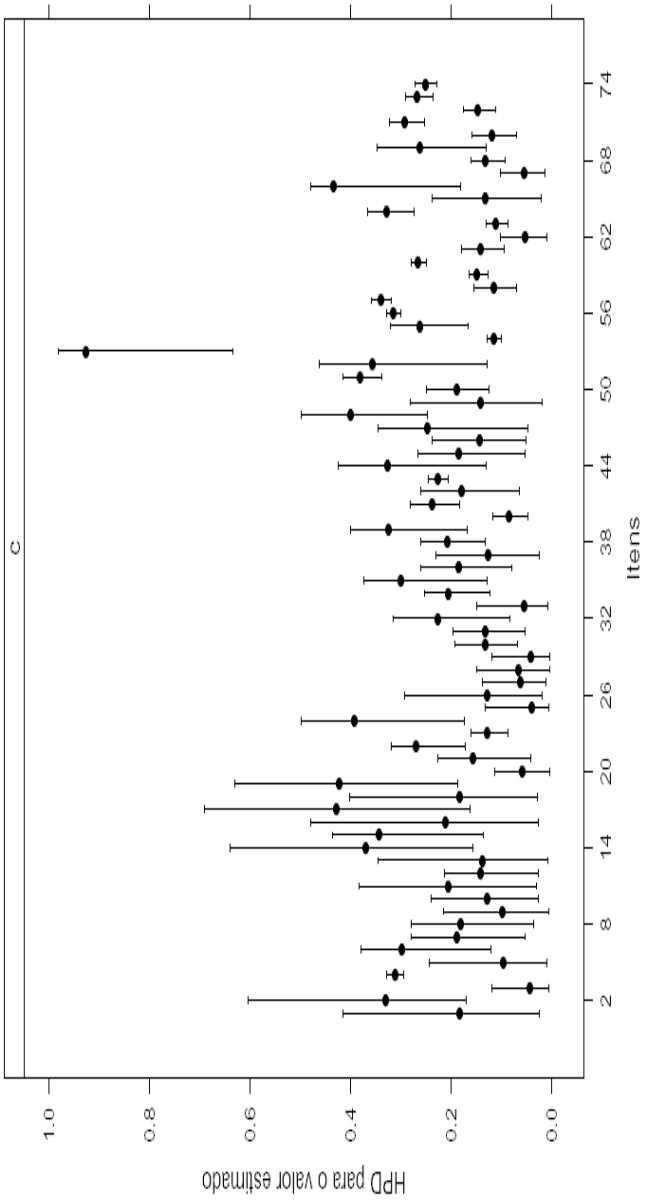


Figura 21 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2008-1 da UFLA e respectivos intervalos de credibilidade HPD a 95%

anulada, todos os candidatos recebem 1 ponto para ela, independente da resposta. Assim, a estimativa para seu grau de dificuldade, já que todos a acertam é baixa. No entanto, como qualquer alternativa assinalada está "correta", a probabilidade de "acertá-la" (ganhar ponto para esta questão) por candidatos com baixa habilidade tem que ser altíssima mesmo, o que, por meio da Figura 21, observa-se que foi praticamente 100%.

Em relação ao parâmetro *b*, pode-se observar que as provas mais difíceis tendem a ser as que apresentam maior poder de discriminação. Estas provas, para este vestibular de 2008-1, foram as de Espanhol, Inglês, Biologia, Física, Matemática e Química, as quais encontram-se representadas à direita das Figuras 19 e 20, referentes aos itens de 27 para cima.

Quanto ao parâmetro *c*, 37 itens tiveram intervalos que não abrangeram o valor 0,25, ou seja, 50% dos itens possuem valores de *c* estatisticamente diferentes de 0,25, sendo que 31 deles estão abaixo desse valor e apenas 6 acima.

Na Figura 22 está representado o histograma das habilidades estimadas dos candidatos ao vestibular 2008-1 juntamente com o gráfico da FII e da CCI com seu respectivo intervalo de credibilidade HPD para três itens, o item 4 (Português), que foi um dos que apresentou maior valor para o parâmetro *a*, o item 2 (Português), que foi um dos que apresentou menor valor e o item 72 por ter sido considerado um item bom. Outros itens apresentaram essas mesmas características, como o caso dos itens 56 e 60, citados acima, que, assim como o item 4 também apresentaram alto poder de discriminação, elevado grau de dificuldade e elevada probabilidade de acerto por indivíduos com baixa habilidade. No entanto, como a representação gráfica de todos eles é semelhante, a apresentação de apenas um deles é suficiente para exemplificar o que se deve ou não existir num vestibular. Assim, foi escolhido o item 4 para representar os que apresentaram maior poder de discriminação, foram

mais difíceis, mas tiveram alta probabilidade de acerto por indivíduos com baixa habilidade.

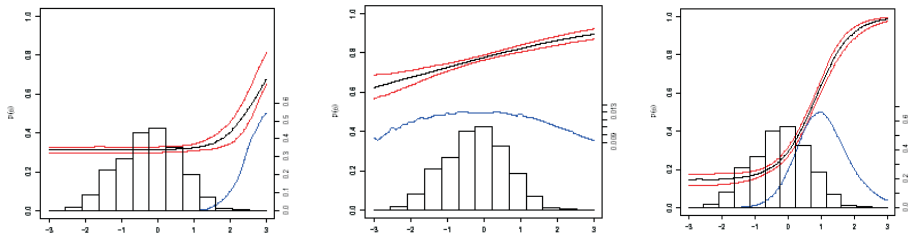


Figura 22 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para exemplos de um item muito difícil (item 4), um item muito ruim (item 2) e um item bom (item 72), do vestibular 2008-1, dispostos nessa ordem

De acordo com essa Figura, observa-se que:

1)item 4 (Português: a = 2,29; b = 2,86; c = 0,31) - é um item que possui bom poder de discriminação mas é muito difícil. Há poucos candidatos com habilidade suficiente para acertar essa questão;

2) item 2 (Português: a = 0,35; b = -1,57; c = 0,38) - é um item fácil para o grupo mas com baixa discriminação e pouco conteúdo de informação, sendo, portanto, um item muito ruim;

3) item 72 (Química: a = 1,89; b = 0,83; c = 0,14) - item bom. Possui boa discriminação e grau de dificuldade compatível com um grupo razoável dos melhores candidatos.

Na Figura 23 encontram-se representadas a curva da FIT e da CCT com seu respectivo intervalo de credibilidade HPD para cada disciplina juntamente com o histograma das estimativas das habilidades, do vestibular 2008-1.

A prova do vestibular 2008-1 que obteve o maior valor para o máximo

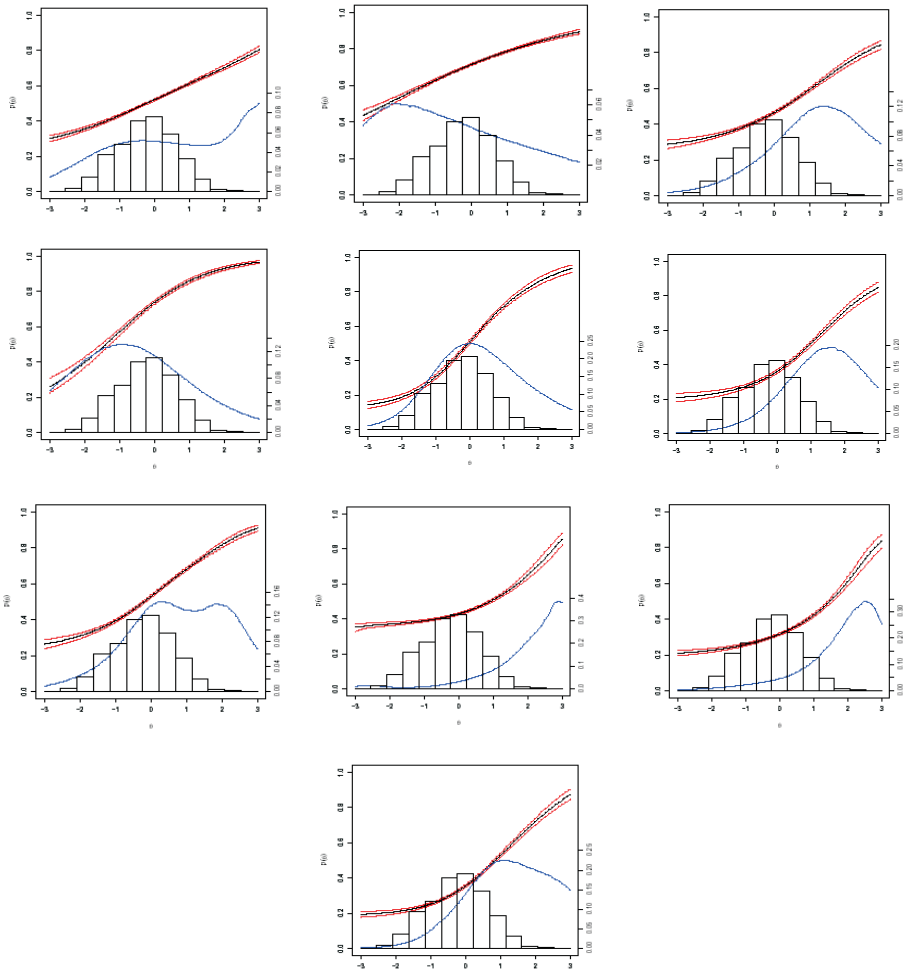


Figura 23 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2008-1, dispostas nessa ordem

da FIT foi a prova de Física. No entanto, seu grau de dificuldade foi superior ao nível de habilidade da grande maioria dos candidatos. Consequentemente, sob a

região mais informativa dessa curva não há praticamente nenhum grupo de habilidades. As provas de Geografia e de Português também apresentaram baixo poder de discriminação apesar de fáceis para o grupo de candidatos. Como a FIT está diretamente relacionada a esses dois parâmetros, pode-se perceber, também, que foram pouco informativas. Provas um pouco mais difíceis, mas condizentes com um grupo razoável dos melhores candidatos, que apresentaram bom poder de discriminação e, consequentemente, foram mais informativas para o vestibular de 2008-1, foram as provas de Matemática, Química, Inglês, Biologia e História, nessa ordem.

Quanto às provas de Língua Estrangeira, pode-se verificar que, nesse vestibular de 2008-1, a prova de Espanhol foi mais informativa para candidatos com habilidade média, pois seu grau de dificuldade está de acordo com habilidades medianas. Pela inclinação da CCT observa-se que seu poder de discriminação permite separar os candidatos que estão acima e abaixo da média da população. A prova de Inglês também apresentou bom poder de discriminação, porém, com uma dificuldade maior. No entanto, essa dificuldade não ultrapassou o nível de habilidade dos melhores candidatos, sendo, portanto, mais informativa para o grupo de interesse.

Todos os candidatos ao vestibular 2008-1 tiveram suas habilidades estimadas e convergências obtidas. A média dessas habilidades foi 0 (zero) e o desvio padrão 0,91. Houve 700 indivíduos com habilidade superior a 0,91, que corresponde a um desvio padrão acima da média dos candidatos. A correlação com as notas foi de 0,96. Na Tabela 9, encontram-se a estimativa do menor e do maior valor encontrado.

Dentre os candidatos ao vestibular 2008-1 da UFLA, 2078 optaram pela disciplina de Espanhol, correspondente a 49,42%. A média desse grupo foi de -

Tabela 9 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2008-1 da UFLA e seus respectivos erros de Monte Carlo (EMC)

Candidato	Habilidade estimada	I.C. - HPD		EMC.
		inferior	superior	
2605	-2,70	-3,88	-1,48	0.021
1049	3,06	2,38	3,72	0.013

0,21 e o desvio padrão de 0,86. Os que optaram pelo Inglês foram 2127 candidatos, com uma média de 0,20 e desvio padrão de 0,90, o que representa 50,58% do total.

As médias das habilidades estimadas de cada curso do vestibular 2008-1 e seus respectivos desvios padrão estão representados na Tabela 10.

Tabela 10 Médias das habilidades por curso e respectivo desvio padrão (sd), referentes ao vestibular 2008-1

Curso	Média	sd	Curso	Média	sd
AD	-0,08	0,86	ED	-0,69	0,76
AG	0,05	0,85	EF	0,19	0,83
AL	0,32	0,89	MA	-0,28	0,88
CC	0,15	0,94	MV	0,28	0,90
SI	-0,35	0,75	QI	-0,07	0,74
CB	0,31	0,86	ZO	-0,11	0,76
EA	-0,11	0,81			

Observa-se que, no vestibular 2008-1, o curso que obteve menor habilidade média foi o de Educação Física e a maior o de Engenharia de Alimentos. Também para este vestibular as variâncias podem ser representadas por uma variância comum, pois os valores de seus desvios padrão foram semelhantes.

A Figura 24 representa os resultados do vestibular 2008-1 como um todo, isto é, possui o histograma das habilidades estimadas dos candidatos juntamente com o gráfico da FIT e da CCT com seu respectivo intervalo de credibilidade HPD.

Pode-se observar que o vestibular de 2008-1 apresentou moderado poder

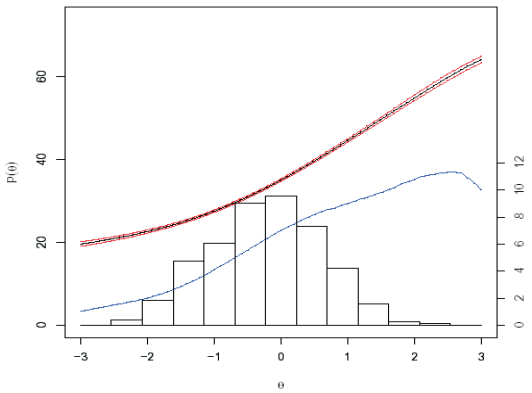


Figura 24 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), do vestibular 2008-1

de discriminação e grau de dificuldade maior que o nível médio da habilidade dos candidatos. Por exemplo, de acordo com sua CCT, espera-se que, para acertar 60% da prova (que corresponde a aproximadamente 45 questões) o candidato deve possuir habilidade igual a 1 ($\theta = 1$). O vestibular foi informativo para o grupo dos melhores candidatos, pois o gráfico da FIT se encontra à direita da média das habilidades.

3.5 Análise do vestibular 2008-2

Na Figura 25 a 27 estão representadas as estimativas dos parâmetros dos 74 itens do vestibular 2008-2, com seus respectivos intervalos de credibilidade HPDs. No que diz respeito à convergência, todos convergiram, de acordo com os métodos adotados.

Analisando a Figura 25 pode-se observar que, para o vestibular 2008-2, houve um grupo de itens que apresentaram estimativas pontuais mais elevadas

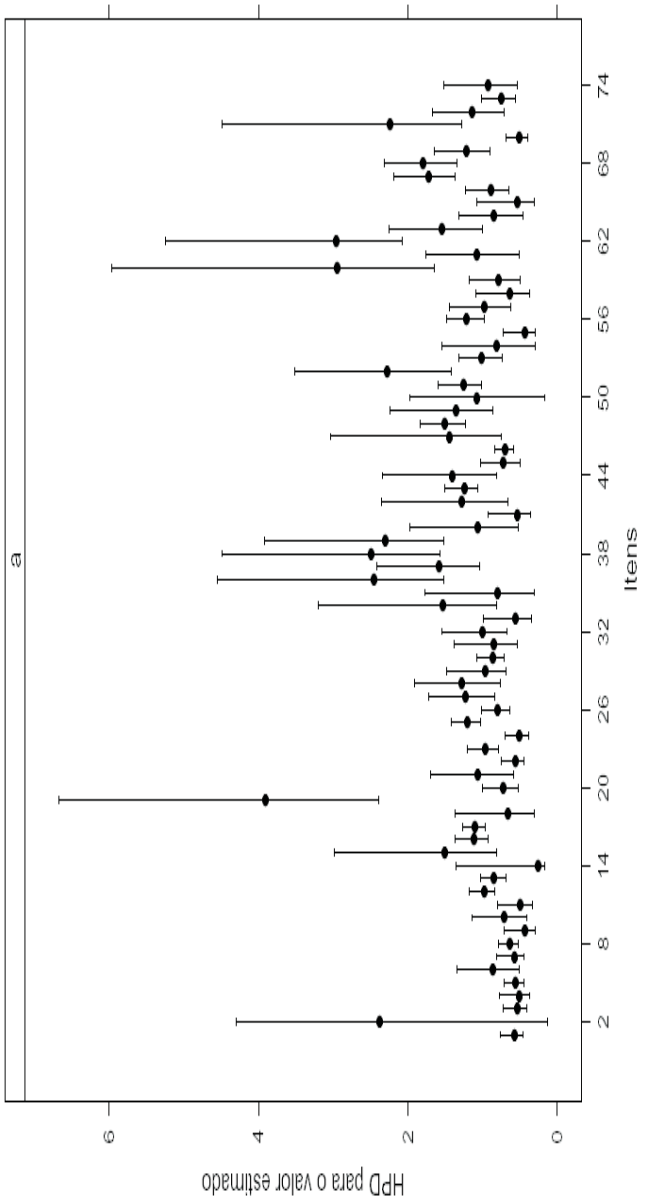


Figura 25 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2008-2 da UFLA e respectivos intervalos de credibilidade HPD a 95%

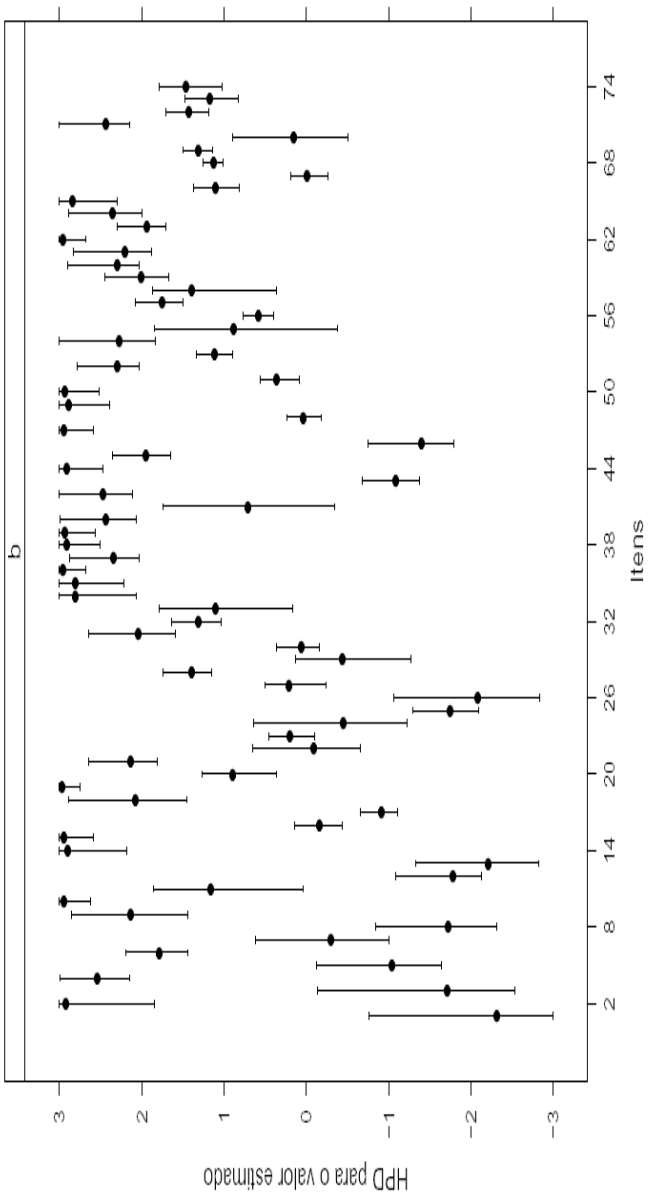


Figura 26 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2008-2 da UFLA e respectivos intervalos de credibilidade HPD a 95%

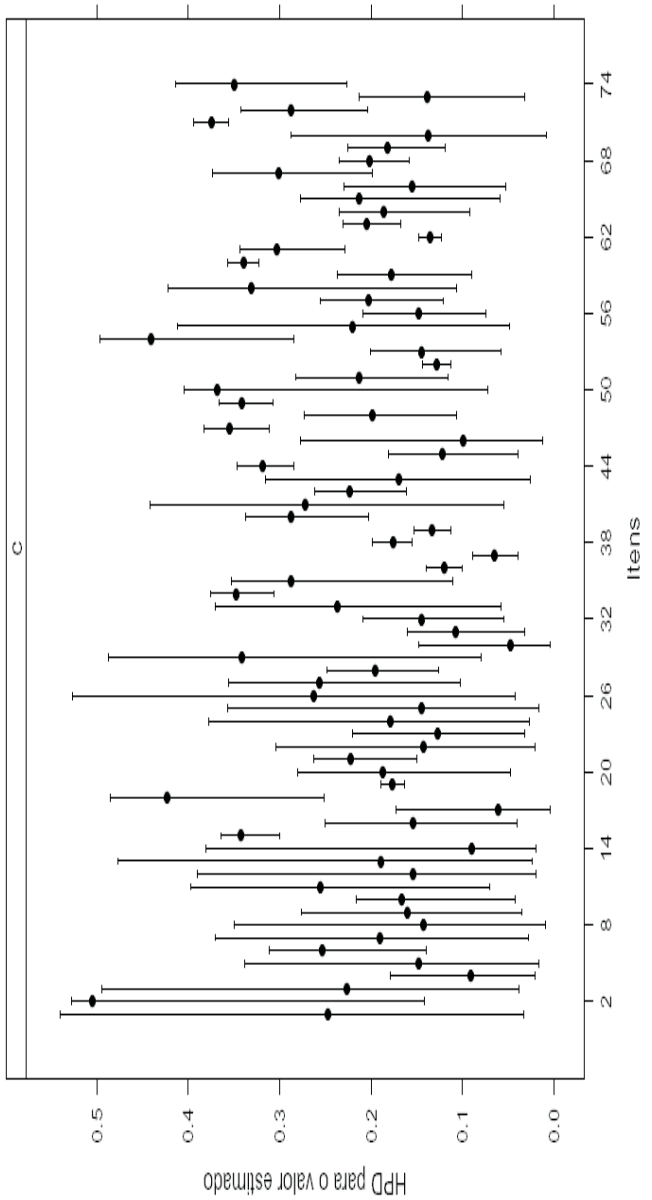


Figura 27 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2008-2 da UFLA e respectivos intervalos de credibilidade HPD a 95%

para o parâmetro *a* em relação aos demais itens. Estes itens foram os de número 2 (Português), 19 (História), 36, 38 e 39 (todos de Inglês), 52 (Física), 60 e 62 (ambos de Matemática) e 71 (Química). Atentando-se para a Figura 26, nota-se que todos eles tiveram grau de dificuldade elevado, sendo que os itens 2, 60 e 71 também apresentaram elevada probabilidade de acerto por indivíduos com baixa habilidade. De forma geral, o item que apresentou menor estimativa pontual para o parâmetro *a* foi o item 14 (Geografia), que foi um item fácil e de baixa probabilidade de acerto por indivíduos com baixa habilidade, e maior o item 19 (História), também com baixa probabilidade de acerto por indivíduos com baixa habilidade. De acordo com a classificação apresentada na seção 2.1.2 do capítulo 1, houve apenas 1 item com muito baixa discriminação, 16 com baixa discriminação, 38 com discriminação moderada, 8 com alta e 11 com muito alta.

Para o parâmetro *b*, o item mais fácil foi o item 1 (Português) e o mais difícil o item 19 (História). Por meio da Figura 26, referente a esse parâmetro, também pode-se observar que do número 34 para cima, assim como tem ocorrido nos vestibulares analisados até agora, encontram-se os itens mais difíceis, os quais se referem às provas de Inglês, Biologia, Física, Matemática e Química.

Quanto ao parâmetro *c*, houve 34 itens com intervalos que não abrangeram o valor 0,25, ou seja, 46% do total, sendo que 26 deles são menores que esse valor. Os itens com estimativas pontuais mais elevadas foram os itens 2 (Português), 18 (Geografia) e 54 (Física). O item 2 teve bom poder de discriminação mas foi muito difícil. O item 18, apesar de mais fácil que o item 2, também foi um item difícil. Apresentou poder de discriminação moderado. Como já observado, itens que não possuem boa discriminação, também não são muito informativos. O item 54, além da moderada discriminação, foi também muito difícil.

Na Figura 28 encontram-se representados, juntamente com o histograma

das habilidades estimadas dos candidatos ao vestibular 2008-2, o gráfico da função informação do item e a CCI com seu respectivo intervalo de credibilidade HPD para três itens: o item 19 (História) que foi o mais difícil e o que mais discriminou, o item 3 (Português) que foi um item fácil e com baixa discriminação, e o item 68 (Química) por ter sido um item bom.

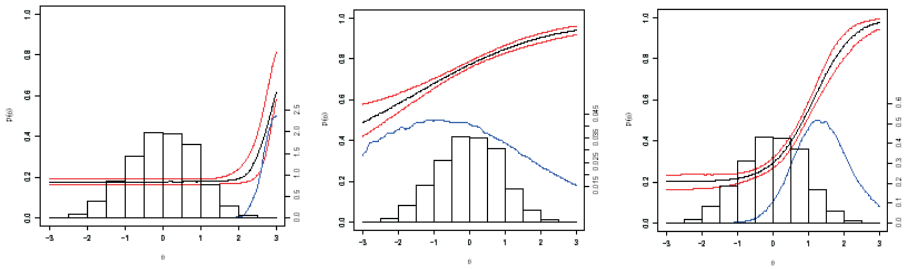


Figura 28 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para exemplos de um item muito difícil (item 19), um item muito ruim (item 3) e um item bom (item 68), do vestibular 2008-2, dispostos nessa ordem

De acordo com essa Figura, pode-se observar que:

1) item 19 (História: a = 4,30; b = 2,91; c = 0,18) - é um item muito difícil. Apesar de ter sido o que apresentou maior estimativa para o parâmetro de discriminação, essa discriminação só é ótima para indivíduos com habilidade muito elevada. Isso se deve ao fato de que esse item apresenta um elevado grau de dificuldade. Observa-se que, entre os que têm habilidade abaixo de 2, a discriminação é zero;

2) item 3 (Português: a = 0,55; b = -1,46; c = 0,26) - item muito ruim. Fornece pouco conteúdo de informação e apesar de fácil, tem baixo poder de discriminação;

3) item 68 (Química: a = 1,81; b = 1,13; c = 0,20) - item bom, pois possui boa discriminação e grau de dificuldade condizente com o nível de habilidade dos melhores candidatos.

As médias da FIT e da CCT com seu respectivo intervalo de credibilidade HPD para cada disciplina do vestibular 2008-2, junto com o histograma das estimativas das habilidades estão representadas na Figura 29.

Pode-se observar que, no vestibular 2008-2, o gráfico da FIT da prova de Inglês atinge um valor máximo superior a todas as demais disciplinas. Isto significa que se trata de uma prova muito informativa. No entanto, isso é verdade apenas para candidatos com uma habilidade muito alta, pois trata-se de uma prova muito difícil. O mesmo pode-se dizer da prova de História e quase isso, da de Matemática. Consequentemente, a discriminação entre os candidatos é praticamente nula para aqueles que não tiverem valores de habilidades muito elevados.

A prova menos informativa desse vestibular de 2008-2 foi a de Português. A mesma apresentou baixo poder de discriminação, pois indivíduos com habilidades distintas acertam praticamente a mesma quantidade de itens. Essas características também podem ser detectadas na prova de Geografia.

As provas de Filosofia, Português e Geografia tiveram alta probabilidade de acerto de forma casual. Pode-se verificar, por meio da CCT, que indivíduos com baixa habilidade acertam, aproximadamente, 40% de suas questões.

Comparando-se as duas provas de Língua Estrangeira do vestibular 2008-2, a prova de Espanhol foi mais informativa que a prova de Inglês a qual apresentou um grau de dificuldade mais condizente com o nível dos candidatos que têm habilidades acima da média. Também proporcionou boa discriminação entre eles.

Quanto às habilidades dos candidatos ao vestibular 2008-2, foi obtida uma média de -0,02 e desvio padrão de 0,87. De acordo como os critérios adotados,

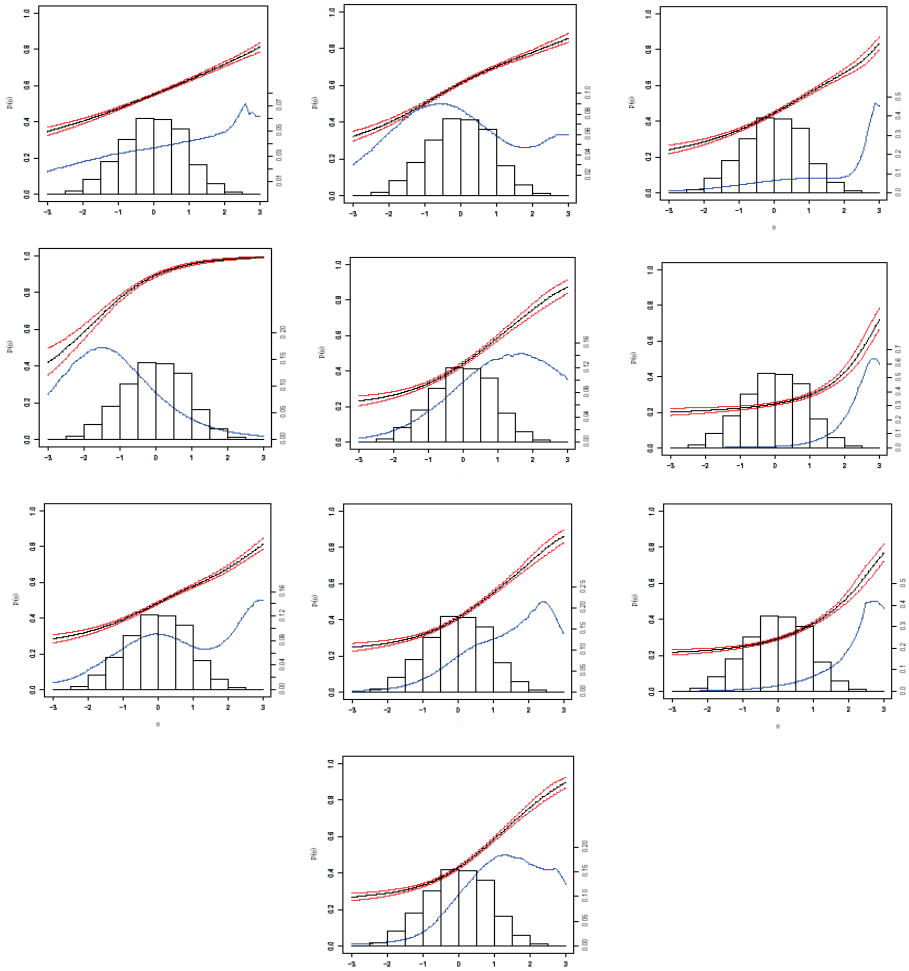


Figura 29 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2008-2, dispostas nessa ordem

todos os valores convergiram. Houve 536 candidatos com habilidades acima de 0,85 que corresponde ao valor de um desvio padrão acima da média geral. A

correlação com as notas foi de 0,95. O menor e o maior valor das estimativas obtidas para as habilidades encontram-se na Tabela 11.

Tabela 11 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidades de dois candidatos do vestibular 2008-2 da UFLA e seus respectivos erros de Monte Carlo (EMC)

Candidato	Habilidade estimada	I.C. - HPD		EMC.
		inferior	superior	
1806	-2,56	-3,79	-1,41	0.013
1527	2,52	1,95	3,09	0.007

O número dos que optaram pela disciplina de Espanhol no vestibular 2008-2 foi de 1735, o que corresponde a 53,33% dos candidatos. A média das habilidades estimadas desse grupo foi de -0,20 com um desvio padrão de 0,85. Para os que optaram pelo Inglês, houve 1518 indivíduos, representando 46,67% do total de inscritos, com habilidade média de 0,18 e desvio padrão de 0,85.

As estimativas das médias das habilidades para cada curso do vestibular 2008-2 e seus respectivos desvios padrão encontram-se na Tabela 12.

Tabela 12 Médias das habilidades por curso do vestibular 2008-2 e respectivo desvio padrão (sd)

Curso	Média	sd	Curso	Média	sd
AD	-0,03	0,89	ED	-0,60	0,66
AG	-0,09	0,79	FS	-0,29	0,76
AL	0,31	0,80	MA	-0,49	0,79
CB	0,21	0,86	MV	0,26	0,85
CC	0,36	0,89	QI	-0,24	0,92
EA	0,06	0,77	SI	-0,26	0,83
EB	-0,62	0,82	ZO	-0,18	0,81
EF	0,12	0,81			

Observa-se que, no vestibular 2008-2, o curso com menor habilidade média foi o de Educação Física e Esportes e com maior o de Ciências da Computação.

O valor do desvio padrão de cada um dos cursos não apresenta muita variação podendo ser considerada, portanto, uma variância comum para eles.

Na Figura 30 está representado o histograma das habilidades estimadas dos candidatos ao vestibular 2008-2, juntamente com o gráfico da FIT e da CCT com seu respectivo intervalo de credibilidade HPD, para o vestibular 2008-2 completo.

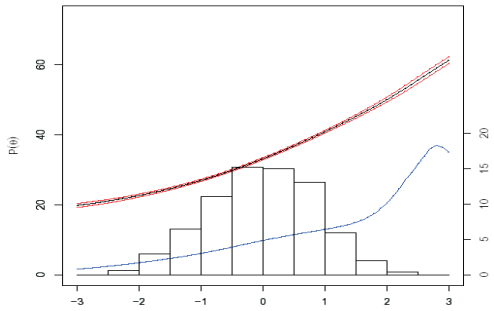


Figura 30 Histograma das habilidades, $\hat{\theta}$CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), referente ao vestibular 2008-2

Observando-se essa Figura, pode-se notar que esse vestibular de 2008-2, como um todo, foi difícil. Consequentemente, foi mais informativo para candidatos com nível de habilidade mais elevada. Nesse vestibular, indivíduos com baixa habilidade acertam, aproximadamente, apenas 20 questões. Olhando a inclinação da CCT, pode-se dizer que teve uma discriminação moderada.

3.6 Análise do vestibular 2009-1

Na Figura 31 a 33 estão representadas as estimativas dos parâmetros dos 74 itens do vestibular 2009-1 com seus respectivos intervalos de credibilidade HPDs. De acordo com os métodos de análise de convergência adotados, todos os parâmet-

ros de todos esses itens obtiveram convergência.

Por meio dessa Figura, pode-se observar que, no vestibular 2009-1, os itens que apresentaram maior poder de discriminação foram os itens 1 e 10 (ambos de Português), 12 (Geografia), 21 (História), 33 (Espanhol), 45 e 50 (Biologia), 51, 56 e 58 (de Física) e 66 (Matemática). Com exceção dos itens 10 e 50, todos esses itens também apresentaram elevado grau de dificuldade. Esse fato tem ocorrido em todos os vestibulares, ou seja, em geral os itens mais difíceis são também os que apresentam mais altos valores para o parâmetro a. Se observarmos a Figura 32 podemos verificar que os itens mais difíceis se encontram do número 33 para cima e se referem às provas de Inglês, Biologia, Física, Matemática e Química, repetindo o ocorrido nos vestibulares anteriores. Comparando com a Figura 31, referente ao parâmetro a percebe-se que essas provas são também as que apresentam maiores valores para esse parâmetro.

Itens muito difíceis foram os itens 1, 5, 8 (Português), 20, 21, 23 (História), 42 (Inglês) e 51 (Física). Destes, destaca-se o item 21 que apresentou elevado poder de discriminação e baixa probabilidade de acerto por indivíduos com baixa habilidade. Este seria um item considerado muito bom. No entanto, por meio de gráficos representativos de itens com tais características e apresentados em vestibulares anteriores, pôde-se verificar que itens com graus de dificuldade muito elevado, que superem os níveis de habilidade do grupo, não são considerados bons para selecionar os indivíduos, pois a diferença de probabilidade de acertos entre os melhores será nula ou muito pequena, dificultando saber qual deles seria o mais hábil.

Do grupo de itens citado como os que apresentaram maior poder de discriminação, excetuou-se dois deles, os de número 10 e 50, pois, apesar da estimativa elevada para o parâmetro a, foram itens muito fáceis (os dois mais fáceis de

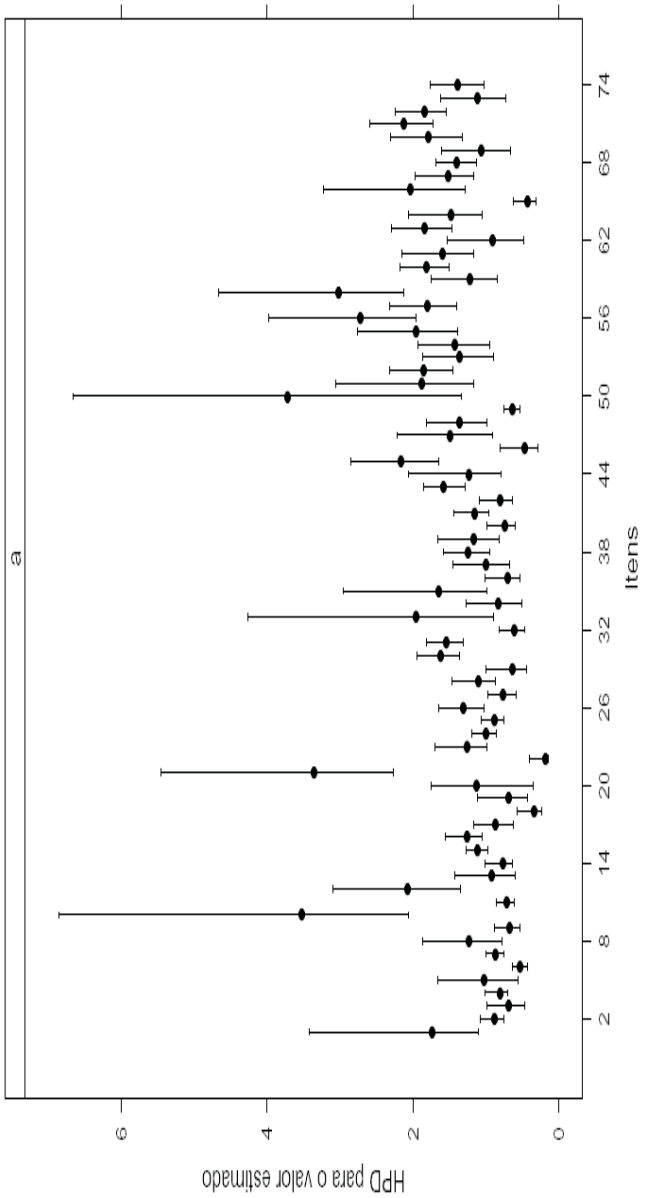


Figura 31 Estimativas pontuais do parâmetro *a* dos itens do vestibular 2009-1 da UFLA e respectivos intervalos de credibilidade HPD a 95%

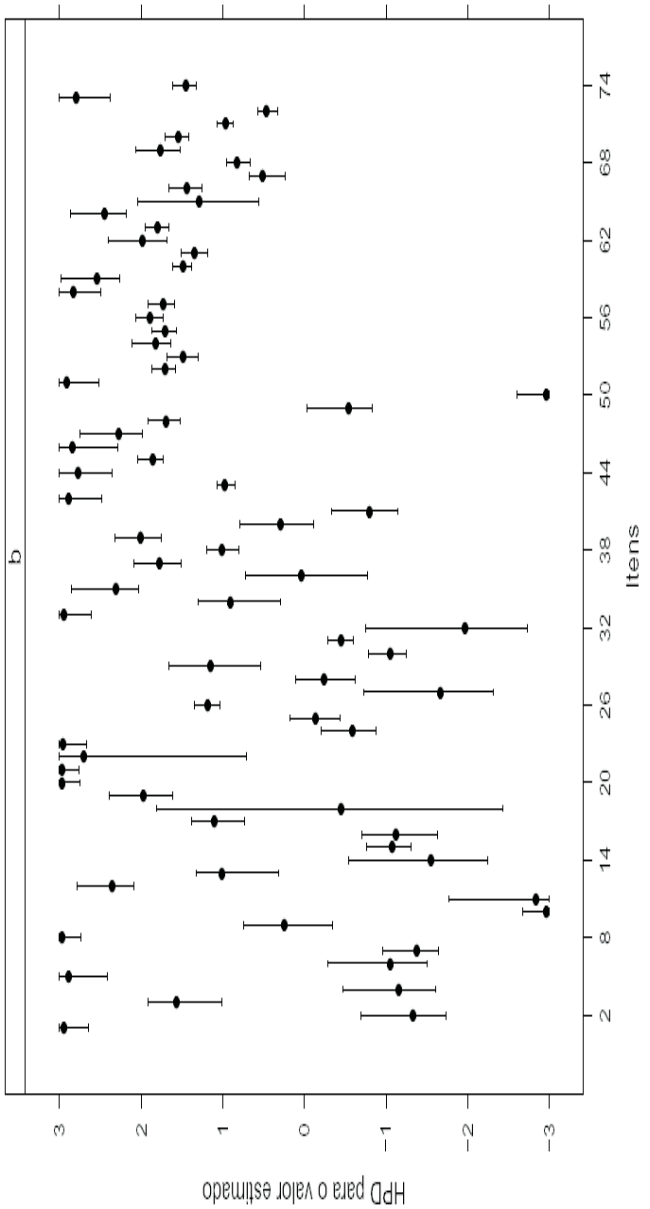


Figura 32 Estimativas pontuais do parâmetro *b* dos itens do vestibular 2009-1 da UFLA e respectivos intervalos de credibilidade HPD a 95%

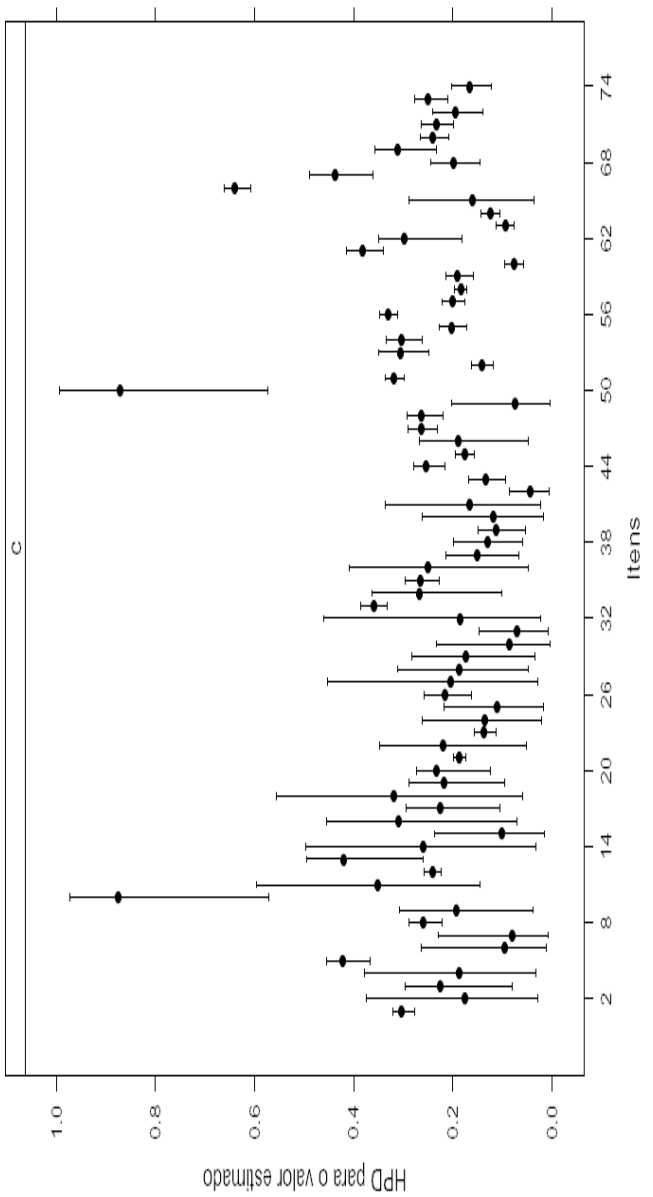


Figura 33 Estimativas pontuais do parâmetro *c* dos itens do vestibular 2009-1 da UFLA e respectivos intervalos de credibilidade HPD a 95%

todo o vestibular 2009-1). Observando a Figura 33 verifica-se que ambos apresentaram uma elevada probabilidade de acerto por indivíduos com baixa habilidade. Buscando uma explicação para essa discrepância, verificou-se que se tratava de itens que foram anulados. Assim, a explicação é a mesma dada para os vestibulares 2007-1 e 2008-1 em que tal fato também ocorreu, isto é, como todos os candidatos receberam 1 ponto para esses itens independente da alternativa assinalada, a estimativa para o parâmetro de acerto por indivíduos com baixa habilidade torna-se elevada e o grau de dificuldade apresenta-se baixo.

De acordo com a classificação apresentada na seção 2.1.2 do capítulo 1, 2 itens tiveram muito baixa discriminação nesse vestibular de 2009-1, 5 baixa discriminação, 35 discriminação moderada, 13 alta e 19 muito alta.

Quanto ao parâmetro c, nesse vestibular de 2009-1, houve 36 itens com intervalos não abrangendo o valor 0,25, correspondendo a 49% dos itens, sendo que 25 deles estão abaixo desse valor. Excluindo-se os itens 10 e 50 devido a serem questões anuladas, ainda tem-se um item que apresentou elevada probabilidade de acerto por indivíduos com baixa habilidade. Trata-se do item de número 66 (Matemática). Este item apresentou alto poder de discriminação e dificuldade razoável, porém, devido a ter elevada probabilidade de acerto por indivíduos com baixa habilidade não é um item bom. Sua representação gráfica encontra-se na Figura 34. Nessa Figura também estão representados o histograma das habilidades estimadas dos candidatos ao vestibular 2009-1, juntamente com o gráfico da FII e da CCI com seu respectivo intervalo de credibilidade HPD para mais dois itens. Esses dois itens foram escolhidos por apresentarem valores para os parâmetros a e c semelhantes (ambos com boa discriminação e baixa probabilidade de acerto por indivíduos com baixa habilidade), mas com graus de dificuldade diferentes. O intuito é tornar mais claro o que a diferença, apenas quanto ao grau de dificuldade,

pode implicar na elaboração de novos itens para provas futuras.

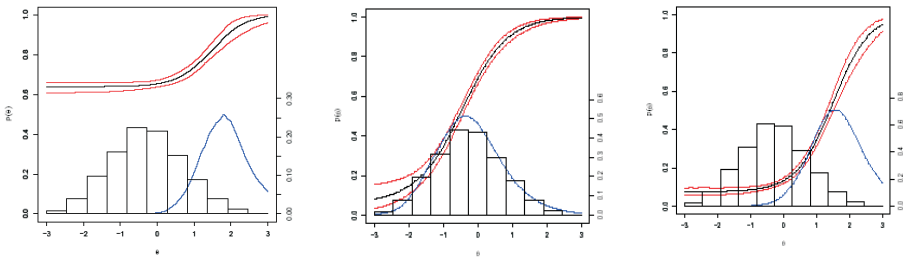


Figura 34 Histograma das habilidades, CCI dos itens (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação do item (linha azul; conteúdo de informação nas escalas à direita) para os itens 66, 31 e 60, do vestibular 2009-1

Pode-se observar por meio da Figura, que:

1)item 66 (Matemática: a = 2,19; b = 1,46; c = 0,64) - é um item ruim, pois apesar de possuir elevado poder de discriminação e grau de dificuldade condizente com um grupo razoável dos melhores candidatos, apresentou alta probabilidade de acerto por indivíduos com baixa habilidade.

Esse item é ruim porque apresenta alta probabilidade de que um indivíduo com baixa habilidade o acerte meramente pelo acaso. Observa-se que um candidato com habilidade igual a -3, que é muito baixa, tem 64% de probabilidade de acertar a questão. No entanto, como explicar o elevado poder de discriminação ($a = 2,19$)? Como já comentado, a discriminação é maior em torno do parâmetro b. Comparando-se, pois, as probabilidades de acerto de dois indivíduos que possuem habilidades em torno de 1,46 (que é o valor do parâmetro b e onde se tem maior informação sobre a habilidade do candidato) tem-se a seguinte situação: um indivíduo com nível de habilidade $\theta = 1,16$ e outro com $\theta = 1,76$. O candidato com $\theta = 1,16$ possui probabilidade de acerto igual a 0,76 e o candidato

com $\theta = 1,76$, igual a 0,88. A diferença entre essas probabilidades é de 0,12.

Vamos supor que esse item tivesse $a = 1,70$, que já é considerado muito alto e manter os mesmos valores para os outros dois parâmetros. Calculando-se as probabilidades de acerto para os mesmos indivíduos ter-se-á: para $\theta = 1,16$, probabilidade de acerto igual a 0,78; e para $\theta = 1,76$, igual a 0,86 - uma diferença de 0,08, que é menor que a anterior (o que é óbvio, pois diminui-se o valor de *a*). Vamos alterar também o valor do parâmetro *c* para 0,25, que é o valor esperado para questões com 4 alternativas e continuar com mesmo valor de *b*. Teremos, então, para $\theta = 1,16$, probabilidade de acerto igual a 0,53 e para $\theta = 1,76$, igual a 0,72, diferença de 0,19. Pode-se perceber que, diminuindo-se somente o valor de *a*, a diferença entre as probabilidades de acerto diminuiu, mas quando diminuiu-se também o valor de *c*, essa diferença entre as probabilidades voltou a elevar-se. Isso indica que quando um item possui elevado valor para o parâmetro *a*, mas também elevado valor para o parâmetro *c*, seu poder de discriminação é o mesmo de outro item que possui um valor de *a* menor mas um valor de *c* também menor, isto é, para que se possa obter o mesmo valor para a diferença das probabilidades de acerto entre esses dois indivíduos, isto é, mesma discriminação entre eles, é necessário que o valor de *a* seja maior quando o valor de *c* também aumenta. Assim, um item que tenha alta probabilidade de "chute", para que ele mantenha o mesmo poder de discriminação, tem que possuir a inclinação de sua CCI mais íngreme, isto é, tem que apresentar maior valor para o parâmetro *a*. Portanto, apesar do valor do parâmetro *a* ser diferente, seu significado é o mesmo. O valor de $a = 1,70$ para $c = 0,25$ não significa a mesma coisa que para $c = 0,64$. Da mesma forma, o valor de *a* ter sido igual a 2,19, não está significando que o item possui um exagerado poder de discriminação, nesse caso.

Essas considerações indicam que na escolha de quais são os bons itens,

deve-se atentar para suas características como um todo, ou seja, analisar os valores de todos os seus parâmetros.

Considerando essas observações, torna-se ainda mais claro compreender por que um item pode ser considerado ruim, mesmo apresentando elevado valor para o parâmetro *a*, como foi o caso do item 21 (História), destacado acima. Com dificuldade muito elevada, mesmo que haja alguns candidatos, com níveis de habilidades muito elevadas, a diferença de probabilidades de acerto entre eles será muito pequena e, portanto, será difícil concluir qual deles seria o mais hábil.

2) item 31 (Espanhol: a = 1,56; b = -0,44; c = 0,07) - item bom para discriminar entre os indivíduos que estão abaixo e acima da habilidade média, devido ao alto valor da estimativa do parâmetro *a*. Possui baixa probabilidade de acerto por indivíduos com baixa habilidade e um grau de dificuldade adequado ao nível dos candidatos com habilidade média.

A diferença entre as probabilidades de acerto de candidatos com habilidades pouco abaixo do valor de *b* e pouco acima da média, isto é, entre $\theta = -1$ e $\theta = 0.5$ é de 0,48, ou seja, é possível uma boa discriminação entre eles. Como também apresenta baixa probabilidade de acerto por indivíduos com baixa habilidade, pode ser classificado como um item bom.

3) item 60 (Matemática: a = 1,83; b = 1,49; c = 0,08) - item bom para discriminar os melhores candidatos, tendo, também um grau de dificuldade compatível com os mesmos.

Verificando-se qual seria o poder de discriminação entre os mesmos indivíduos considerados para o item 31, obtêm-se para $\theta = -1$, probabilidade de acerto igual a 0,09 e para $\theta = 0,5$, probabilidade igual a 0,20. A diferença entre essas probabilidades é muito baixa (igual a 0,11), ou seja, não é um item que discrimina bem os candidatos com habilidades abaixo da média ou em torno dela.

Mesmo entre habilidades discrepantes, como por exemplo, entre $\theta = 0$ e $\theta = -3$, essa diferença entre probabilidades de acerto será praticamente nula (igual a 0,06). Portanto, esse item não é bom para discriminar entre níveis de habilidades inferiores à média das habilidades. No entanto, para indivíduos com habilidades acima de 1, a diferença entre as probabilidades já são bem maiores.

Para comparar as características do item 31 com as do item 60, tomemos valores para $\theta = 1$ e $\theta = 2$, que são dois níveis de habilidade em torno do valor do parâmetro *b* do item 60. A diferença da probabilidade do acerto para esses candidatos será de 0,39 para o item 60 e 0.07 para o item 31, ou seja, o item 60 não é útil para distinguir entre candidatos com habilidade abaixo da média desse grupo, mas é bom para distinguir entre aqueles com maiores habilidades. Já o item 31 não é útil para distinguir entre aqueles com maiores habilidades, mas é útil para distinguir entre aqueles com habilidade abaixo da média. Assim, o conhecimento das características desses itens auxiliará na elaboração de outros novos para próximos exames.

Para uma avaliação os itens mais interessantes são, portanto, aqueles que discriminem bem, possuam baixa probabilidade de acerto por indivíduos com baixa habilidade (valores compatíveis com 0,25, no contexto do número de alternativas dos vestibulares analisados) e com diferentes graus de dificuldade a fim de que torne possível obter informações sobre os candidatos em todos os níveis de habilidade, lembrando que a informação será maior quanto mais discriminativo for o item.

É importante que numa avaliação haja questões que apresentem diferentes graus de dificuldades. Isto porque se ela for fácil, será respondida por quase todos os que estão mais preparados e por parte dos que se mostram menos preparados; se for difícil, será respondida somente por alguns dos mais hábeis. Como a dis-

criminação traduz a eficácia com que o item distingue entre os mais e os menos hábeis, desde que um item tenha boa discriminação, os diversos graus de dificuldade servirão para saber o quanto um indivíduo com baixo nível de habilidade sabe e o quanto um indivíduo com alto nível de habilidade não sabe. Comparando-se os itens 31 e 60, os quais diferiram praticamente apenas quanto ao grau de dificuldade, pode-se notar que o item 31, por apresentar dificuldade menor, será respondido por quase todos com maiores níveis de habilidades e por parte daqueles com níveis de habilidades médias. Portanto, é um item interessante para discriminar candidatos com habilidades médias. O item 60, como apresenta dificuldade maior, será útil para discriminar entre os de indivíduos com maiores habilidades.

De acordo com essas considerações, os itens mais interessantes desse Vestibular de 2009-1, que podem ser utilizados como modelos para formulação de novos itens são aqueles que apresentaram como características o ter boa discriminação e baixa probabilidade de acerto por indivíduos com baixa habilidade (ou no máximo com um valor de 0,25). Quanto ao valor do parâmetro b, numa prova é importante que se tenha questões com graus de dificuldades variados. Os itens com essas qualificações são: 12, 15, 17, 24, 25, 26, 28, 29,30, 31, 37, 38, 39, 40, 41,43, 45, 52, 55, 57, 60, 63, 64, 68, 70, 71, 72 e 74.

Desde que os itens discriminem bem e não possuam probabilidade de acerto por indivíduos com baixa habilidade elevado, mesmo alguns itens com elevado grau de dificuldade ou o extremo oposto têm sua utilidade, pois eles cumprem a função de identificar o limite até onde o mais hábil sabe e o quanto o menos hábil desconhece. Nestas condições, os mais difíceis foram: 8, 20, 21, 23, 44, 58 e 73.

Um diagnóstico completo sobre os porquês da qualidade melhor ou pior de cada item, assim como possíveis problemas de formulação, só poderão ser feitos por um conjunto de especialistas em cada uma das disciplinas. Entretanto, a iden-

tificação dessas características que a TRI proporciona muito pode auxiliar na melhoria da qualidade de futuras provas.

As médias da FIT e da CCT com seu respectivo intervalo de credibilidade HPD para cada disciplina do vestibular 2009-1, junto com o histograma das estimativas das habilidades estão representadas na Figura 35.

As mesmas discussões feitas para avaliar as características dos itens e que os classificam como tendo as devidas propriedades de interesse para os avaliadores valem para as provas como um todo.

De acordo com o que se espera de um processo seletivo, pode-se dizer que as 4 últimas provas (Biologia, Física, Matemática e Química) do vestibular 2009-1 foram as que mais atingiram seus objetivos. Isso porque foram provas bem informativas para o grupo dos melhores candidatos, tornaram possível uma boa discriminação entre eles e obtiveram um grau de dificuldade condizente com o nível de habilidade dos mesmos. As provas de Filosofia e Inglês tiveram grau de dificuldade um pouco menor, mas também foram informativas para um número razoável dos melhores candidatos. A prova de História foi muito difícil nesse vestibular.

As provas de Português e Geografia foram as provas mais fáceis do vestibular de 2009-1, apresentaram baixo poder de discriminação e elevado valor para o parâmetro c. Devido à relação que esses parâmetros têm com a função de informação, elas também foram as menos informativas nesse vestibular.

A prova de Espanhol foi boa para discriminar entre indivíduos com uma habilidade mediana.

Estimou-se as habilidades de todos os candidatos ao vestibular 2009-1, obtendo-se convergência para todos os valores. Obteve-se média geral de -0,01 com desvio padrão de 0,88. O valor da correlação entre as habilidades estimadas

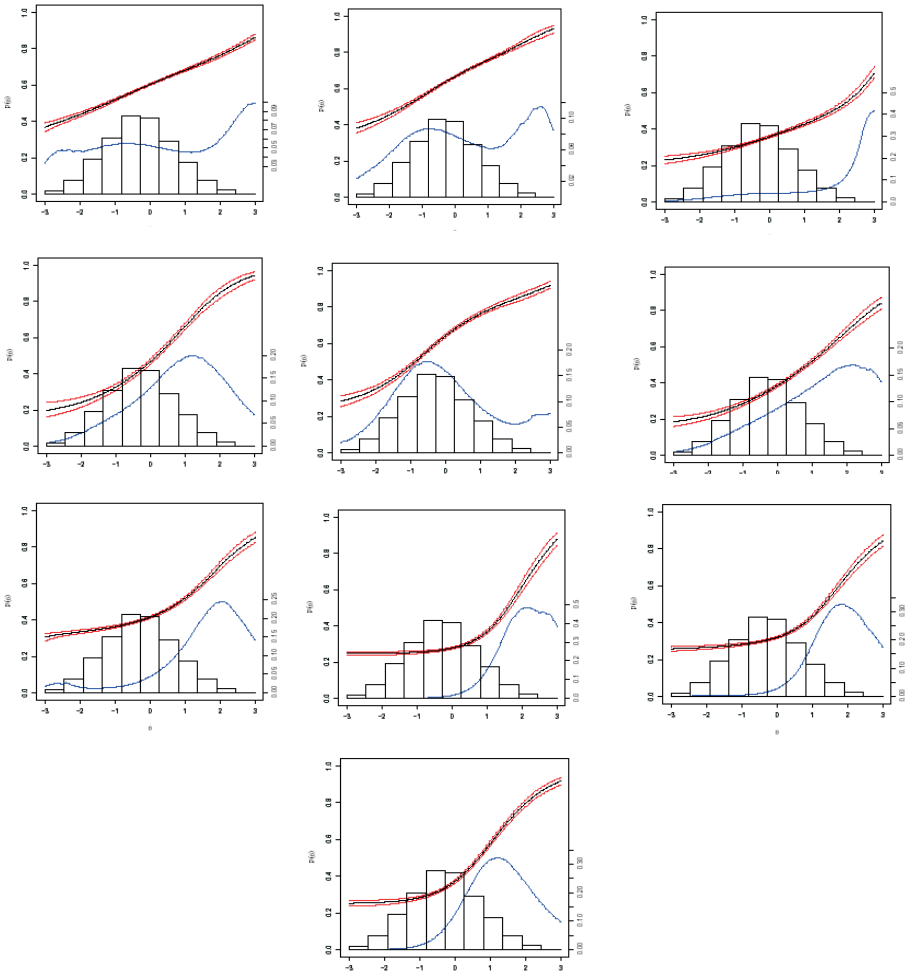


Figura 35 Histograma das habilidades, CCT por disciplina (linha preta) e seus intervalos de credibilidade HPD a 95% (linha vermelha; probabilidades nas escalas à esquerda), curva de informação por disciplina (linha azul; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês, Biologia, Física, Matemática e Química, do vestibular 2009-1, dispostas nessa ordem

e as notas foi de 0,94. Houve 680 candidatos com habilidade estimada acima de um desvio padrão da média, ou seja, acima de 0,87. Na Tabela 13 encontram-se

relacionadas a maior e a menor estimativa da habilidade.

Tabela 13 Estimativas a *posteriori* pontuais e por intervalo para os parâmetros de habilidade de dois candidatos do vestibular 2009-1 da UFLA e seus respectivos erros de Monte Carlo (EMC)

Candidato	Habilidade estimada	I.C. - HPD		EMC.
		inferior	superior	
763	-2,43	-3,57	-1,38	0.012
2689	2,74	2,20	3,26	0.006

Dos candidatos inscritos ao vestibular 2009-1, 2074 optaram pela disciplina de Espanhol (50,75%), obtendo uma habilidade média de -0,22 e desvio padrão de 0,83. Para a disciplina de Inglês houve 2013 candidatos (49,25%) com uma habilidade média de 0,20 e um desvio padrão de 0,89.

Na Tabela 14 encontram-se as médias das habilidades estimadas para cada curso do vestibular 2009-1 com seus respectivos desvios padrão.

Tabela 14 Médias das habilidades por curso do vestibular 2009-1 e respectivo desvio padrão (sd)

Curso	Média	sd	Curso	Média	sd
AD	0,01	0,82	EF	0,13	0,88
AG	-0,03	0,84	FS	-0,16	0,94
AL	0,30	0,89	MA	-0,40	0,79
CB	0,28	0,84	MV	0,32	0,88
CC	0,08	0,90	QI	-0,13	0,88
EA	0,04	0,94	SI	-0,31	0,72
ED	-0,65	0,77	ZO	-0,09	0,71

Por meio desses resultados, observa-se que, no vestibular 2009-1, o curso que obteve menor habilidade média foi o de Educação Física e a maior o de Engenharia de Alimentos. Verifica-se também que houve pouca diferença entre os valores dos desvios padrão podendo ser considerada uma variância comum a todos eles.

Pode-se constatar que, a partir do Vestibular de 2007-1, quando passou a ser oferecido o curso de Educação Física, este curso tem predominado como tendo candidatos que apresentam menores médias de habilidades. O curso de Engenharia de Alimentos também apresentou as melhores médias de habilidades em mais dois vestibulares - 2007-1 e 2008-1.

Na Figura 36 está representado o histograma das habilidades estimadas dos candidatos ao vestibular 2009-1, a curva da FIT e a CCT com seu respectivo intervalo de credibilidade HPD do vestibular todo.

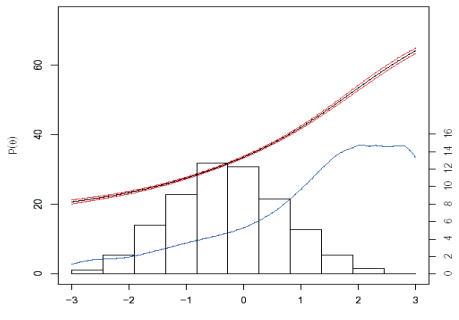


Figura 36 Histograma das habilidades, CCT (linha preta) junto ao seu intervalo de credibilidade HPD a 95% (linha vermelha; probabilidade na escala à esquerda) e curva de informação do teste (linha azul; conteúdo de informação na escala à direita), do vestibular 2009-1

Por meio dessa Figura, pode-se observar que, no vestibular 2009-1, maior informação é obtida para indivíduos com habilidade acima da média da população e em uma região onde há candidatos de interesse que sejam selecionados. O grau de dificuldade foi condizente com esse nível de habilidade. Apresentou, também, boa discriminação e baixa probabilidade de acerto por indivíduos com baixa habilidade. O número de questões respondidas corretamente por indivíduos com baixa habilidade é de aproximadamente 20.

4 CONCLUSÃO

Por meio da TRI é possível reunir elementos para discutir a qualidade das questões e das provas dos vestibulares da UFLA de 2006-2 a 2009-1 com vistas à futura melhoria da qualidade dos exames.

Pôde-se verificar que, tanto em relação aos itens quanto em relação às provas, os(as) mais informativos(as) foram aqueles(as) em que o grau de dificuldade não superou o nível do grupo dos melhores candidatos e que estes(as) foram também os(as) que mais discriminaram. Pode-se concluir que um item é mais informativo para o grupo de interesse quando possui grau de dificuldade condizente com o nível de habilidade desse grupo e possui alto poder de discriminação. Pode-se ainda concluir que:

a) questões mais difíceis são úteis para discriminar entre indivíduos com habilidades maiores, que são os de interesse para uma primeira chamada no preenchimento das vagas de um vestibular;

b) conforme diminui o grau de dificuldade, as questões passam a ser úteis para discriminar entre os candidatos com os níveis de habilidades correspondentes a essa dificuldade.

REFERÊNCIAS

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. **Teoria da resposta ao item**: conceitos e aplicações. São Paulo: Associação Brasileira de Estatística - SINAPE, 2000.

BAKER, F. B. **The basics of item response theory**. 2^{nd}. ed. Wisconsin: University of Wisconsin, 2001.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. In: LORD, F. M.; NOVICK, M. R. (Ed.). **Statistical theories of mental test scores**. Reading, MA: Addison-Wesley, 1968. p. 397-549.

BRAGION, M. L. L.; BUENO FILHO, J. S. S. Análise dos candidatos e do vestibular 2006-2, do curso de agronomia da UFLA, usando um modelo de teoria de resposta ao item (TRI). **Revista Matemática e Estatística**, São Paulo, v. 25, n. 3, p. 39-55, 2007.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, Hayward, v. 7, n. 4, p. 457-511, 1992.

HASTINGS, W. K. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika**, London, v. 57, n. 1, p. 97-109, Mar. 1970.

METROPOLIS, N. et al. Equations of state calculations by fast computing machines. **Journal of Chemical Phisics**, New York, v. 21, n. 6, p. 1087-1092, June 1953.

RAFTERY, A. L.; LEWIS, S. Comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. **Statistical Science**, Hayward, v. 7, n. 4, p. 493-497, 1992.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing, reference index: version 2.9.0. Vienna: R Foundation for Statistical Computing, 2009. Disponível em: <http://www.R0project.org>. Acesso em: 05 ago. 2009.

CAPÍTULO 4

Análise combinada dos vestibulares 2006-2 a 2009-1 da UFLA

RESUMO

Neste capítulo foi feita a análise combinada de uma série de exames vestibulares da Universidade Federal de Lavras (UFLA - Vestibular 2006-2 a 2009-1). Neste período os vestibulares ocorreram duas vezes por ano, sendo que os conhecimentos testados por cada um deles envolveram 10 disciplinas e 74 itens por exame. O número de cursos oferecidos variou de 10 até 15 com um total de 22.253 candidatos em todos os 6 exames (média de 3.700 por vestibular). O objetivo foi investigar qual a contribuição que cada prova tem dado para a seleção dos melhores candidatos, verificar progressos de tendências de informação ao longo dos anos, assim como averiguar a evolução das habilidades médias dos cursos durante o período. Os vestibulares apresentaram baixa probabilidade de acerto por indivíduos com baixa habilidade. Os resultados confirmaram alguns padrões esperados como o fato de que candidatos a cursos noturnos foram os que possuíram as menores habilidades. As melhores habilidades foram obtidas pelos candidatos aos cursos de Engenharia de Alimentos e Medicina Veterinária. Quanto aos parâmetros dos itens, o poder de discriminação e o grau de dificuldade das provas variaram de acordo com cada vestibular.

Palavras-chave: Análise combinada. Informação do item. Parâmetros do item. Vestibular.

ABSTRACT

In this chapter the combined analysis was made of a series of entrance examinations (*vestibular*) at the Federal University of Lavras (UFLA - *Vestibular* 2006-2 to 2009-1). During this period the *vestibular* occurred twice a year, and the knowledge tested by each involved 10 subjects and 74 items for examination. The number of courses ranged from 10 to 15 with a total of 22.253 candidates in all six tests (average of 3.700 per *vestibular*). The objective was to investigate which the contribution that each test has given to the selection of the best candidates, check progress of trend information over the years, as well as investigate the evolution of average abilities of the courses during the period. The *vestibular* didn't show a high probability by guess. The results confirmed some expected patterns as the fact that candidates the evening courses were those who possessed the lowest average ability. Candidates to Food Engineering and Veterinary has the highest average ability. The best abilities were obtained by the candidates to the courses of Food Engineering and Veterinary Medicine. Regarding the parameters of the items, the power of discrimination and the degree of difficulty tests varied according to each *vestibular*.

Keywords: Combined analysis. Entrance examinations (*vestibular*). Item information function. Parameters of the item.

1 INTRODUÇÃO

O principal meio de acesso ao ensino superior no Brasil nos últimos 40 anos é o processo seletivo denominado exame vestibular, que se caracteriza como uma prova de aferição dos conhecimentos adquiridos no ensino fundamental e médio. Embora polêmicos, os vestibulares são adotados tanto pelas instituições públicas quanto pelas privadas. Seu principal objetivo é selecionar, dentre os candidatos, aqueles que possuem as melhores habilidades.

Considerando apenas este aspecto técnico, em que pesem as distintas opiniões a respeito do acesso ao ensino superior e à validade dos exames vestibulares, as instituições tentam continuamente aprimorar este instrumento de seleção. Para isto é interessante observar em que medida as diferentes disciplinas têm contribuído para a seleção final, o que pode orientar eventuais alterações na estrutura das provas e no ensino médio e preparatório.

De grande utilidade, pois, é fazer um estudo de como tem sido o comportamento dessas provas ao longo dos anos, isto é, como cada disciplina tem contribuído para a discriminação desses candidatos; como tem sido a tendência do grau de dificuldade de cada uma; quanto cada uma delas tem trazido de informação sobre a habilidade dos candidatos ao longo do tempo.

A teoria de resposta ao item (TRI) tem sido usada como uma poderosa ferramenta para a avaliação educacional. Conforme Baker (1992) e Hambleton e Cook (1977), alguns de seus principais benefícios são que ela permite obter características dos itens que possibilitam identificar as questões que realmente contribuem para a avaliação do conhecimento; permite comparar o grau de dificuldade das questões assim como seu poder de discriminação e permite comparar indivíduos que não realizaram uma mesma prova.

Portanto, por meio da TRI, é possível obter as estimativas dos parâmetros dos itens de uma série de exames vestibulares de uma mesma instituição e em seguida, realizar um estudo que vise responder às indagações acima referidas.

Como um dos objetivos propostos neste trabalho, este capítulo traz o estudo dos exames vestibulares de 2006-2 a 2009-1 da Universidade Federal de Lavras (UFLA) de forma combinada a fim de comparar as provas ao longo dos anos e investigar qual a contribuição que cada uma delas tem dado para a seleção dos melhores candidatos, verificar progressos, tendências da informação ao longo dos anos, como também averiguar a evolução das habilidades médias dos cursos durante o período.

Nas próximas seções estão descritos o material e a metodologia, seguidas dos principais resultados e da discussão.

2 METODOLOGIA

2.1 Material

Os dados utilizados foram referentes aos 6 vestibulares da Universidade Federal de Lavras (UFLA). Cada vestibular constou de 74 itens divididos entre 10 disciplinas, assim distribuídos: Português (10 itens), Geografia (8 itens), História (6 itens), Filosofia (2 itens), Espanhol (8 itens), Inglês (8 itens), Biologia (8 itens), Física (8 itens), Matemática (8 itens) e Química (8 itens), cada item contendo 4 alternativas.

O número de cursos oferecidos em cada vestibular variou entre 10 a 15 e teve um total de 22.253 candidatos, distribuídos de acordo com a Tabela 1.

Tabela 1 Relação de candidatos por vagas inscritos aos diversos cursos oferecidos pela UFLA para os vestibulares 2006-2 a 2009-1

	Candidatos(Candidatos/vagas)					
Cursos	**2006-2**	**2007-1**	**2007-2**	**2008-1**	**2008-2**	**2009-1**
Administração (AD)	284(11,1)	380(23,0)	2959(11,4)	388(25,6)	299(8,9)	387(18,4)
Agornomia (AG)	651(9,2)	826(18,1)	580(7,9)	751(17,3)	602(6,7)	807(14,3)
Eng. Alimentos (AL)	229(10,2)	288(19,9)	223(10,1)	288(21,1)	227(10,5)	289(20,6)
Ciênc. Biológicas (CB)	326(15,2)	424(30,1)	316(14,1)	421(32,3)	288(12,9)	371(13,5)
Cinc. Computação (CC)	235(17,9)	252(19,3)	214(9,6)	263(19,6)	176(5,1)	263(11,5)
Ed. Física (ED)	-	284(10,2)	220(5,1)	293(22,6)	52(2,0)	221(7,2)
Eng. Agrícola (EA)	154(6,5)	121(8,2)	113(5,0)	135(9,9)	118(5,3)	138(9,9)
Eng. Florestal (EF)	238(10,8)	285(19,5)	273(12,1)	339(26,7)	271(12,5)	317(23,0)
Física (FS)	-	-	-	-	58(2,2)	44(2,3)
Matemática (MA)	-	153(6,1)	107(3,1)	97(7,4)	70(2,5)	63(3,9)
Med. Veterinária (MV)	604(26,3)	639(47,4)	529(23,5)	663(51,8)	543(23,9)	686(24,4)
Química (QI)	104(4,7)	135(8,7)	118(5,6)	113(8,1)	90(4,0)	108(7,9)
Sist. Informação (SI)	-	280(11,8)	190(5,6)	244(12,0)	207(5,7)	196(8,7)
Zootecnia (ZO)	228(10,0)	266(18,5)	164(7,2)	210(16,0)	155(7,0)	197(7,1)
Ed. Fís. e Esportes (EB)	-	-	-	-	97(3,9)	
Total	3053	4333	3342	4205	3253	4087

Para a obtenção dos resultados utilizou-se um computador equipado com processador Core i7 - 965 - 3.20 ghz com 12 gb de memória RAM.

2.2 Metodologia

Utilizou-se as estimativas de cada parâmetro dos itens, obtidas através dos procedimentos desenvolvidos no capítulo anterior (seção 5.1 a 5.6). Para combinar essas estimativas de parâmetros de item dos diferentes vestibulares, fez-se o uso da Análise de Variância (ANAVA) e de testes de médias para estudar diferenças significativas. Quando necessário foi empregado, para comparações múltiplas, o teste de Tukey.

Para a realização da ANAVA utilizou-se o seguinte modelo de Quadro:

Variável	gl	SQ	QM	F_c	valor-p
Vestibular	5	SQVestibular			
Prova	9	SQProva			
Vest*Prova	45	SQVest*Prova			
Resíduos	384	SQResíduos			

em que:

a) gl: graus de liberdade;

b) SQ: soma de quadrados;

c) QM: quadrado médio, dado por SQ/gl;

d) F_c: teste F calculado. É o quociente entre o QM da variável em análise e o QM do resíduo;

e) valor-p: trata-se de uma probabilidade e representa o nível de significância do teste F de Snedecor.

As variáveis analisadas foram os 6 vestibulares, as 10 provas e a interação entre elas. As repetições se referem ao número de itens de cada prova, totalizando 74, sendo, portanto, o grau de liberdade total igual a 443 $[(74 \times 6) - 1]$.

O modelo estatístico utilizado para esse delineamento é dado por:

$$z_{ijk} = \mu + V_i + P_j + VP_{ij} + \varepsilon_{ijk}$$

em que:

a) z_{ijk}: é o valor observado do item *k* da prova *j* no vestibular *i*;

b) μ: é uma constante inerente a toda observação;

c) V_i: se refere ao vestibular (*i* = 1, ..., 6);

d) P_j: se refere à prova (*j* = 1, ..., 10);

e) VP_{ij}: é o efeito da interação entre os vestibulares e as provas;

f) ε_{ijk}: representa o erro experimental.

Maiores detalhes sobre a ANAVA e os testes utilizados podem ser encontrados em Montgomery (2001).

Análises semelhantes foram realizadas considerando o máximo da informação que cada item forneceu e também da informação ponderada pelas habilidades dos candidatos. Procurou-se, com essas análises, verificar progressos de tendências de informação por prova ao longo dos anos.

Para maior visualização dos resultados, foram plotados gráficos das provas ao longo dos vestibulares. O mesmo procedimento foi realizado para identificar a evolução das habilidades médias ao longo dos anos. A ANAVA também foi utilizada para detectar diferenças entre os parâmetros dos itens dos vestibulares realizados no primeiro e no segundo semestre.

Os resultados obtidos dessas análises estão apresentados na próxima seção.

3 RESULTADOS

Na Tabela 2 estão representados os resultados da análise da variância para as provas e os vestibulares com relação ao parâmetro a.

Tabela 2 Quadro-resumo da ANAVA das provas e vestibulares da UFLA de 2006-2 a 2009-1 com relação ao parâmetro a

Variável	gl	SQ	QM	F_c	valor-p
Vestibular	5	47,30	9,46	15,02	1,82e-13***
Prova	9	43,99	4,89	7,76	1,66e-10***
Vest*Prova	45	73,93	1,64	2,60	4,41e-07***
Resíduos	384	242,07	0,63		

*** significativo à 0,001% pelo teste F

Observa-se que a interação entre provas $\times$ vestibular foi significativa. Assim sendo, não se pode dizer de forma geral, qual prova ou qual vestibular mais discriminou ao longo desses anos. Procedeu-se, pois, às comparações entre as médias das provas. A apresentação destes resultados encontra-se na Figura 1 por meio da qual pode-se verificar, também, a tendência na variabilidade das provas ao longo dos vestibulares.

Pode-se observar por meio dessa Figura (1) que, para o vestibular 2006-2, a prova de Biologia foi a mais discriminativa, não diferindo estatisticamente, apenas da prova de Química. As provas de Química, Física, Matemática e Inglês também apresentaram igualdade estatística quanto ao poder de discriminação. Em escala descendente de valores, a prova de Geografia foi a que obteve menor valor.

Atentando-se para o comportamento de cada disciplina ao longo dos anos, observa-se que as provas de Português, Geografia, História, Filosofia e Espanhol não apresentaram nenhuma tendência, mantendo-se mais constantes quanto às estimativas desse parâmetro (a). A prova de Biologia foi muito discriminativa no vestibular 2006-2, caiu bastante nos 2 vestibulares seguintes e apresentou uma

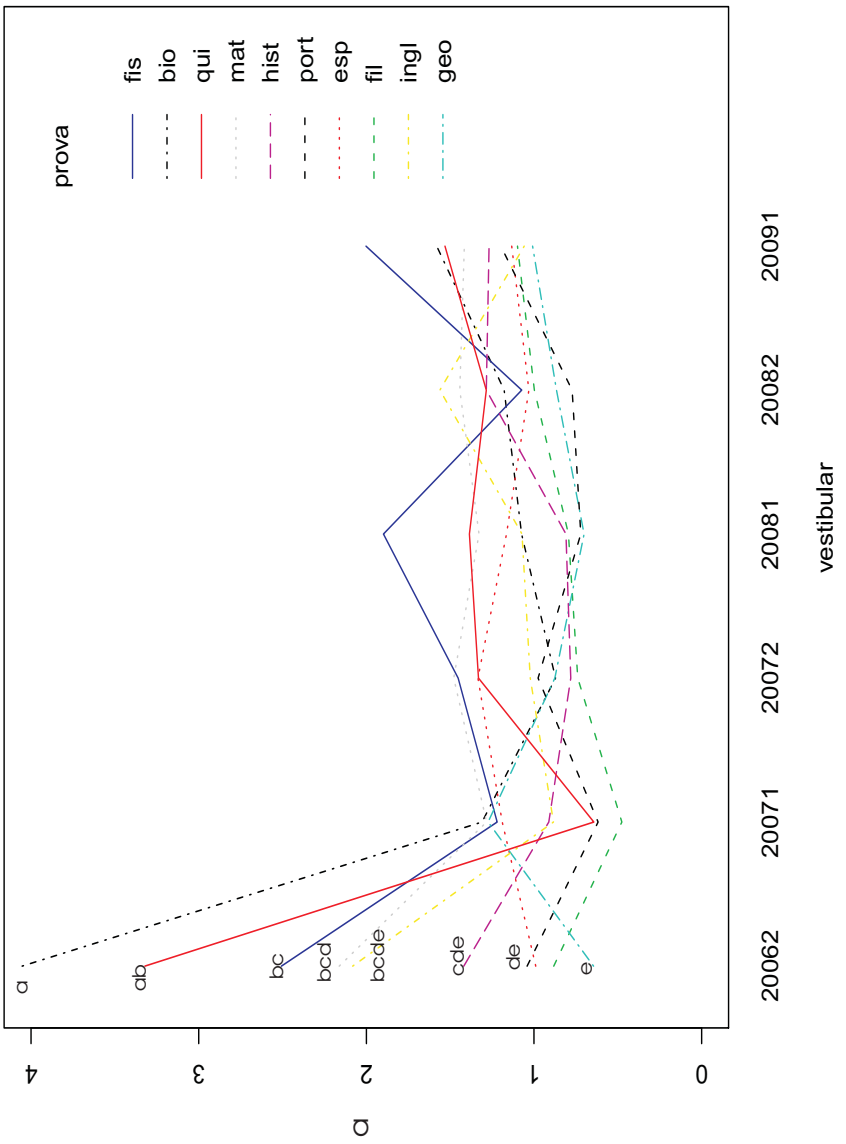


Figura 1 Poder de discriminação das provas dos vestibulares de 2006-2 a 2009-1 da UFLA. Letras iguais na coluna indicam que não há diferenças significativas pelo teste de Tuckey. Para cursos que se encontram sem letras, repetem-se as letras do curso imediatamente superior

tendência ascendente a seguir. A prova de Química foi também bastante discriminativa no vestibular 2006-2, caiu no vestibular 2007-1, subiu um pouco no vestibular 2007-2 e a partir daí, tendeu a se manter estável com uma leve ascensão em 2009-1. A prova de Física foi a que apresentou maior variabilidade na estimativa do valor desse parâmetro de discriminação.

Como apenas o vestibular 2006-2 apresentou diferenças significativas entre as provas, uma outra ANAVA foi realizada excluindo-se esse vestibular. Os resultados obtidos encontram-se na Tabela 3.

Tabela 3 Quadro-resumo da ANAVA das provas e vestibulares da UFLA de 2007-1 a 2009-1 com relação ao parâmetro a

Variável	gl	SQ	QM	F_c	valor-p
Vestibular	4	4,37	1,09	3,21	0,01*
Prova	9	16,13	1,79	5,26	1,24e-06***
Vest*Prova	36	15,25	0,42	1,24	0,17
Resíduos	320	109,89	0,34		

* significativo à 0,05%, *** significativo à 0,001% pelo teste F

Observa-se que, quando não foi incluído esse vestibular na análise (2006-2), a interação deixa de ser significativa, confirmando que esse ano é o que está sendo responsável pela interação significativa. No entanto, houve diferenças significativas para as provas. Assim sendo, isto indica que, para os últimos cinco vestibulares analisados, algumas provas são melhores que outras em geral.

Os resultados da análise da variância das provas e dos vestibulares, com relação ao parâmetro b, estão representados na Tabela 4.

Observa-se que houve diferenças significativas para a interação entre as provas e os vestibulares, o que significa que não se pode classificar o grau de dificuldade de cada prova de forma geral. Na Figura 2 está representada graficamente essa variabilidade.

Nessa Figura 2, pode-se observar que as provas da área de exatas (Física,

Tabela 4 Quadro-resumo da ANAVA das provas e vestibulares da UFLA de 2006-2 a 2009-1 com relação ao parâmetro *b*

Variável	gl	SQ	QM	F_c	valor-p
Vestibular	5	17,20	3,44	1,86	0,10
Prova	9	203,21	22,58	12,21	2e-16***
Vest*Prova	45	118,30	2,63	1,42	0,04*
Resíduos	384	711,59	1,85		

* significativo a 0,05%, *** significativo a 0,001% pelo teste F

Matemática e Química) sempre estão classificadas entre as provas mais difíceis para todos os vestibulares. Inglês também manteve uma dificuldade elevada e apresentou uma tendência em ser mais difícil que a prova de Espanhol. O grau de dificuldade das provas de Geografia, Espanhol e Filosofia, apesar de terem suas estimativas oscilando entre os vestibulares, tenderam a estar entre as provas mais fáceis ao longo dos anos. A prova mais estável foi a prova de Química.

Comparando-se essa Figura com a Figura 1, pode-se perceber uma relação entre o grau de dificuldade e o poder de discriminação. Provas mais difíceis tendem a ser as que mais discriminam.

Para o parâmetro *c*, os resultados da análise da variância das provas e dos vestibulares encontram-se na Tabela 5.

Tabela 5 Quadro-resumo da ANAVA das provas e vestibulares da UFLA de 2006-2 a 2009-1 com relação ao parâmetro *c*

Variável	gl	SQ	QM	F_c	valor-p
Vestibular	5	0,08	0,02	2,00	0,25
Prova	9	0,35	0,04	4,00	1,00e-03**
Vest*Prova	45	0,70	0,02	2,00	0,16
Resíduos	384	4,83	0,01		

** significativo à 0,01% pelo teste F

Esta análise reforça a idéia de que observa-se diferenças apenas entre as provas e que essas diferenças não estão dependendo do vestibular. Portanto, pode-

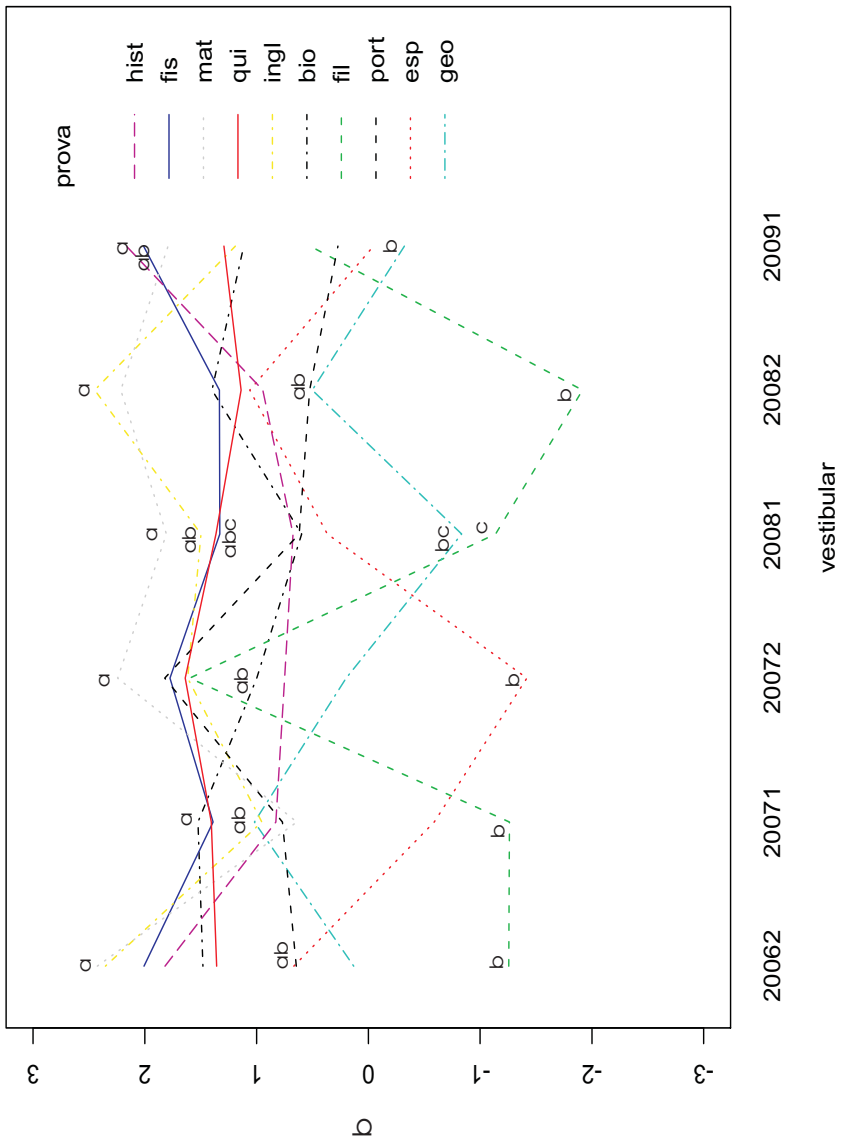


Figura 2 Grau de dificuldade das provas dos vestibulares de 2006-2 a 2009-1 da UFLA. Letras iguais nas colunas indicam que não há diferenças significativas pelo teste de Tuckey. Para cursos que se encontram sem letras, repetem-se as letras do curso imediatamente superior

se dizer, de forma geral, quais foram as que tiveram maior probabilidade média de acerto por indivíduos com baixa habilidade ao longo desses anos. Esses resultados estão apresentados na Tabela 6. Nesta mesma Tabela também se encontram as médias em relação aos parâmetros *a* e *b* de forma geral.

Para os vestibulares, a média geral foi de 0,21, valor abaixo da esperada uma vez que as questões possuem 4 alternativas.

Tabela 6 Médias das provas dos vestibulares de 2006-2 a 2009-1 da UFLA com relação aos parâmetros *a*, *b* e *c*

Provas	*a*	*b*	*c*
Port	0,89 c	0,77 bc	0,21 ab
Geo	0,90 c	0,12 cd	0,25 a
Hist	1,08 bc	1,20 ab	0,21 ab
Fil	0,83 c	-0,01 d	0,13 b
Esp	1,14 abc	0,01 cd	0,17 ab
Ingl	1,28 abc	1,68 ab	0,19 ab
Bio	1,68 a	1,19 ab	0,24 a
Fis	1,69 a	1,64 ab	0,24 a
Mat	1,52 ab	1,85 a	0,22 a
Qui	1,58 ab	1,37 ab	0,21 a

Médias seguidas pela mesma letra não diferem entre si pelo teste de Tukey

Conquanto o resultado tenha apresentado ambiguidade quanto à estimativa do parâmetro de acerto por indivíduos com baixa habilidade entre as diversas provas de forma geral, dificultando dizer qual delas teve maior valor para esse parâmetro, pode-se fazer uma análise quanto a tendência de suas médias. Apesar de, estatisticamente, as provas de Filosofia e Espanhol não diferirem, (Espanhol também não difere de nenhuma outra disciplina), contudo percebe-se que a média da probabilidade de acerto por indivíduos com baixa habilidade apresentada pela disciplina de Filosofia foi a mais baixa.

Essa situação, à primeira vista, pode parecer um pouco estranha. Como explicar que uma prova fácil (de acordo com a análise dos resultados referentes ao

parâmetro *b* anteriormente) possa ter baixa probabilidade de acerto ao acaso? O que se segue é uma tentativa de interpretar este resultado aparentemente paradoxal.

Consideremos novamente o significado do parâmetro *c*. Ele é definido como a probabilidade de que um indivíduo, com habilidade indefinidamente baixa venha acertar a questão. Ao realizar uma prova há duas possibilidades: o candidato "sabe que não sabe" e "chuta" ou ele "pensa que sabe" e não "chuta", mas erra com alta probabilidade (pois tem baixa habilidade). No caso dos vestibulares analisados, cada item possui 4 alternativas. Logo, espera-se que a probabilidade de acerto por quem nada sabe esteja em torno de 0,25. Valores maiores que esse são ruins pois indicam questões mal formuladas. Se o candidato sabe que uma das alternativas não é a resposta correta, sobram 3 outras para ele escolher aleatoriamente e isso faz com que aumente a probabilidade para que ele venha a acertar essa questão no "chute". Por outro lado, valores muito abaixo de 0,25 indicam que a probabilidade do indivíduo acertar aleatoriamente é muito baixa. Isso pode estar acontecendo devido ao fato de que o candidato pensa que sabe mas não sabe. Ele tem baixa habilidade, mas crê que entendeu o item. Consequentemente, terá a tendência de assinalar uma alternativa errada achando que acertou, induzido por algum conhecimento superficial do assunto. Logo, a probabilidade de acertar essa questão é baixa.

A probabilidade de acerto por indivíduos com baixa habilidade, portanto, tem a ver com o nível de habilidade do candidato. Se ele "chuta" sabendo que não sabe tem mais chance de acertar do que se ele não "chuta" por achar que sabe mas não sabe. Assim sendo, faz sentido uma prova considerada fácil ter baixo valor para o parâmetro *c*. A Figura 3 abaixo, comparada com as duas anteriores (Figura 1 e 2), deixa mais clara essa relação entre os parâmetros.

Pode-se verificar que as provas mais difíceis tendem a ter probabilidade de

acerto casual dentro do esperado para questões com 4 alternativas (0,25) e provas mais fáceis, têm baixa probabilidade de que um candidato com baixa habilidade venha acertar suas questões. Entretanto, vale frisar que valores acima de 0,25 para esse parâmetro podem indicar que a prova "não é boa", isto é, pode apresentar alguns problemas de formulação. No entanto, cabe observar que a média geral dos vestibulares quanto a esse parâmetro foi de 0,21, um valor abaixo do esperado (o qual seria 0,25). Sendo assim, pode-se dizer que os vestibulares da UFLA, de 2006-2 a 2009-1, apresentaram baixa probabilidade de acerto por indivíduos com baixa habilidade.

Por meio dessa Figura pode-se observar também o comportamento das provas, quanto a esse parâmetro, ao longo dos vestibulares. Nota-se que maior variabilidade foi apresentada pelas provas de Filosofia e Espanhol. A prova de Geografia teve um valor elevado no vestibular 2007-1 e Física no de 2008-1. As demais provas mantiveram-se mais estáveis e com médias abaixo de 0,25.

Uma outra análise julgada interessante de se fazer, foi a comparação entre os vestibulares realizados no primeiro e no segundo semestre, a fim de saber se havia diferenças entre seus parâmetros. Os resultados obtidos estão apresentados na Tabela 7.

Tabela 7 Médias dos parâmetros de discriminação (a), dificuldade (b) e probabilidade de acerto por indivíduos com baixa habilidade (c) das provas dos vestibulares 2006-2 a 2009-1 da UFLA, sendo consideradas por semestre

Semestre	a	b	c
I	1,42 a	1,21 a	0,21 a
II	1,16 b	0,86 b	0,22 a

Médias seguidas pela mesma letra não diferem estatisticamente pelo teste F

Observa-se que, com relação aos parâmetros a e b, os vestibulares do primeiro semestre foram mais difíceis e discriminaram mais que os do segundo.

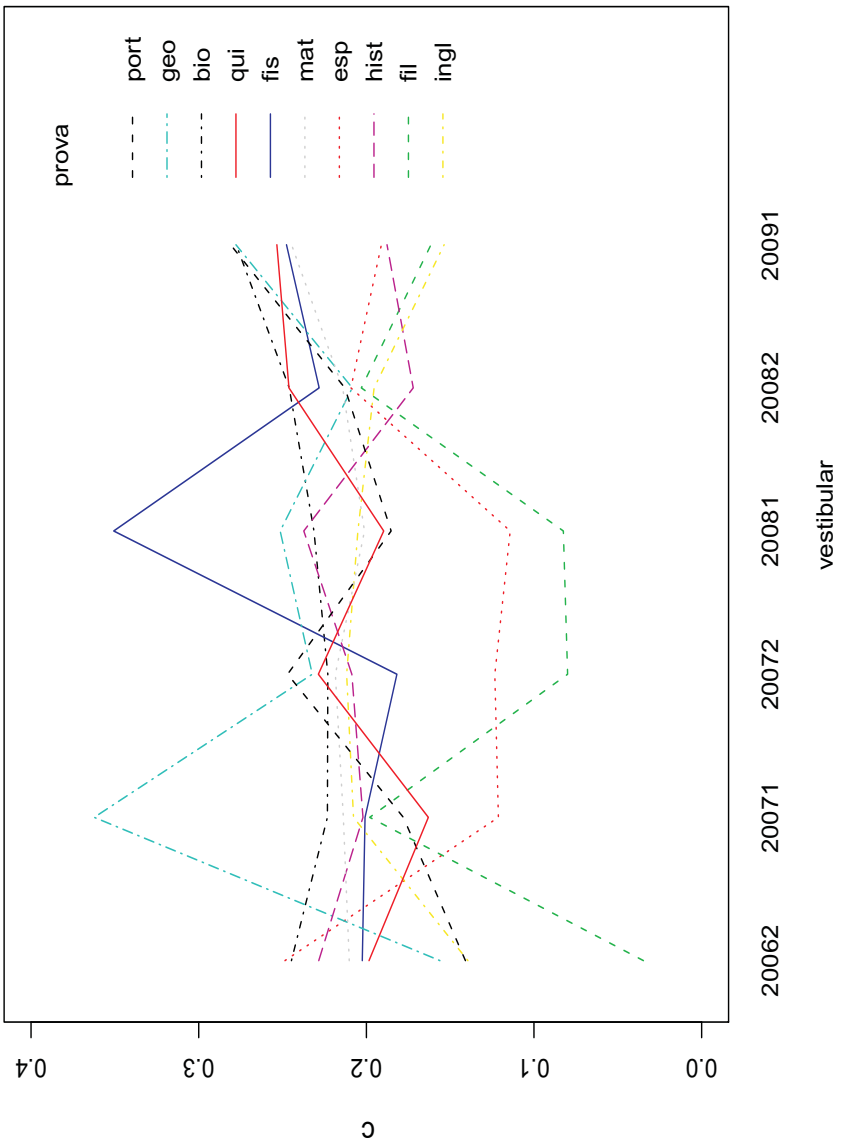


Figura 3 Probabilidade de acerto por indivíduos com baixa habilidade das provas dos vestibulares de 2006-2 a 2009-1 da UFLA

No entanto, quanto ao parâmetro *c*, não houve nenhuma diferença entre eles.

Na Tabela 8 estão os resultados da análise da variância para a informação

que cada prova forneceu, sendo considerado o máximo de seu valor para cada item.

Tabela 8 Quadro-resumo da ANAVA das provas e os vestibulares de 2006-2 a 2009-1 da UFLA com relação ao máximo da informação

Variável	gl	SQ	QM	F_c	valor-p
Vestibular	5	39,54	7,91	13,18	00,00e-04***
Prova	9	19,60	2,18	3,63	2,00e-04***
Vest*Prova	45	72,66	1,61	2,68	0,00e-04***
Resíduos	384	231,90	0,60		

* significativo à 0,001%

O resultado dessa Tabela demonstra que não se pode concluir de forma geral quais provas foram as mais informativas. A representação gráfica do comportamento de cada prova ao longo dos vestibulares, quanto a essa variável, encontra-se na Figura 4.

Pode-se observar que o único vestibular relevante heterogêneo é o primeiro (vestibular 2006-2). Para este vestibular, a prova mais informativa foi a prova de Biologia. Em seguida, foi a prova de Química. Apesar de estatisticamente não ter havido diferenças significativas, graficamente pode-se notar que as provas de Filosofia e Geografia foram, de forma geral, as menos informativas.

Como apenas o vestibular 2006-2 apresentou diferenças significativas entre a informação das provas, uma outra ANAVA foi realizada excluindo-se esse vestibular. Os resultados obtidos encontram-se na Tabela 9.

Tabela 9 Quadro-resumo da ANAVA das provas e os vestibulares de 2007-1 a 2009-1 da UFLA com relação ao máximo da informação

Variável	gl	SQ	QM	F_c	valor-p
Vestibular	4	85,72	21,43	0,97	0,43
Prova	9	163,29	18,14	0,82	0,60
Vest*Prova	36	728,48	20,23	0,91	0,62
Resíduos	320	7093,59	22,17		

Observa-se que ao ser excluído o vestibular 2006-2 da análise, tanto a in-

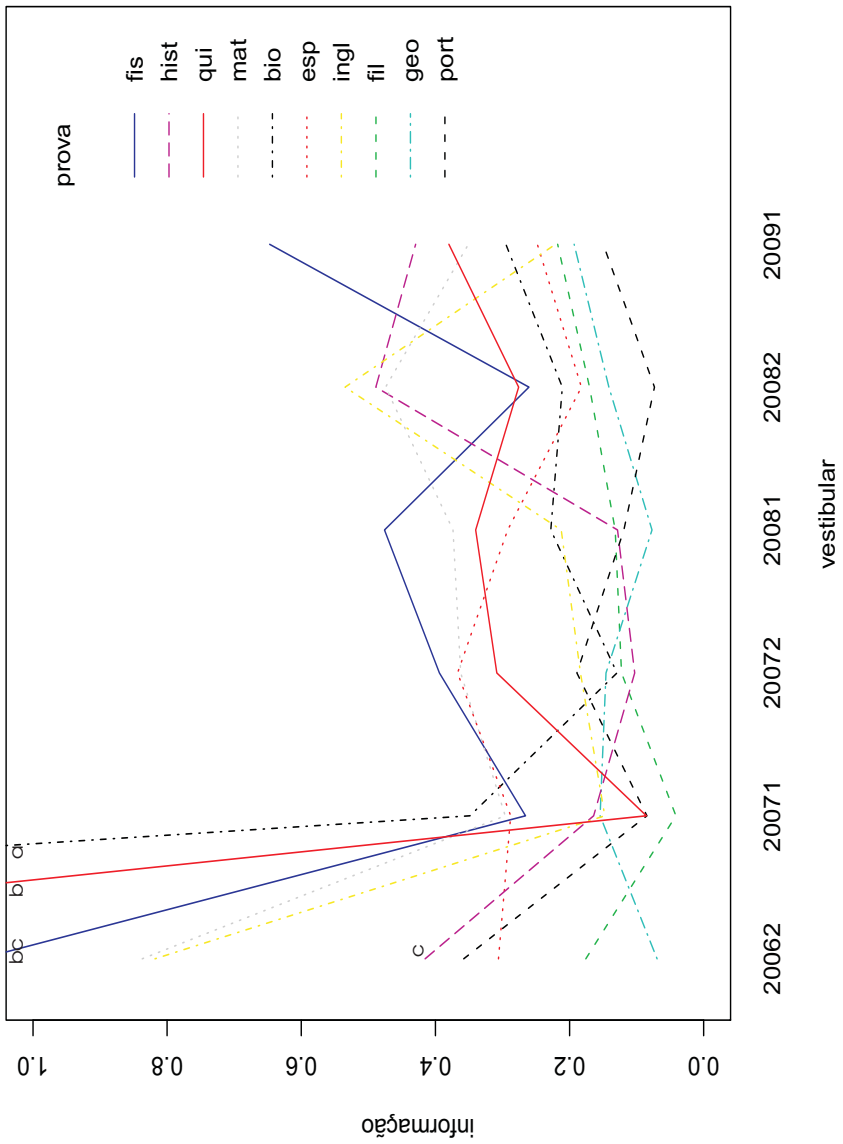


Figura 4 Gráfico do máximo de informação dadas pelas provas ao longo dos Vestibulares de 2006-2 a 2009-1 da UFLA. Letras iguais na coluna indicam que não há diferenças significativas pelo teste de Tuckey. Para cursos que se encontram sem letras, repetem-se as letras do curso imediatamente superior

teração deixa de ser significativa como não há nenhuma prova ou vestibular em particular que sejam mais informativos que os demais. Assim sendo, o vestibular que está sendo responsável pela interação significativa é apenas o de 2006-2. Nos demais vestibulares, em média, todas as provas contribuem igualmente para informação sobre o nível de habilidade dos candidatos.

A análise da variância dessa mesma variável informação, porém, considerando-a, agora, ponderada pelas habilidades dos candidatos, encontra-se na Tabela 10.

Tabela 10 Quadro-resumo da ANAVA das provas e os vestibulares de 2006-2 a 2009-1 da UFLA com relação à informação ponderada pelas habilidades

Variável	gl	SQ	QM	F_c	valor-p
Vestibular	5	16.896,18	3.379,24	3,88	1,90e-03**
Prova	9	58.228,78	6.469,86	7,44	0,00e-04***
Vest*Prova	45	121.190,94	2.693,13	3,10	0,00e-04***
Resíduos	384	334.070,52	869,98		

** significativo a 0,01%; *** significativo a 0,001% pelo teste F

A interação entre provas e vestibulares também foi verificada ser significativa aqui. Foi realizada portanto, uma análise em separado sobre o comportamento das provas em cada vestibular e ao longo deles. Seus resultados estão indicados na Figura 5.

Pode-se perceber por meio dessa Figura (5) que apenas nos vestibulares 2006-2, 2007-1 e 2008-1 houve diferenças significativas entre suas médias.

Observando a tendência ao longo dos anos, percebe-se uma estabilidade maior para as disciplinas de Português, Geografia, História, Filosofia e Física. As demais disciplinas apresentaram-se ora mais, ora menos informativas, sem contudo manifestar nenhum progresso de informação ao longo dos anos.

Apesar de terem sido encontradas diferenças quanto à informação pondera-

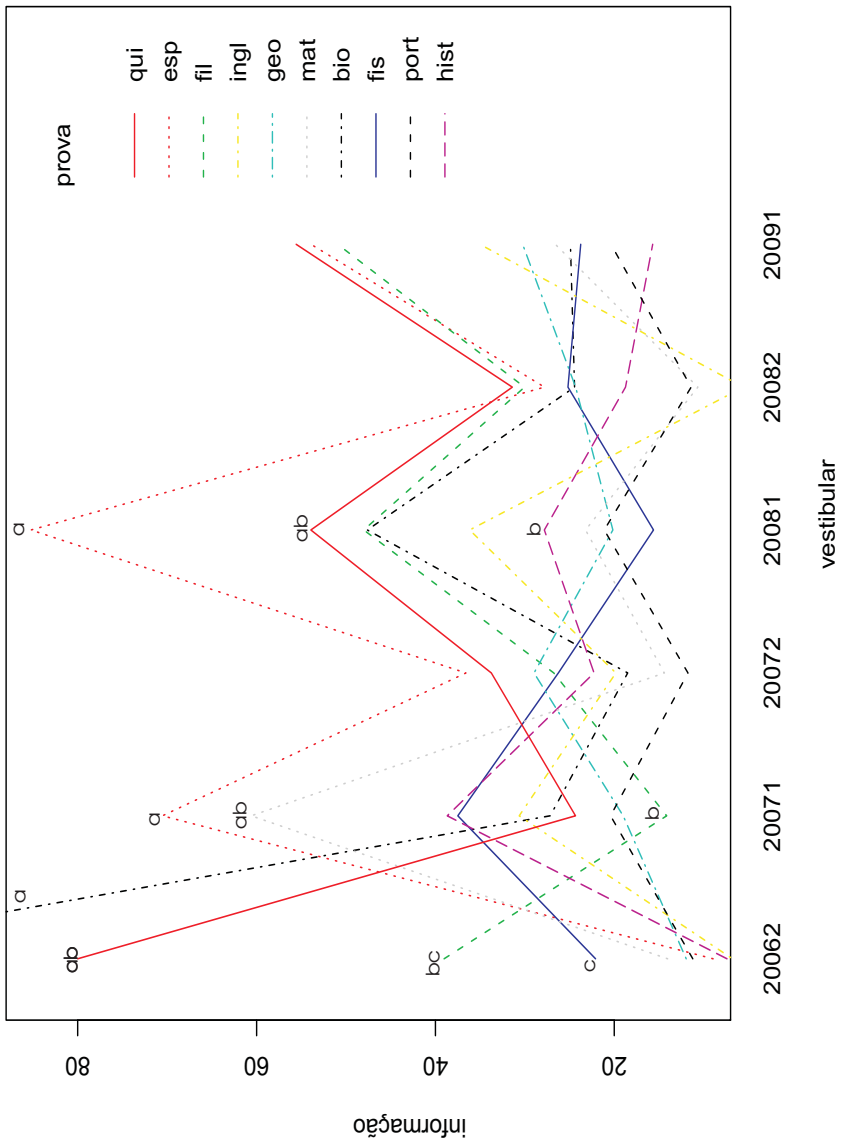


Figura 5 Gráfico da informação ponderada pelas habilidades dadas pelas provas ao longo dos Vestibulares de 2006-2 a 2009-1 da UFLA. Letras iguais na coluna indicam que não há diferenças significativas pelo teste de Tuckey. Para cursos que se encontram sem letras, repetem-se as letras do curso imediatamente superior

da pelas habilidades nos vestibulares 2006-2, 2007-1 e 2008-1, devido à ambiguidade de resultados é difícil dizer qual prova foi a mais ou a menos informativa. Entretanto, graficamente pode-se observar que a prova de Biologia foi a mais informativa do vestibular 2006-2 e a de Espanhol dos outros dois.

Comparando essa Figura 5 com a Figura 4 anterior pode-se observar que as provas da área de exatas - Física, Matemática e Química - estão entre as mais informativas. No entanto, quando essa informação é ponderada pela habilidade, as provas de Física e Matemática passam a apresentar menores valores. Note-se que o nível de habilidade dos candidatos estão aquém do nível das provas, isto é, essas disciplinas são muito informativas mas, como são poucos os candidatos mais hábeis, o valor ponderado passa a ser baixo. Pode-se dizer que são provas boas para informar a respeito dos melhores candidatos.

A prova de Química já foi consistentemente informativa, tanto quanto ao máximo valor quanto quando essa informação foi ponderada pela habilidade. O oposto pode-se verificar quanto à prova de Português que foi menos informativa tanto de forma absoluta como de forma ponderada.

A prova mais informativa quando ponderada pela habilidade tendeu a ser a prova de Espanhol. Isso quer dizer que foi a prova que mais informou para candidatos com média habilidade, pois é onde se encontra a maior massa de candidatos.

A representação gráfica da evolução das habilidades médias dos cursos ao longo desses anos encontra-se na Figura 6.

Pode-se observar que a habilidade média dos candidatos por curso não variou muito ao longo dos anos, isto é, não foi detectado nenhum curso que apresentasse evolução das médias das habilidades de seus candidatos. A exceção é apenas para o curso de Ciência da Computação, que teve um crescimento rápido

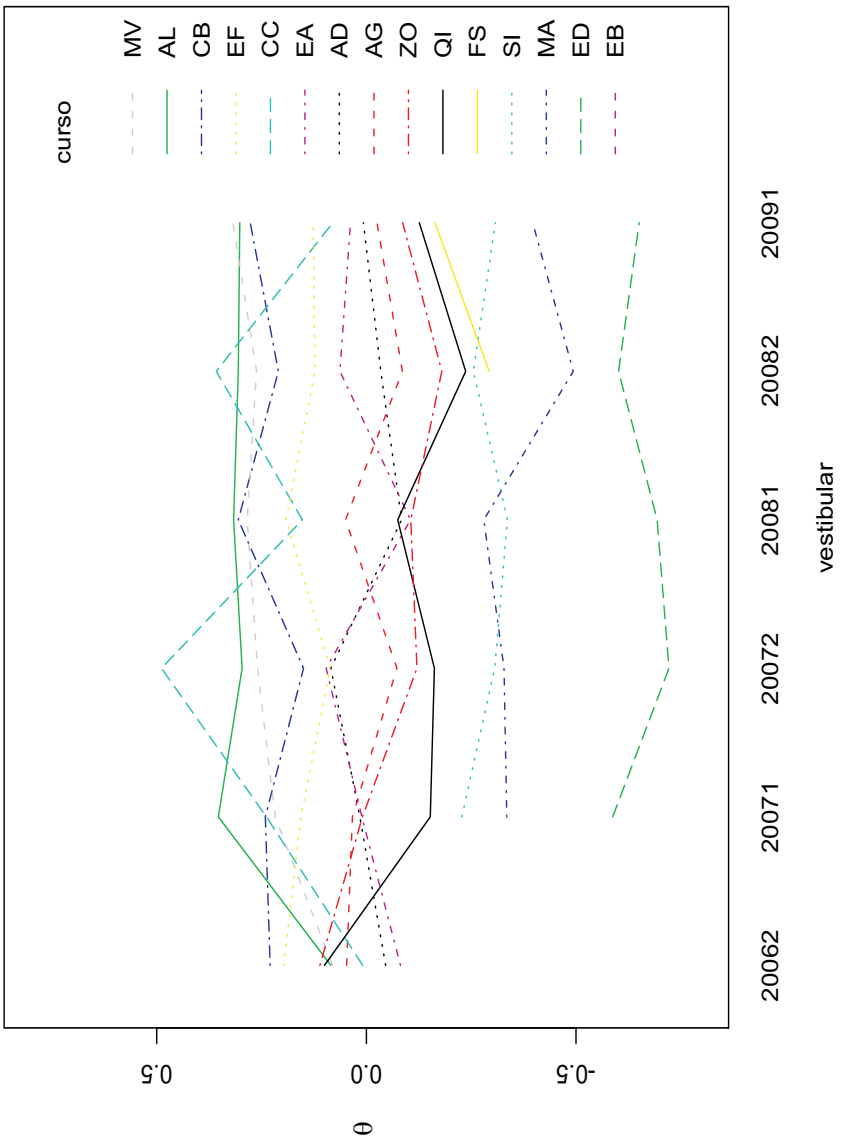


Figura 6 Gráfico da média das habilidades por curso ao longo dos vestibulares de 2006-2 a 2009-1 da UFLA

de habilidade e depois voltou a cair. Os candidatos do curso de Engenharia de Alimentos e de Medicina Veterinária foram os que apresentaram médias de habili-

dades mais estáveis e com maiores valores, sendo que o nível mais elevado de conhecimento ficou com os candidatos a Engenharia de Alimentos. O nível mais baixo foi obtido pelos candidatos ao curso de Educação Física. Baixos valores também foram obtidos pelos candidatos aos cursos de Física, Sistemas de Informação e Matemática. Esses quatro cursos referidos como os que apresentaram menor habilidade em média são cursos noturnos. Os demais são cursos diurnos.

Na Figura 7 encontra-se representado a relação entre o número de candidatos por vagas e respectivas habilidades médias para os cursos oferecidos ao longo desses vestibulares.

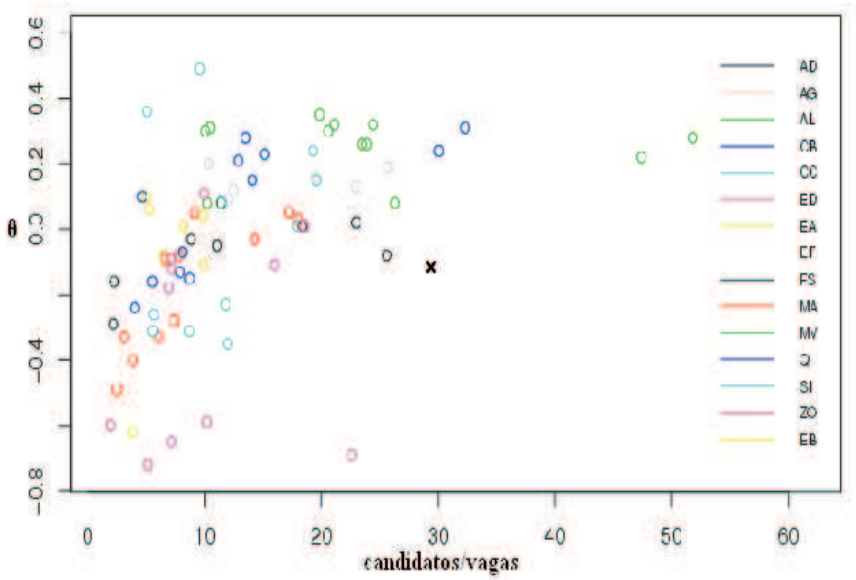


Figura 7 Gráfico do números de candidados por vaga × média das habilidades por curso ao longo dos vestibulares de 2006-2 a 2009-1 da UFLA

Pode-se observar que a habilidade média de cada curso é diretamente proporcional à concorrência para ele, isto é, quando a concorrência aumenta, aumenta-se também as habilidades médias dos candidatos aos respectivos cursos. Isto pode

ser explicado devido ao fato de que as habilidades estão associadas à procura por ele, ou seja, quando um curso está na "moda", a concorrência aumenta. Consequentemente, os candidatos têm que se preparar mais para enfrentá-la e isso implica diretamente no aumento do valor das estimativas de suas habilidades. A correlação entre essas duas variáveis foi de 0,49 com valor-p de $5,76 \times 10^{-6}$, ou seja, essa correlação é altamente significativa, reafirmando o fato da direta proporcionalidade entre a procura pelo curso e o nível de habilidade média dos candidatos ao mesmo.

4 CONCLUSÃO

As provas do vestibular da UFLA apresentaram baixa probabilidade de acerto por indivíduos com baixa habilidade e grau heterogêneo de dificuldade, discriminação e informação ponderada pela habilidade sendo que esta diferença varia com o vestibular.

O grau de dificuldade das provas está positivamente associado ao parâmetro de acerto por indivíduos com baixa habilidade.

Os cursos em que os candidatos possuem maiores níveis de habilidades são cursos diurnos, estando as médias dessas habilidades diretamente associadas à concorrência para os mesmos. Nos vestibulares de 2006-2 a 2009-1 da UFLA destacaram-se como os mais hábeis os candidatos de Engenharia de Alimentos e Medicina Veterinária.

REFERÊNCIAS

BACKER, F. B. **Item response theory**: parameter estimation techniques. New York: M. Dekker, 1992.

HAMBLETON, R. K.; COOK, L. L. Latent trait models and their use in the analysis of educational test data. **Journal of Educational Measurement**, Washington, v. 14, n. 2, p. 75-96, 1977.

MONTGOMERY, D. C. **Design and analysis of experiments**. 5. ed. New York: J. Wiley, 2001.

APÊNDICE

APÊNDICE A - Demonstração de fórmula da função informação do item

$$\pi_{ij} = c_j + \frac{1 - c_j}{1 + e^{-a_i(\theta_i - b_j)}} = \frac{c_j + c_j e^{-a_j(\theta_i - b_j)} + 1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}}$$

$$\pi_{ij} = \frac{1 + c_j e^{-a_j(\theta_i - b_j)}}{1 + e^{-a_j(\theta_i - b_j)}}$$

$$1 - \pi_{ij} = 1 - c_j - \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}} = \frac{(1 - c_j)\left[1 + e^{-a_j(\theta_i - b_j)}\right] - (1 - c_j)}{1 + e^{-a_j(\theta_i - b_j)}}$$

$$1 - \pi_{ij} = \frac{(1 - c_j)e^{-a_j(\theta_i - b_j)}}{1 + e^{-a_j(\theta_i - b_j)}}$$

$$L = \left[\frac{1 + c_j e^{-a_j(\theta_i - b_j)}}{1 + e^{-a_j(\theta_i - b_j)}}\right]^{Y_{ij}} \left[\frac{(1 - c_j)e^{-a_j(\theta_i - b_j)}}{1 + e^{-a_j(\theta_i - b_j)}}\right]^{1 - Y_{ij}}$$

$$lnL = Y_{ij} ln\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right] - Y_{ij} ln\left[1 + e^{-a_j(\theta_i - b_j)}\right] +$$

$$+(1 - Y_{ij}) ln(1 - c_j) - (1 - Y_{ij}) a_j(\theta_i - b_j) - (1 - Y_{ij}) ln\left[1 + e^{-a_j(\theta_i - b_j)}\right]$$

$$\frac{\partial lnL}{\partial \theta_i} = Y_{ij} \frac{c_j e^{-a_j(\theta_i - b_j)}(-a_j)}{1 + c_j e^{-a_j(\theta_i - b_j)}} - Y_{ij} \frac{e^{-a_j(\theta_i - b_j)}(-a_j)}{1 + e^{-a_j(\theta_i - b_j)}} - (1 - Y_{ij}) a_j +$$

$$-(1 - Y_{ij}) \frac{e^{-a_j(\theta_i - b_j)}(-a_j)}{1 + e^{-a_j(\theta_i - b_j)}}$$

$$= -\frac{a_j c_j Y_{ij} e^{-a_j(\theta_i - b_j)}}{1 + c_j e^{-a_j(\theta_i - b_j)}} + \frac{a_j Y_{ij} e^{-a_j(\theta_i - b_j)}}{1 + e^{-a_j(\theta_i - b_j)}} - (1 - Y_{ij}) a_j + \frac{(1 - Y_{ij}) a_j e^{-a_j(\theta_i - b_j)}}{1 + e^{-a(\theta - b)}}$$

$$\frac{\partial^2 lnL}{\partial \theta_i^2} = \frac{a_j c_j Y_{ij} e^{-a_j(\theta_i - b_j)}(-a_j)\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]}{\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]^2} +$$

$$+ \frac{a_j c_j Y_{ij} e^{-a_j(\theta_i - b_j)} c_j e^{-a_j(\theta_i - b_j)}(-a_j)}{\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]^2} +$$

$$+ \frac{a_j Y_{ij} e^{-a_j(\theta_i - b_j)}(-a_j)\left[1 + e^{-a_j(\theta_i - b_j)}\right] - a_j Y_{ij} e^{-a_j(\theta_i - b_j)} e^{-a_j(\theta_i - b_j)}(-a_j)}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2} +$$

$$+ \frac{(1 - Y_{ij}) a_j e^{-a_j(\theta_i - b_j)}(-a_j)\left[1 + e^{-a_j(\theta_i - b_j)}\right] - (1 - Y_{ij}) a_j e^{-a_j(\theta_i - b_j)} e^{-a_j(\theta_i - b_j)}(-a_j)}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2} =$$

$$= \frac{a_j^2 c_j Y_{ij} e^{-a_j(\theta_i - b_j)}\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right] - a_j^2 c_j^2 Y_{ij} e^{-2a_j(\theta_i - b_j)}}{\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]^2} +$$

$$\frac{-a_j^2 Y_{ij} e^{-a_j(\theta_i - b_j)}\left[1 + e^{-a_j(\theta_i - b_j)}\right] + a_j^2 Y_{ij} e^{-2a_j(\theta_i - b_j)}}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2} +$$

$$\frac{-a_j^2 (1 - Y_{IJ}) e^{-a_j(\theta_i - b_j)}\left[1 + e^{-a_j(\theta_i - b_j)}\right] + a_j^2 (1 - Y_{IJ}) e^{-2a_j(\theta_i - b_j)}}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2}$$

$$= \frac{a_j^2 c_j Y_{ij} e^{-a_j(\theta_i - b_j)} + a_j^2 c_j^2 Y_{ij} e^{-2a_j(\theta_i - b_j)} - a_j^2 c_j^2 Y_{ij} e^{-2a_j(\theta_i - b_j)}}{\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]^2} +$$

$$\frac{-a_j^2 Y_{ij} e^{-a_j(\theta_i - b_j)} - a_j^2 Y_{ij} e^{-2a_j(\theta_i - b_j)} + a_j^2 Y_{ij} e^{-2a_j(\theta_i - b_j)}}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2} +$$

$$\frac{-a_j^2 (1 - Y_{ij}) e^{-a_j(\theta_i - b_j)} - a_j^2 (1 - Y_{ij}) e^{-2a_j(\theta_i - b_j)} + a_j^2 (1 - Y_{ij}) e^{-2a_j(\theta_i - b_j)}}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2}$$

$$= \frac{a_j^2 c_j Y_{ij} e^{-a_j(\theta_i - b_j)}}{\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]^2} - \frac{a_j^2 Y_{ij} e^{-a_j(\theta_i - b_j)} + a_j^2 (1 - Y_{ij}) e^{-a_j(\theta_i - b_j)}}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2} =$$

$$= \frac{a_j^2 c_j Y_{ij} e^{-a_j(\theta_i - b_j)}}{\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]^2} - \frac{a_j^2 Y_{ij} e^{-a_j(\theta_i - b_j)} + a_j^2 e^{-a_j(\theta_i - b_j)} - a_j^2 Y_{ij} e^{-a_j(\theta_i - b_j)}}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2}$$

$$\frac{\partial^2 lnL}{\partial \theta_i^2} = \frac{a_j^2 c_j Y_{ij} e^{-a_j(\theta_i - b_j)}}{\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]^2} - \frac{a_j^2 e^{-a_j(\theta_i - b_j)}}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2}$$

$$I(\theta_i) = -E\left[\frac{\partial^2 lnL}{\partial \theta_i^2}\right]$$

$$E\left[\frac{\partial^2 lnL}{\partial \theta_i^2}\right] = E\left[\frac{a_j^2 c_j Y_{ij} e^{-a_j(\theta_i - b_j)}}{[1 + c_j e^{-a_j(\theta_i - b_j)}]^2}\right] - E\left[\frac{a_j^2 e^{-a_j(\theta_i - b_j)}}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2}\right] =$$

$$= \frac{a_j^2 c_j e^{-a_j(\theta_i - b_j)}}{\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]^2}\left[\frac{1 + c_j e^{-a_j(\theta_i - b_j)}}{1 + e^{-a_j(\theta_i - b_j)}}\right] - \frac{a^2 e^{-a(\theta - b)}}{\left[1 + e^{-a_j(\theta_i - b_j)}\right]^2} =$$

$$= \frac{\left[a_j^2 c_j e^{-a_j(\theta_i - b_j)} + a_j^2 c_j^2 e^{-2a_j(\theta_i - b_j)}\right]\left[1 + e^{-a_j(\theta_i - b_j)}\right]}{\left[1 + c_j e^{-a_j(\theta_i - b_j)}\right]^2 [1 + e^{-a_j(\theta_i - b_j)}]^2} +$$

$$-\frac{-a_j^2 e^{-a_j(\theta_i-b_j)}\left[1+2c_j e^{-a_j(\theta_i-b_j)}+c_j^2 e^{-2a_j(\theta_i-b_j)}\right]}{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]^2\left[1+e^{-a_j(\theta_i-b_j)}\right]^2}=$$

$$=\frac{a_j^2 c_j e^{-a_j(\theta_i-b_j)}+a_j^2 c_j^2 e^{-2a_j(\theta_i-b_j)}+a_j^2 c_j e^{-2a_j(\theta_i-b_j)}+a_j^2 c_j^2 e^{-3a_j(\theta_i-b_j)}}{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]^2\left[1+e^{-a_j(\theta_i-b_j)}\right]^2}+$$

$$+\frac{-a_j^2 e^{-a_j(\theta_i-b_j)}-2a^2 c e^{-2a(\theta-b)}-a_j^2 c_j^2 e^{-3a_j(\theta_i-b_j)}}{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]^2\left[1+e^{-a_j(\theta_i-b_j)}\right]^2}=$$

$$=\frac{a_j^2 c_j e^{-a_j(\theta_i-b_j)}+a_j^2 c_j^2 e^{-2a_j(\theta_i-b_j)}-a_j^2 c_j e^{-2a_j(\theta_i-b_j)}-a_j^2 e^{-a_j(\theta_i-b_j)}}{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]^2\left[1+e^{-a_j(\theta_i-b_j)}\right]^2}=$$

$$=a_j^2\frac{c_j e^{-a_j(\theta_i-b_j)}\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]-e^{-a_j(\theta_i-b_j)}\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]}{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]^2\left[1+e^{-a_j(\theta_i-b_j)}\right]^2}=$$

$$=-a_j^2\frac{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]}{\left[1+e^{-a_j(\theta_i-b_j)}\right]}\underbrace{\frac{e^{-a_j(\theta_i-b_j)}(1-c_j)}{\left[1+e^{-a_j(\theta_i-b_j)}\right]}}_{1-\pi_{ij}}\frac{1}{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]^2}$$

$$=-a_j^2\frac{1}{\left[1+e^{-a_j(\theta_i-b_j)}\right]}(1-\pi_{ij})\frac{1}{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]}$$

$$=-a_j^2\,(1-\pi_{ij})\frac{1}{\left[1+e^{-a_j(\theta_i-b_j)}\right]}\frac{1}{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]}\frac{\pi_{ij}}{(1-\pi_{ij})}$$

$$= -a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \frac{1}{\left[1+e^{-a_j(\theta_i-b_j)}\right]} \frac{1}{\left[1+c_j e^{-a_j(\theta_i-b_j)}\right]} \frac{1+c_j e^{-a_j(\theta_i-b_j)}}{1+e^{-a_j(\theta_i-b_j)}}$$

$$= -a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \left[\frac{1}{1+e^{-a_j(\theta_i-b_j)}}\right]^2$$

$$= -a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \left\{\frac{1-c_j}{(1-c_j)\left[1+e^{-a_j(\theta_i-b_j)}\right]}\right\}^2$$

$$= -a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \left\{\frac{1-c_j+c_j e^{-a_j(\theta_i-b_j)}-c_j e^{-a_j(\theta_i-b_j)}}{(1-c_j)\left[1+e^{-a_j(\theta_i-b_j)}\right]}\right\}^2$$

$$= -a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \left\{\frac{1+c_j e^{-a_j(\theta_i-b_j)}}{(1-c_j)\left[1+e^{-a_j(\theta_i-b_j)}\right]} - \frac{c_j\left[1+e^{-a_j(\theta_i-b_j)}\right]}{(1-c_j)\left[1+e^{-a_j(\theta_i-b_j)}\right]}\right\}^2$$

$$= -a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \left\{\frac{p_{ij}(\theta_i)}{1-c_j} - \frac{c_j}{1-c_j}\right\}^2$$

$$E\left[\frac{\partial^2 lnL}{\partial\theta_i^2}\right] = -a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \left\{\frac{p_{ij}-c_j}{1-c_j}\right\}^2$$

$$I(\theta_i) = -E\left[\frac{\partial^2 lnL}{\partial\theta_i^2}\right]$$

$$\therefore I_j(\theta_i) = a_j^2 \frac{(1-\pi_{ij})}{\pi_{ij}} \left[\frac{\pi_{ij}-c_j}{1-c_j}\right]^2$$