Análise do vestibular 2009-I da UFLA usando a TRI

MARIA DE LOURDES LIMA BRAGION*
JÚLIO SÍLVIO DE SOUSA BUENO FILHO**
FÁBIO MATHIAS CORRÊA***

RESUMO

Este trabalho teve como objetivo analisar as questões de múltipla escolha do vestibular 2009-I da Universidade Federal de Lavras (UFLA) por meio da TRI e utilizando a inferência bayesiana, a fim de identificar quais propriedades dos itens o classificam como melhor ou pior, com vistas a orientar novas provas. Verificou-se que os itens mais difíceis tendem a ser mais discriminativos e informativos para selecionar os candidatos mais hábeis. As provas de Inglês, Biologia, Física, Matemática e Química foram as mais difíceis e apresentaram maior poder de discriminação. Em geral o vestibular não apresentou itens com alta chance de acerto casual. Alguns itens tiveram suas estimativas analisadas em maior detalhe e revelam aspectos de planejamento que podem melhorar a qualidade desse tipo de prova.

Palavras-chave: Concurso vestibular, Teoria de resposta ao item, Prova de múltipla escolha, Análise qualitativa.

* Professora de Matemática do Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas – Campus Machado – MG (lourdinha.bragion@gmail.com).

** Professor Adjunto do Departamento de Ciências Exatas da Universidade Federal de Lavras – MG (jssbueno@dex.ufla.br).

*** Professor do Departamento de Estatística da Universidade Estadual de Santa Cruz – BA (fmcron@gmail.com).

RESUMEN

Este trabajo tuvo como objetivo analizar las preguntas de selección múltiple del examen de ingreso 2009-I de la Universidad Federal de Lavras (UFLA) por medio de la teoría de respuesta al ítem (TRI) y utilizando la inferencia bayesiana, con el fin de identificar qué propiedades de los ítems lo clasifican como mejor o peor y con el objetivo de orientar nuevas propuestas. Se verificó que los ítems más difíciles tienden a ser más discriminatorios e informativos para seleccionar a los candidatos más hábiles. Las pruebas de inglés, biología, física, matemática y química fueron las más difíciles y presentaron mayor poder de discriminación. En general, la prueba de ingreso no presentó ítems con alta chance de acierto casual. El análisis más detallado de las estimativas de algunos ítems reveló aspectos del planeamiento que pueden mejorar la calidad de ese tipo de prueba.
Palabras clave: Exámenes de ingreso, Teoría de la respuesta al ítem, Prueba de selección múltiple, Análisis cualitativo.

ABSTRACT

This study aimed to analyze the multiple-choice questions of the 2009-I entrance examination ("vestibular") of the Federal University of Lavras (UFLA) by using IRT and Bayesian inference in order to identify which properties of the items qualify as better or worse, in order to inform new exams. It was found that the most difficult items tend to be more discriminative and more informative to select the most skilled candidates. Tests of English, Biology, Physics, Mathematics and Chemistry were the most difficult and had the highest discrimination power. Overall, this "vestibular" did not show that guessing would ensure a high probability of success. Some items estimates were analyzed in detail to unveil planning properties which could be used to enhance the quality of this kind of exam.
Keywords: Entrance exam, Item response theory, Multiple choice test, Qualitative analysis.

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 esse 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 isso, um importante fator a ser considerado é a qualidade dos itens, pois, além do interesse mais imediato dos examinadores, que é selecionar candidatos, há um interesse adicional em identificar que itens ou tipos de itens são mais eficientes para essa seleção, assim como observar em que medida as diferentes disciplinas têm contribuído para isso, o que pode orientar eventuais alterações na estrutura das provas e no ensino médio e preparatório.

A teoria de resposta ao item (TRI) tem sido usada como uma poderosa ferramenta para a avaliação educacional. 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 em razão de discussões sobre a viabilidade de se compararem as habilidades de indivíduos submetidos a provas diferentes (Hambleton, Swaminathan e Rogers, 1991), que sempre foi uma das limitações da teoria clássica dos testes (TCT). Conforme Baker (1992) e Hambleton e Cook (1977), alguns dos principais benefícios da TRI 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.

Apesar de, nas últimas décadas, a teoria de resposta ao item (TRI) ter sido aplicada com sucesso para a construção e análise de diferentes tipos de testes, poucos estudos ainda são encontrados com relação à sua aplicação para questões de vestibular, sendo utilizada uma metodologia bayesiana. Um exemplo de TRI aplicada a dados de vestibular, fazendo uso dessa metodologia, pode ser encontrado em Bragion e Bueno Filho (2007). No entanto, o estudo realizado por eles é feito apenas para uma prova isolada (candidatos da Agronomia do vestibular 2006-2 da UFLA),

em que foram consideradas apenas as questões comuns a todos os estudantes, a fim de garantir um delineamento balanceado.

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*), também conhecido como parâmetro de acerto casual ou de "chute"; e os das habilidades (θ), às características dos candidatos que respondem a estas questões.

O modelo mais utilizado na literatura é o modelo logístico de três parâmetros (ML3P), conhecido como modelo de Birnbaum (1968), cuja expressão é dada por:

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

em que:

Y_{ij}: variável dicotômica associada ao acerto (igual a 1) ou erro (igual a 0) na resposta do indivíduo *i* ao item *j*, *i* = 1, ..., *n*, *j* = 1, ..., *k*;

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.

b_j: parâmetro do grau de dificuldade, ou índice de locação. Representa o ponto na escala de habilidade no qual o indivíduo, com habilidade $\theta = b_j$, possue $(c_j + 1)/2$ de probabilidade de responder corretamente a esse item;

c_j: parâmetro da probabilidade de que um indivíduo com baixa habilidade responda corretamente o item *j*;

θ_i: habilidade do indivíduo *i*;

$P(\theta) = P(Y_{ij} = 1|\theta, a, b, c)$: probabilidade condicional de que o indivíduo *i* responda corretamente ao item *j*;

n: número de indivíduos;

k: número de itens.

O ML3P pode ser expresso graficamente pela curva característica do item (CCI). Cada item do teste terá sua própria CCI.

Outras importantes ferramentas utilizadas pela TRI são a curva característica do teste (CCT), a função informação do item (FII) e a função informação do teste

(FIT). A CCT envolve todo o conjunto de questões. É obtida somando-se todas as probabilidades computadas com a CCI de todos os itens (Baker, 2001). Pode ser comparada à nota de um indivíduo num determinado teste. A FII permite analisar quanto um item traz de informação para a medida de habilidade. É útil para identificar as questões que são realmente relevantes. Para o ML3P é dado pela expressão:

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

em que: $\pi_{ij} = P(\theta)$

A FIT, assim como a CCT, envolve o teste todo e é obtida pela soma das informações de cada item que compõe o mesmo (Baker, 2001). É um recurso extremamente útil da TRI, pois, por meio dela, é possível saber quão bom está sendo o teste todo em fornecer a informação sobre a habilidade de interesse.

INFERÊNCIA BAYESIANA

A estimação dos parâmetros do modelo utilizado pode ser feita via inferência frequentista ou bayesiana. Na inferência bayesiana, toda informação que se tenha sobre o parâmetro, antes de ser observada a amostra, deve ser considerada e incorporada aos dados, por meio de uma distribuição a *priori*. A distribuição dos dados é chamada de função de verossimilhança. A distribuição resultante entre a junção da distribuição a *priori* com a função de verossimilhança é chamada distribuição a *posteriori* e é com base nela que é feita toda inferência bayesiana (O'Hagan, 1994). Para o caso contínuo, por envolver integrações trabalhosas ou mesmo impossíveis, a inferência sobre o parâmetro de interesse é feita por intermédio de métodos de aproximações numéricos. Um método bastante utilizado é o método de Monte Carlo por via da cadeia de Markov (MCMC), pelo qual é obtida uma amostra da função de interesse. Dos métodos de simulação que utilizam cadeias de Markov, destacam-se o amostrador de Gibbs (AG) e o algoritmo Metropolis-Hastings (MH). Maiores detalhes podem ser encontrados em Gamerman e Lopes (2006).

Este trabalho teve como objetivo analisar as questões de múltipla escolha do vestibular 2009-1 da UFLA por meio da TRI e utilizando a inferência bayesiana, a fim de identificar quais propriedades dos itens o classificam como melhor ou pior, com vistas a orientar novas provas. Vale ressaltar o fato de que um estudo aplicado a

dados de vestibular virá a contribuir para que se tenha 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.

MATERIAL E MÉTODOS

Material

Os dados utilizados referem-se ao vestibular 2009-1 da Universidade Federal de Lavras (UFLA) composto de 66 itens de múltipla escolha com 4 alternativas cada um. Como a disciplina de Língua Estrangeira possui a opção de escolher entre Inglês e Espanhol, duplicou-se o número de itens referentes a ela, alterando-se para 74 no total. A distribuição desses itens[1] por disciplina encontra-se na tabela 1.

Tabela 1 – Disciplinas e itens do vestibular 2009-I da UFLA

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

Os cursos oferecidos e o número de candidatos inscritos estão na tabela 2.

[1] O conteúdo de cada um dos itens do vestibular analisado está disponível em: <$http://www.copese.ufla.br/copese/provasAnteriores.asp?tipo=V$>.

Tabela 2 – Relação de candidatos e cursos do vestibular 2009-I da UFLA

Curso	Candidatos inscritos
Administração (AD)	387
Agronomia (AG)	807
Engenharia de Alimentos (AL)	289
Ciências Biológicas (CB)	371
Ciências da Computação (CC)	263
Educação Física (ED)	221
Engenharia Agrícola (EA)	138
Engenharia Florestal (EF)	317
Física (FS)	44
Matemática (MA)	63
Medicina Veterinária (MV)	686
Química (QI)	108
Sistema de Informação (SI)	196
Zootecnia (ZO)	197
Total	4087

Foi elaborada uma rotina de análise[2] para implementar a metodologia escrita em linguagem C programada para rodar na linguagem R (R Development Core Team, 2009). Utilizou-se para isso um computador equipado com processador Core i7 - 965 - 3.20 ghz com 12 gb de memória RAM.

Métodos

O modelo matemático utilizado foi o ML3P. A estimação dos parâmetros foi feita via inferência bayesiana usando o algoritmo MH para obter amostras da distribuição a *posteriori* dos parâmetros.

As distribuições a *priori* escolhidas para cada parâmetro foram: distribuição normal (0,1) para cada elemento de θ; distribuição log-normal (0;0,5) para cada elemento de *a;* distribuição uniforme (-3,3) para cada elemento de *b;* e distribuição beta (2,4) para cada elemento de *c*.

Para identificação de quais itens eram bons ou não e realmente relevantes para selecionar os candidatos, assim como os que mais informavam sobre eles, foram construídas a CCI e a FII para cada um deles.

[2] A rotina de análise implementada em R pode ser encontrada na tese de doutorado da primeira autora e está sendo submetida a publicação.

Em um mesmo gráfico foram plotados a CCI com seu respectivo HPD (intervalo de máxima densidade a *posteriori*), a FII e o histograma das habilidades estimadas, a fim de melhor visualizar não somente as características dos itens quanto ao seu grau de dificuldade, poder de discriminação e probabilidade de acerto por indivíduos com baixa habilidade, mas 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 para 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 tabela 1), foi dividido cada valor total da CCT e FIT pelo número de itens de cada prova, a fim de torná-las com iguais condições de comparação.

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, para verificar quanto o teste, como um todo, foi informativo.

RESULTADOS E DISCUSSÃO

Nas figuras 1 a 3 estão representadas as estimativas dos parâmetros dos 74 itens com seus respectivos intervalos de credibilidade HPDs. Pode-se observar que vários itens apresentaram elevados valores para o poder de discriminação, sendo 19 o número de itens com valores de $a > 1{,}70$. Esses itens foram os de número 1 e 10 (ambos de Português), 12 (Geografia), 21 (História), 33 (Espanhol), 45 e 50 (Biologia), 51, 52, 55, 56, 57 e 58 (Física), 60, 63 e 66 (Matemática) e 70, 71 e 72 (Química). Com exceção dos itens de número 10 e 50, todos esses itens apresentaram elevado grau de dificuldade. Comparando-se as figuras 1 e 2, percebe-se que, em geral, os itens mais difíceis são também os que apresentam mais altos valores para o parâmetro *a*. A mesma observação pode ser feita em relação às provas. Observando-se a figura 2 pode-se verificar que os itens mais difíceis encontram-se, em geral, do número 33 para cima e se referem às provas de Inglês, Biologia, Física, Matemática e Química. Comparando-se com a figura 1 percebe-se que essas provas foram também as que apresentam maior poder de discriminação.

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

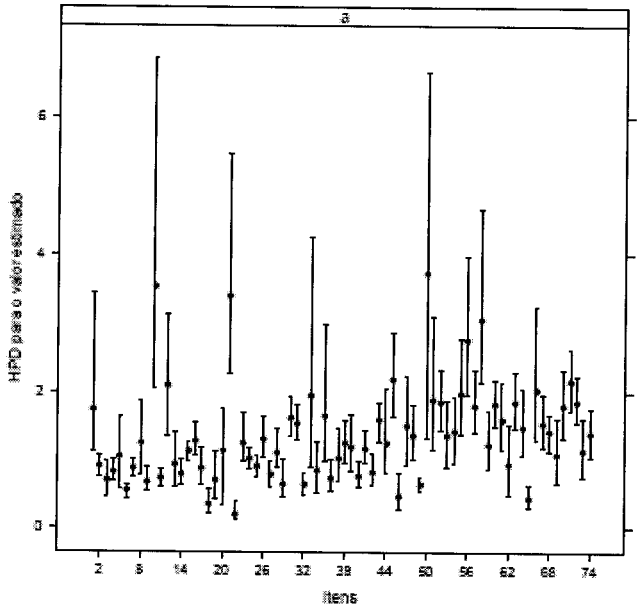


Figura 2 – Estimativas pontuais do parâmetro *b* dos itens do vestibular 2009-I da UFLA e respectivos intervalos de credibilidade HPD a 95%

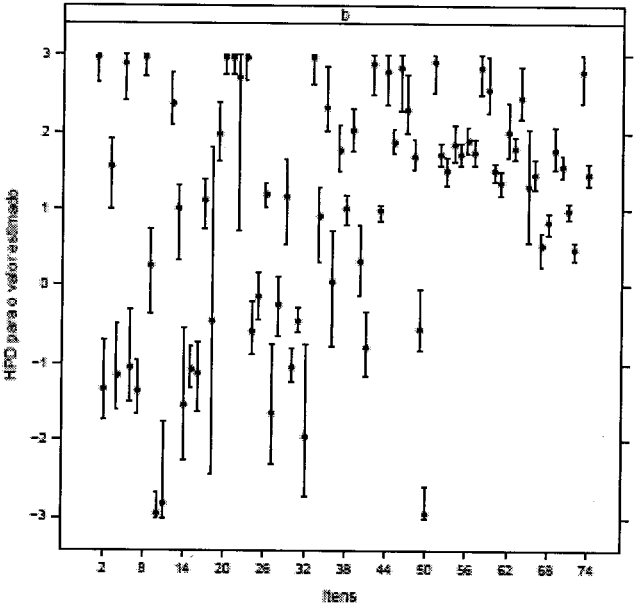
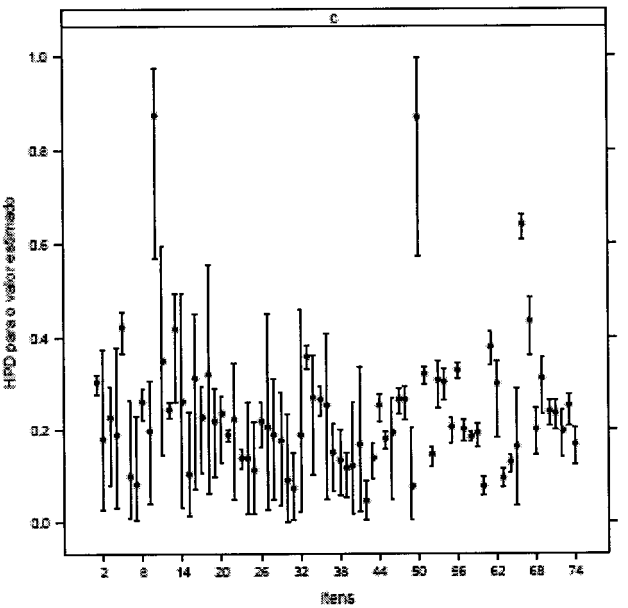
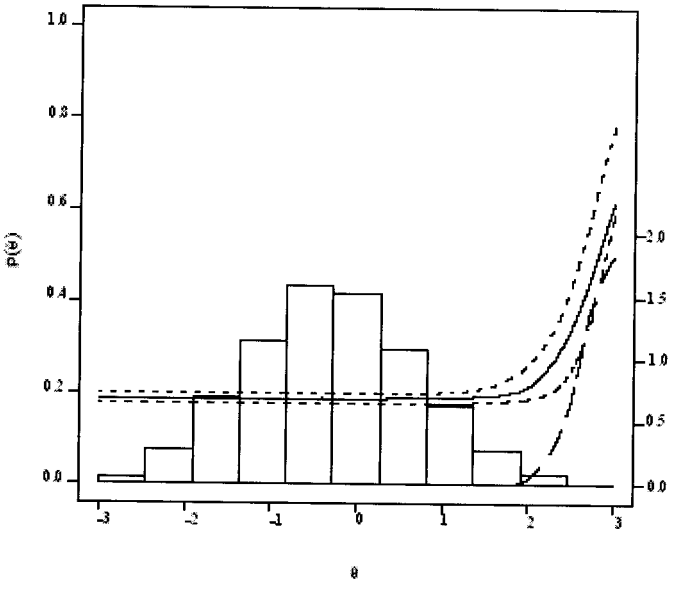


Figura 3 – Estimativas pontuais do parâmetro *c* dos itens do vestibular 2009-I da UFLA e respectivos intervalos de credibilidade HPD a 95%



Itens muito difíceis foram os de número 1, 5, 8 (Português), 20, 21, 23 (História), 33 (Espanhol), 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. Por essas características, poderia ser considerado um item muito bom. No entanto, olhemos sua representação gráfica que se encontra na figura 4. Pode-se observar que, apesar de ter apresentado um elevado valor para a estimativa do parâmetro de discriminação, essa discriminação só é informativa para indivíduos com habilidade muito elevada. Isso se deve ao fato de que esse item é muito difícil. Observa-se que, entre os candidatos que têm habilidade abaixo de 2, a probabilidade de acerto para todos eles é baixa. Assim, apesar do interesse em que se tenha itens com elevado poder de discriminação, se o grau de dificuldade for muito elevado, mesmo que haja alguns candidatos com níveis de habilidades muito elevados, a diferença de probabilidades de acerto entre eles será muito pequena e, portanto, será difícil concluir qual deles seria o mais hábil.

Figura 4 – Histograma das habilidades, CCI (____) e intervalo de credibilidade HPD a 95% (- - -; probabilidades na escala à esquerda), curva de informação do item (__ __ __; conteúdo de informação na escala à direita) do item 21. (História: $a = 3{,}36$; $b = 2{,}97$; $c = 0{,}17$)

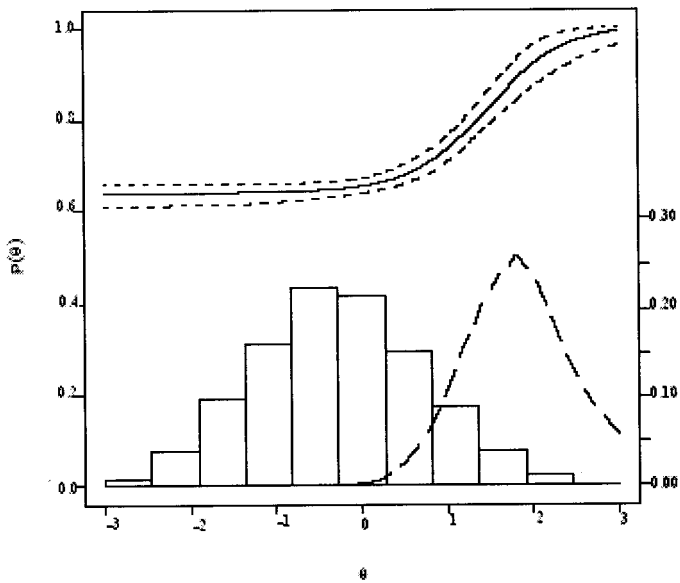


Dos itens citados como os que apresentaram maior poder de discriminação, excetuaram-se dois deles, os de número 10 e 50, pois, apesar da estimativa elevada para o parâmetro *a*, foram itens muito fáceis (os dois mais fáceis de todo o vestibular). Observando-se a figura 3, verifica-se que ambos apresentaram elevada probabilidade de acerto por indivíduos com baixa habilidade. Parece um pouco estranho que itens com grau de dificuldade muito baixo tenham tão elevado valor para o parâmetro *c*. Buscando-se uma explicação para essa discrepância, verificou-se que se tratava de itens que foram anulados. Quando a questão é anulada todos os candidatos recebem 1 ponto para ela, independentemente da alternativa escolhida. Consequentemente, o "acerto" é garantido, "chutando-se" ou não, isto é, como qualquer alternativa assinalada está "correta", mesmo os candidatos com baixa habilidade ganham ponto nessa questão. Isso explica o fato de a estimativa para o valor do parâmetro *c* ter sido tão elevada e para o grau de dificuldade muito baixa.

Essa ocorrência pode ser vista como um fator que confirma a confiabilidade da metodologia utilizada e do algoritmo empregado.

Excluindo-se esses dois itens – 10 e 50 – por serem questões anuladas, ainda se pode observar, na figura 3, que o item de número 66 (Matemática) também apresentou elevada probabilidade de acerto por indivíduos com baixa habilidade. Sua representação gráfica encontra-se na figura 5.

Figura 5 – Histograma das habilidades, CCI (____) e intervalo de credibilidade HPD a 95% (- - -; probabilidades na escala à esquerda), curva de informação do item (__ __ __; conteúdo de informação na escala à direita do item 66 (Matemática: $a = 2{,}19$; $b = 1{,}46$; $c = 0{,}64$)



Pode-se observar que um candidato com habilidade -3, que é muito baixa, tem 64% de probabilidade de acertar essa questão. No entanto, como explicar o elevado poder de discriminação ($a = 2{,}19$)? A maior quantidade de informação é obtida em torno do valor 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* desse item) têm-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.

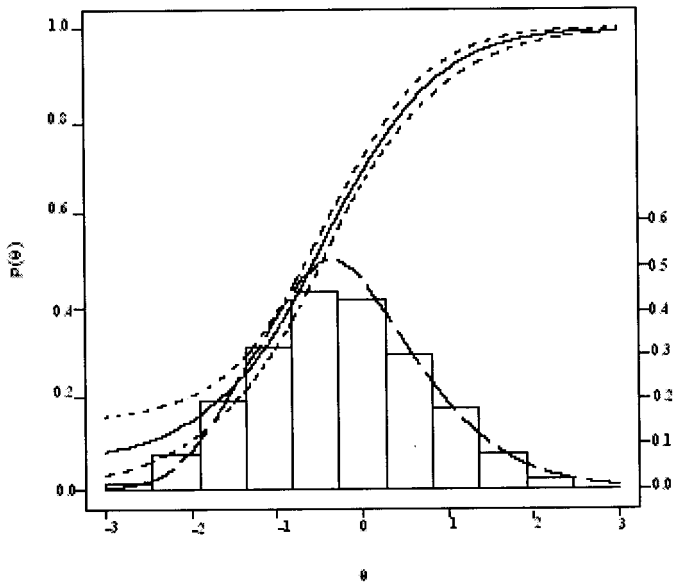
Vamos supor que esse item tivesse $a = 1{,}70$, que já é considerado muito alto, e manter os mesmos valores para os outros dois parâmetros. Calculando-se as

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 alternativa 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 se diminuiu também o valor de *c*, essa diferença entre as probabilidades voltou a elevar-se. Isso indica que quando um item possui elevado valor, tanto para o parâmetro *a* quanto 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, ou seja, a 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 de o 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 exagerado poder de discriminação, nesse caso.

Gráficos como os das figuras 4 e 5 foram desenvolvidos para todos os itens, no entanto apenas mais três deles serão apresentados. Como itens com representação gráfica semelhante possuem as mesmas características, foram escolhidos aqueles que representam situações típicas diferentes, visando a extrair, dos mesmos, características que auxiliarão no planejamento de novas provas. Assim, foram escolhidos os itens 31 (Espanhol), 60 (Matemática) e 22 (História).

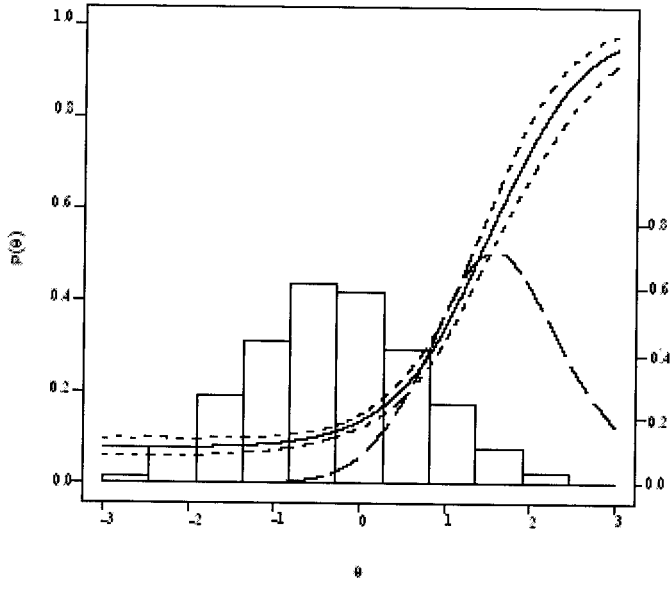
Os itens 31 e 60 estão representados nas figuras 6 e 7, respectivamente. Esses dois itens foram escolhidos por apresentarem valores para os parâmetros *a* e *c* semelhantes (ambos com boa discriminação e baixa probabilidade de acerto por indivíduos com baixa habilidade), mas com graus de dificuldade 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 6 – Histograma das habilidades, CCI (____) e intervalo de credibilidade HPD a 95% (- - -; probabilidades na escala à esquerda), curva de informação do item (__ __ __; conteúdo de informação na escala à direita) do item 31 (Espanhol: $a = 1{,}56$; $b = -0{,}44$; $c = 0{,}07$)



O item 31 (Figura 6) apresenta um alto valor para o parâmetro *a*, grau de dificuldade adequado ao nível dos candidatos com habilidade média e baixa probabilidade de acerto por aqueles com baixa habilidade. Trata-se, portanto, de um item bom para discriminar entre os indivíduos que possuem habilidade mediana. A diferença entre as probabilidades de acerto de candidatos com habilidade 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.

Figura 7 – Histograma das habilidades, CCI (____) e intervalo de credibilidade HPD a 95% (- - -; probabilidades na escala à esquerda), curva de informação do item (__ __ __; conteúdo de informação na escala à direita) do item 60 (Matemática: $a = 1{,}83$; $b = 1{,}49$; $c = 0{,}08$)



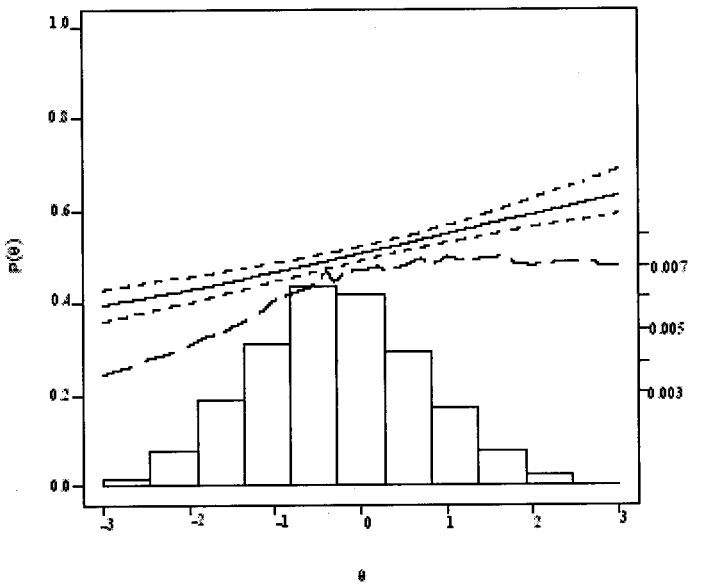
O item 60 (Figura 7) difere mais acentuadamente do item 31 com relação ao grau de dificuldade. Pode-se observar que o item 60 apresentou um grau de dificuldade superior à média da população. No entanto, essa dificuldade é compatível com o nível de habilidade de um grupo razoável dos melhores candidatos. Como também apresenta elevado valor para o parâmetro *a*, trata-se de um item bom para discriminar entre eles.

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 é bem maior.

Por exemplo, 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 de 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 auxiliarão na elaboração de outros novos, para próximos exames.

Observe-se agora o item 22 que se encontra representado na figura 8. Esse é um item muito ruim, pois, apesar de o valor para o parâmetro *c* não ter sido elevado, apresentou baixo poder de discriminação. Indivíduos com diferentes níveis de habilidades possuem praticamente a mesma probabilidade de acerto, dificultando a distinção entre os mais e os menos hábeis.

Figura 8 – Histograma das habilidades, CCI (____) e intervalo de credibilidade HPD a 95 % (- - -; probabilidades na escala à esquerda), curva de informação do item (__ __ __; conteúdo de informação na escala à direita) do item 22 (História: $a = 0{,}19$; $b = 2{,}70$; $c = 0{,}22$)



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.

Com base nessas observações, pode-se dizer que, para uma avaliação, os itens mais interessantes são aqueles que discriminam bem, possuem baixa probabilidade de acerto por indivíduos com baixa habilidade (valores compatíveis com 0,25, no contexto do número de alternativas do vestibular analisado) e apresentam diferentes graus de dificuldade (lembrando que a informação será maior quanto mais discriminativo for o item). Isto porque, se a questão 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 discriminação 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 quanto um indivíduo com baixo nível de habilidade sabe e quanto um indivíduo com alto nível de habilidade não sabe. Referindo-se, novamente, aos 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.

Cabe ressaltar o fato de que 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 identificação 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 junto com o histograma das estimativas das habilidades estão representadas nas figuras 9 e 10.

Figura 9 – Histograma das habilidades, CCT por disciplina (_____) e seus intervalos de credibilidade HPD a 95% (- - - ; probabilidades nas escalas à esquerda) e curva de informação por disciplina (__ __ __ ; conteúdo de informação nas escalas à direita) para as provas de Português, Geografia, História, Filosofia, Espanhol, Inglês), dispostas nessa ordem

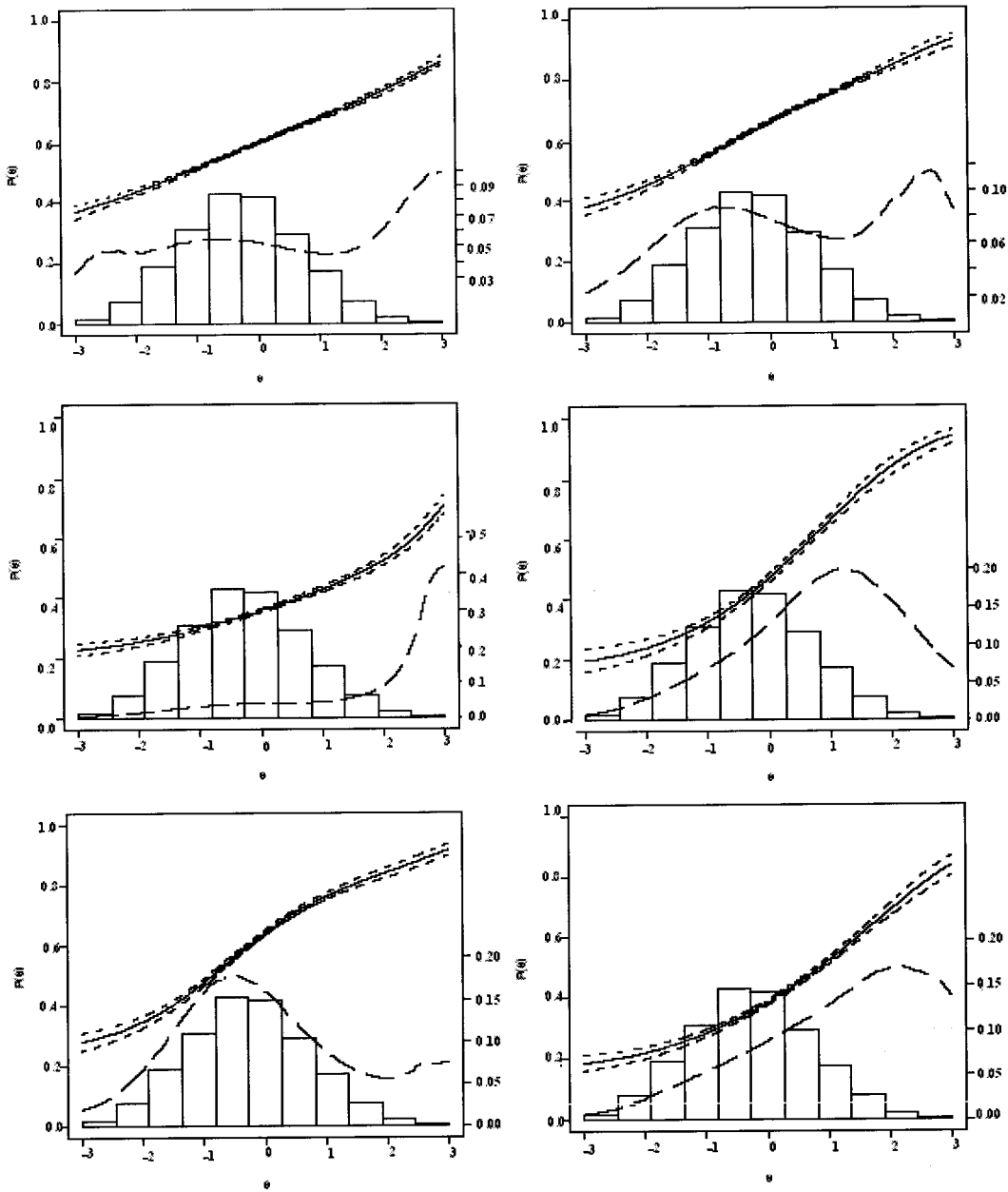
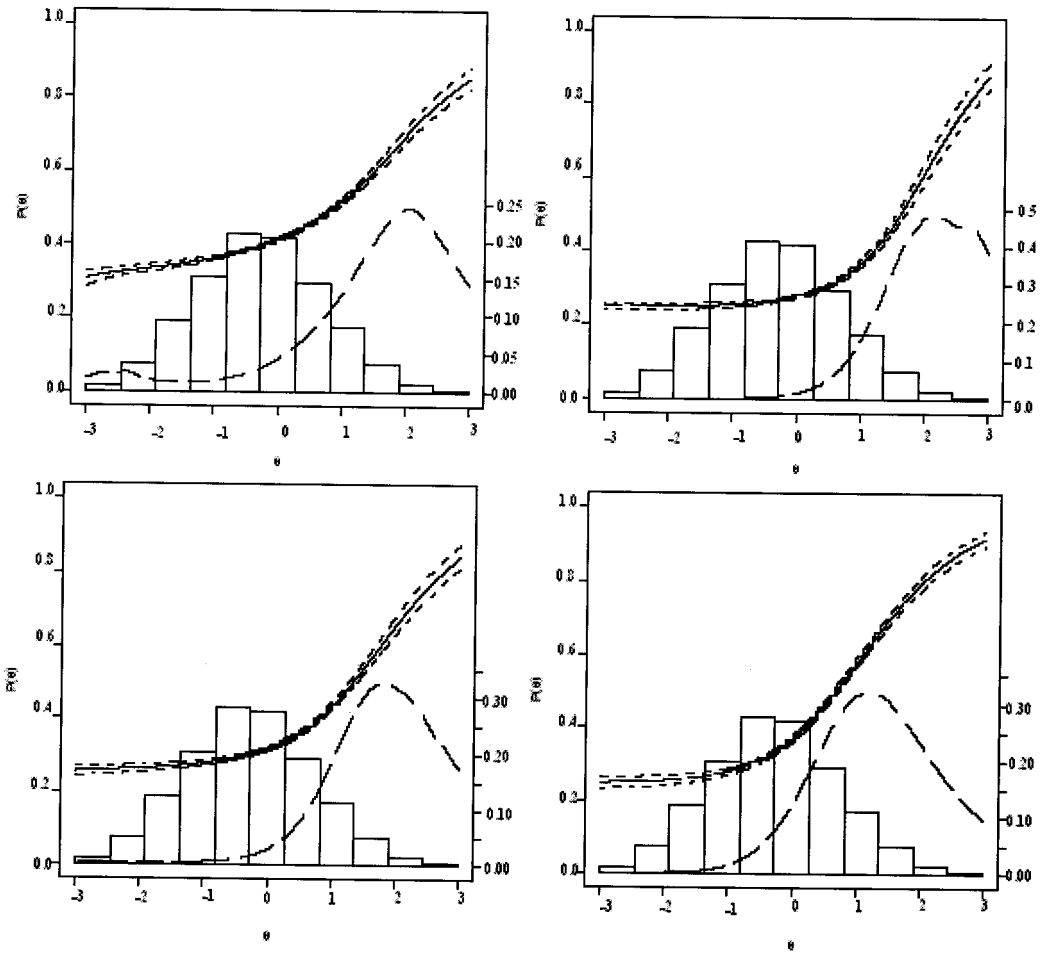


Figura 10 – Histograma das habilidades, CCT por disciplina (____) e seus intervalos de credibilidade HPD a 95% (- - - ; probabilidades nas escalas à esquerda) e curva de informação por disciplina (__ __ __ ; conteúdo de informação nas escalas à direita) para as provas de Biologia, Física, Matemática e Química, dispostas nessa ordem



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.

É interessante relembrar a relação que tem a informação com o grau de dificuldade e com o poder de discriminação. O grau de dificuldade da prova pode ser determinado pela posição da CCT. Quanto mais difícil a prova, mais à direita estará a CCT e quanto mais inclinada estiver a CCT, maior será seu poder de discriminação. Como a informação é maior em torno do parâmetro *b* ocorre que curvas de informação situadas mais à

direita são mais informativas para o grupo dos melhores candidatos, e elas são resultantes de itens mais difíceis. Conforme vai-se diminuindo o grau de dificuldade, passa a ser mais informativa para habilidades menores. Se a prova for muito difícil a informação será apenas para habilidades muito altas e talvez onde haja pouquíssimos indivíduos, não informando praticamente nada a respeito da realidade do grupo dos candidatos em questão. Se o item é muito fácil (seu gráfico 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. Assim, 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. Entretanto, vale ressaltar que 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 quanto o menos hábil desconhece.

Assim, de acordo com o que se espera de um processo seletivo, pode-se dizer que as 4 últimas provas (Biologia, Física, Matemática e Química) foram as que mais atingiram seus objetivos (Figura 10). Isso porque eram 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 sendo muito pouco informativa para aqueles que possuem habilidade abaixo de 2.

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

Quanto às provas de Língua Estrangeira, pode-se verificar que tanto a prova de Espanhol como a de Inglês apresentaram a mesma quantidade de informação. Entretanto, a prova de Inglês foi mais informativa para o grupo de interesse do que a de Espanhol, pois teve um grau de dificuldade maior. Já a prova de Espanhol, em razão do grau de dificuldade menor, foi boa para discriminar entre indivíduos com habilidade mediana.

Foi possível estimar as habilidades de todos os candidatos. A média geral obtida foi de -0,01 com desvio padrão de 0,88. O valor da correlação entre as habilidades estimadas e suas respectivas notas foi de 0,94.

Na tabela 3 encontram-se as médias das habilidades estimadas para cada curso separadamente com seus respectivos desvios padrão.

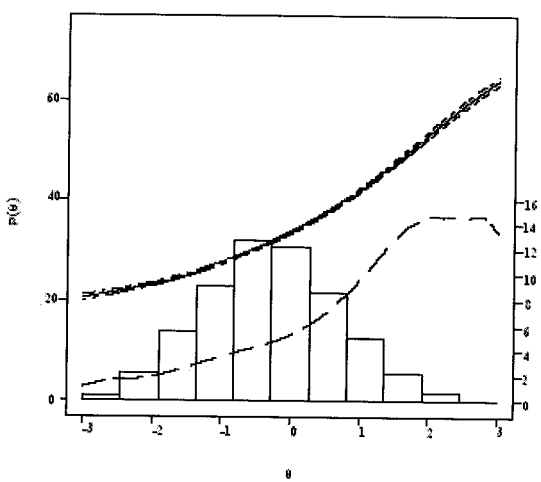
Tabela 3 – Médias das habilidades por curso do vestibular 2009-I e respectivo desvio padrão (sd)

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

Observa-se que o curso que obteve menor habilidade média foi o de Educação Física, e o de 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.

Na figura 11 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 todo.

Figura 11 – Histograma das habilidades, CCT (____) junto ao seu intervalo de credibilidade HPD a 95% (- - -; probabilidade na escala à esquerda) e curva de informação do teste (__ __ __; conteúdo de informação na escala à direita)



Por meio dessa figura, pode-se observar que o maior conteúdo de informação é obtido por indivíduos com habilidade acima da média da população e em uma região onde há interesse em que estes sejam os selecionados. O grau de dificuldade foi condizente com esse nível de habilidade. Considerando-se que acima de um desvio padrão da média de uma distribuição normal há aproximadamente 16% da população, para esse vestibular isso corresponde a buscar quantos indivíduos possuem habilidade acima de 0,87, que é o grupo onde se encontram aqueles que interessa selecionar. Houve 680 candidatos nessas condições. 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.

CONSIDERAÇÕES FINAIS

Por meio da TRI é possível reunir elementos para discutir a qualidade de questões e provas de vestibulares com vistas à futura melhoria da qualidade dos exames. Um item é mais informativo para o grupo de interesse quando possui baixa probabilidade de acerto por indivíduos com baixa habilidade, grau de dificuldade condizente com o nível de habilidade desse grupo e alto poder de discriminação. Pôde-se ainda concluir que:

a) as provas mais informativas foram as de Biologia, Física, Matemática e Química;

b) o curso de maior habilidade média foi o de Engenharia de Alimentos;

c) o vestibular apresentou baixa probabilidade de acerto por indivíduos com baixa habilidade.

REFERÊNCIAS BIBLIOGRÁFICAS

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

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

_____. *The basics of item response theory*. 2. 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.

GAMERMAN, D.; LOPES, H. *Markov chain Monte Carlo*: stochastic simulation for bayesian inference. 2. ed. London: Chapman & Hall, 2006. (CRC Texts in Statistical Science).

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

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.

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

R DEVELOPMENT CORE TEAM *R: a language and environment for statistical computing, reference index: version 2.9.0*. Vienna: R Foundation for Statical Computing, 2009. Disponível em: <http://www.R0project.org>. Acesso em: 5 jul. 2009.

Recebido em: março 2011
Aprovado para publicação em: maio 2011